CLASIFICACIÓN NOPARAMÉTRICA EN DATOS DIRECCIONALES

Por

Santiago Antonio Velasco Forero

Tesis sometida en cumplimiento parcial de los requisitos para el grado de

MAESTRO EN CIENCIAS
en
MATEMÁTICAS
(Estadística)

UNIVERSIDAD DE PUERTO RICO
RECINTO UNIVERSITARIO DE MAYAGUEZ
2004

Aprobado por:

Edgardo Lorenzo, Ph.D Miembro, Comité Graduado	Fecha
Pedro Vásquez, D.Sc Miembro, Comité Graduado	Fecha
Edgar Acuna, Ph.D Presidente, Comité Graduado	Fecha
Rafael Segarra, Ph.D Representante de Estudios Graduados	Fecha
Pedro Vásquez, D.Sc Director de Departamento	Fecha

Abstract

In a supervised classification problem, when the vectors of data are directional, it means, that they take values on a k-dimensional sphere, the application of the algorithms of pattern recognition as k-nearest-neighbour method, discriminant analysis, and kernel discriminant analysis, do not obtain good results in classification error rate. For this type of problems, we propose several algorithms based on directional k-nearest-neighbour, estimation of density for directional kernel and discriminant analysis with assumption of von Mises-Fisher distribution. Additionally we present an extension for these classification methods for directional data in standard sets (not directional), based in the correlation matrix. We illustrate the performance of these methods on simulated data, machine learning datasets and microarray data sets.

RESUMEN

Si se tiene un problema de clasificación supervizada donde los vectores de datos son direccionales, es decir, toman valores sobre una esfera k-dimensional, la aplicación de los métodos de reconocimiento de patrones tales como k-vecinos más cercano, análisis de discriminante y clasificación por estimación de densidad por kernel, no tienen buenos resultados en cuanto a tasa de error en la clasificación. Se propone para este tipo de problemas, algoritmos basados en k-vecinos más cercanos direccionales, en estimación de densidad por kernel direccional y en el análisis de discriminante direccional con el supuesto de distribución de von Mises-Fisher. Adicionalmente se presenta una extención para utilizar los métodos de clasificación para datos direccionales en conjuntos estándar (no direccionales), basándose en la matriz de correlación entre individuos. Se ilustra el rendimiento de los métodos en conjuntos simulados, en datos de aprendizaje automático y en datos de expresión genética tomadas por medio de microareglos.

DEDICATORIA

A Antonio, Rosa, Angélica, Carlos, Foreman, Lola, Violeta, William, Carolina y Diana, por ser el mejor punto de apoyo y el más fuerte viento que me ha empujado hacia nuevas metas.

AGRADECIMIENTOS

Al doctor Edgar Acuña por sus importantes y oportunas sugerencias sobre el presente documento.

A Karen Prieto por su apoyo incondicional en el transcurso de esta maestría.

A Alejo Torres, Marggie González, Víctor López, Ángel Carreras y José Vélez por ser amigos sinceros y estar dispuestos siempre a brindar un abrazo en los momentos difíciles.

A mis compañeros de la maestría por compartir lo mejor y peor de cada momento.

A la Oficina de Préstamo Interbibliotecario del Recinto Universitario de Mayaguez, por su eficiente labor en la búsqueda de información y documentos importantes para esta tesis.

Al Departamento de Matemáticas por el apoyo permanente durante el transcurso de la Maestría.

A todas las personas quienes sonríen por más de un segundo, se alegran en las mañanas y desean brillar toda su vida.

Tabla de Contenido

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

En algunas áreas de la investigación científica es necesario la aplicación de métodos estadísticos a conjuntos de datos donde algunas de sus variables posee características circulares, es decir, donde el rango de valores que puede tomar la variable es finito y puede ser representado en un círculo. Un claro ejemplo de esto, es el registro de dirección de desplazamiento, ya que indicar una dirección de 1^o o de 359^o es equivalente a decir un grado de desplazamiento con respecto a la dirección 0^o. De una manera un poco más general, se pueden considerar variables que pueden ser representadas sobre la superficie de una hiperesfera de grado p, este tipo de datos son denominados direccionales [Mardia y Jupp, 1999], donde el caso de $p = 2$ son datos circulares y $p = 3$ son datos esféricos. En diversos textos como [Jammalamadaka y SenGupta, 2001] y [Fisher, 1993] se han estudiado las modificaciones adecuadas sobre las estadísticas clásicas cuando se poseen datos direccionales, con desarrollos desde estadísticas básicas hasta teoría estadística en datos circulares. Una recopilación de los conceptos principales se describen en el

capítulo dos.

Por otra parte, en la teoría de clasificación estadística multivariada se busca solucionar principalmente los problemas de formación de conglomerados (*cluster* o clasificación no supervisada) y la construcción de reglas a partir de conjuntos etiquetados (discriminación o clasificación supervisada). Este tipo de problema han sido abordados en múltiples textos y artículos en las décadas recientes. Una importante selección de conceptos y resultados es presentado en el tercer capítulo. Adicionalmente, a partir de la necesidad de estudiar conjunto de datos de gran tamaño, principalmente en estudios genéticos, se ha planteado la necesidad de construir mecanismos de clasificación a conjuntos de datos donde el número de variables (p) sea mucho mayor al número de individuos (n), esto es denominado como datos anchos ($p >> n$). Métodos de clasificación para datos direccionales son analizados en detalle en el capítulo 4.

El aporte científico de esta tesis es precisamente la adaptación de los mecanismos clásicos de clasificación a datos direccionales y su aplicación a conjuntos de cualquier tipo de datos. Buscando la validez práctica se aplicarán los mecanismos de clasificación a datos direccionales simulados, conjuntos de datos estándares y a conjunto de datos anchos.

Capítulo 2

Datos Direccionales

2.1. Motivación

Los datos direccionales en dos y tres dimensiones son frecuentemente encontrados en las ciencias naturales tales como biología, medicina, ecología, geología, etc. Por ejemplo, los biologos estudian los registros de dirección del vuelo durante la migración. En un experimento sobre palomas mensajeras, [Jammalamadaka y SenGupta, 2001] se tienen los registros de los ángulos de dirección de vuelo para 15 aves, y estos son: 85^o, 135^o, 135^o, 140^o, 145^o, 150^o, 150^o, 150^o, 160^o, 185^o, 200^o, 210^o, 220^o, 225^o y 270^o. Los datos son representados en la Figura 2.1. Otros ejemplos de utilización de datos circulares son aplicaciones en medicina, en paleontología [Sengupta y Rao, 1967] y en geología [Fuller et al., 1996].

También los eventos periódicos pueden ser representados sobre un círculo donde la circunferencia corresponde al período, por ejemplo, horas en el día, días de la semana, etc.

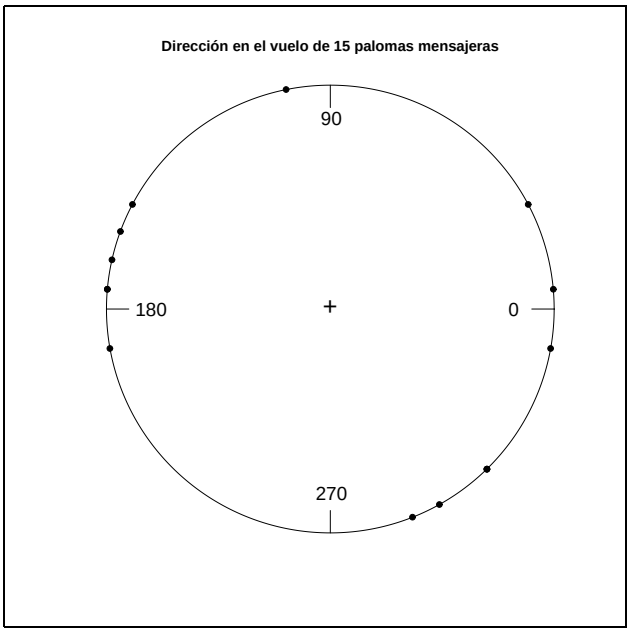

Figura 2.1: Ejemplo de datos circulares

En otro ejemplo, el círculo puede representar los 365 días del año y cada punto representaría la ocurrencia de un accidente aéreo y se desea conocer si la distribución es uniforme en las diferentes épocas del año.

2.2. Estadísticas Descriptivas de datos circulares

Los datos circulares pueden ser representados como ángulos o como puntos sobre una circunferencia. La posición direccional tiene una representación única en un sistema coordenado de dos dimensiones. Es decir, cualquier punto P sobre el plano puede ser representado como (X, Y) en términos de coordenadas rectangulares o como (r, α) en términos de coordenadas polares donde r es la distancia del

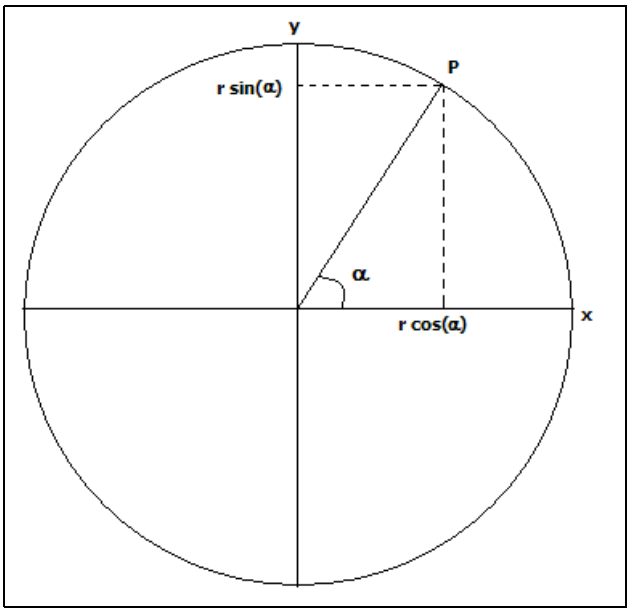

Figura 2.2: Transformada polar

punto al origen y α es la dirección. Para el punto origen, se tiene que $r = 0$ y no tiene dirección, es decir, α no está definida.

La Figura 2.2, representa las coordenadas rectangulares y polares del punto P. Dadas la coordenadas rectangulares de un punto $P = (x, y)$ se tiene que:

$$x = r\cos(\alpha), \quad y = r\sin(\alpha) \tag{2.1}$$

En análisis direccional lo que interesa básicamente es la dirección y no la magnitud del vector, por esta razón usualmente se toma $r = 1$, por conveniencia. Así, cada dirección corresponde a un punto P sobre la circunferencia del círculo unitario. Equivalentemente, cada punto en la circunferencia puede ser representado por un ángulo.

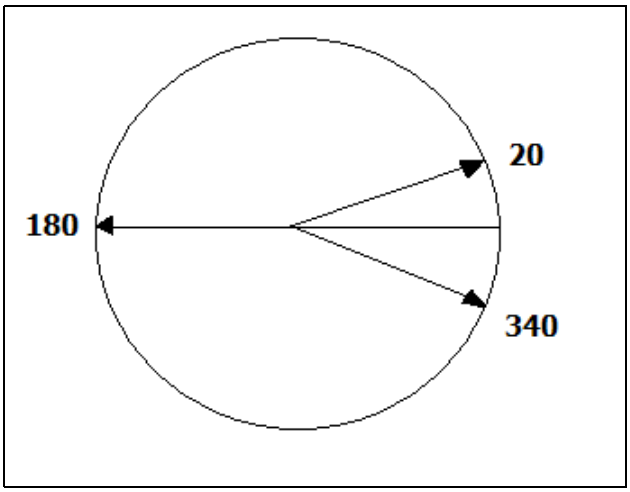

Figura 2.3: La media aritmética es erronea

2.2.1. Medida de Centralidad

Para definir una dirección media, algunas veces llamada la *dirección preferida*, se podría pensar en calcular la media aritmética de los ángulos. Por ejemplo, si se tienen los ángulos 20^o y 340^o, su promedio aritmético sería 180^o, pero como se observa en la Figura 2.3 esto no tiene sentido como registro circular.

Una medida adecuada para la media direccional, en el caso de distribuciones circulares unimodales, es el vector donde se concentran las direcciones. El cómputo de este vector es:

Sea $\alpha_1, \alpha_2, \ldots, \alpha_n$ un conjunto de observaciones circulares dadas en términos de ángulos. Considere la tranformación polar a rectangular de cada observación (ecuación 2.1), es decir $(\cos(\alpha_i), \sin(\alpha_i)), \quad i = 1, \ldots, n$. El vector resultante sería la suma de los n términos sumados componente a componente, es decir:

$$\mathbf{R} = \left(\sum_{i=1}^{n} \cos(\alpha_i), \sum_{i=1}^{n} \sin(\alpha_i) \right) = (C, S) \tag{2.2}$$

Entonces $R = \|\mathbf{R}\| = \sqrt{C^2 + S^2}$ representa la longitud del vector $\mathbf{R}$. La dirección de dicho vector, denotado por $\overline{\alpha}_0$ se obtiene de las siguientes ecuaciones:

$$\cos(\overline{\alpha}_0) = \frac{C}{R}, \qquad \sin(\overline{\alpha}_0) = \frac{S}{R} \tag{2.3}$$

Una definición explícita de $\overline{\alpha}_0$ es dada en [Fisher, 1993] como sigue:

$$\overline{\alpha}_0 = arctan^*(S/C) = \begin{cases} \arctan(S/C) & \text{if} \quad C > 0, S \geq 0, \\ \pi/2 & \text{if} \quad C = 0, S > 0, \\ \arctan(S/C) + \pi & \text{if} \quad C < 0, \\ \arctan(S/C) + 2\pi & \text{if} \quad C \geq 0, S < 0, \\ \text{indefinida} & \text{if} \quad C = 0, S = 0, \end{cases} \tag{2.4}$$

Retomando el ejemplo de la Figura 2.3, se tiene de 2.4 que $C = \cos(20) + \cos(340) = 1{,}879385$ y $C = \sin(20) + \sin(340) = 0$, por tanto, $\overline{\alpha}_0 = \arctan(0) = 0$, siguiendo el sentido natural de una tendencia central(Figura 2.4).

2.2.2. Distancia Circular

En [Jammalamadaka y SenGupta, 2001] se define como una distancia circular adecuada entre dos puntos a la longitud menor de los arcos formados entre los dos puntos en la circunferencia, es decir que para cualquier par de ángulos α y β se tiene que:

$$\rho_0(\alpha, \beta) = \text{mín}(\alpha - \beta, 2\pi - (\alpha - \beta)) = \pi - |\pi - |\alpha - \beta|| \tag{2.5}$$

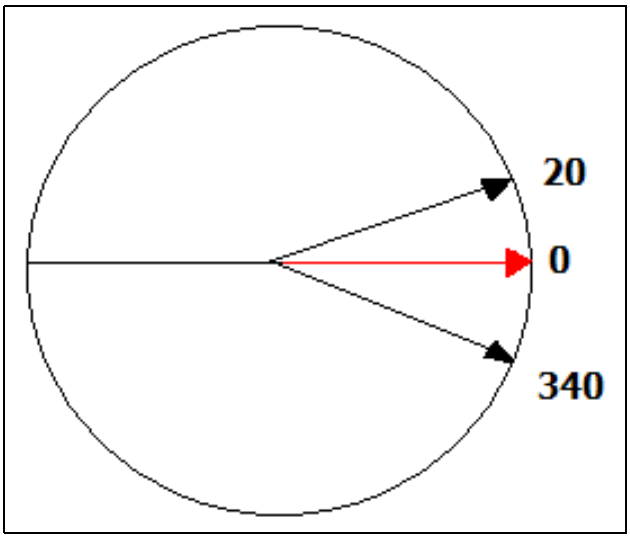

Figura 2.4: La media direccional es adecuada

Por ejemplo en la Figura 2.5, la distancia entre A y B puede ser la longitud del arco ANB o la del arco ASB. Según (2.5), la distancia sería la longitud de arco ANB. Claramente la distancia circular ρ_0 toma valores entre $[0, \pi]$.

En [Jammalamadaka y SenGupta, 2001] se define una segunda distancia circular entre los puntos α y β dada por:

$$\rho_1(\alpha, \beta) = 1 - \cos(\alpha - \beta) \tag{2.6}$$

donde α y β representan los ángulos correspondientes a A y B respectivamente. Si θ es el ángulo entre los puntos A y B, es claro que la función de distancia ρ_1 es monótona creciente con respecto a θ, tomando el valor de 0 cuando $\theta = 0$ y crece hasta 2 si $\theta = \pi$.

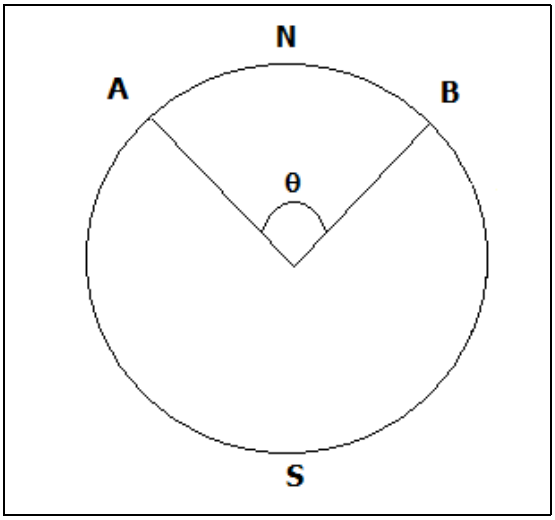

Figura 2.5: La distancia circular ρ_0 es la longitud del arco ANB

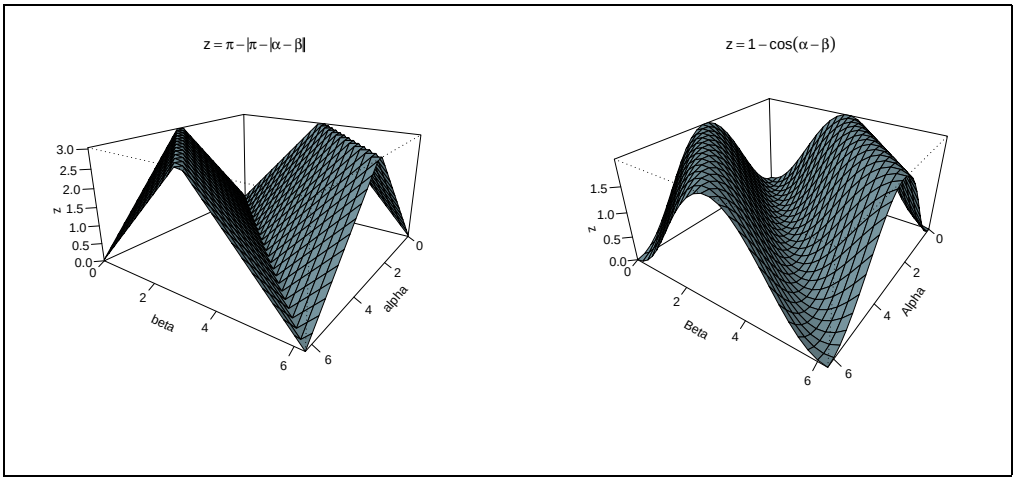

Figura 2.6: Comparación de las medidas de distancias circulares

En la Figura 2.6 se observa, que salvo unidades, ambas distancias tienen el mismo sentido y su diferencia radica en la mayor curvatura que presenta ρ_1.

Es importante, para poder utilizar la distancia como instrumento de decisión, determinar si distancias definidas en (2.5) o (2.6) cumplen las propiedades de medida de disimilaridad. En [Webb, 1999] se indica que una medida ρ entre a y b se dice de disimilaridad si:

- $\rho(a,b) \geq 0 \quad \forall a,b$ (Positiva)
- $\rho(a,a) = 0 \quad \forall a$ (Nulidad)
- $\rho(a,b) = \rho(b,a) \quad \forall a,b$ (Simetría)

Proposición 1 (Disimilaridad Circular de Coseno). *La distancia circular definida en (2.6) por $\rho_1(a,b) = 1 - \cos(a-b)$ es una medida de disimilaridad*

Demostración. La positividad de la distancia se tiene ya que $\cos(a-b)$ está entre $[-1,1]$ por tanto $\rho_1(a,b) \geq 0$. Además, si $\cos(0) = 1$, se tiene que $1 - \cos(0) = 0$, y para cualquier a se tiene que $\rho_1(a) = 1 - \cos(a-a) = 0$. La simetría de la disimilaridad circular se tiene gracias a la paridad de la función coseno. Es decir $\cos(a) = \cos(-a)$ implica directamente $\rho_1(a,b) = 1 - \cos(a-b) = 1 - \cos(b-a) = \rho_1(b,a)$. □

Proposición 2 (Disimilaridad Circular del Valor Absoluto). *La distancia circular definida en (2.5) por $\rho_0(a,b) = \pi - |\pi - |a-b||$ es una medida de disimilaridad.*

Demostración. Dado que el máximo valor que toma la diferencia entre dos ángulos en medidas radianes está entre -2π y 2π se tiene que $|\pi - |a - b||$ toma valores entre $-\pi$ y π por tanto $\pi - |\pi - |a - b||$ tiene como rango $[0, \pi]$ con lo cual se tiene que $\rho_0(a, b) \geq 0$ para todo a y b. La nulidad es obvia y la simetría se obtiene del valor absoluto, ya que $|a - b| = |b - a|$. □

2.3. Distribuciones de Probabilidad Circulares

2.3.1. Introducción

En [Jammalamadaka y SenGupta, 2001] se define una distribución circular como aquella cuya probabilidad total está concentrada sobre la circunferencia de un círculo unitario. El rango de una variable aleatoria (rv) circular θ, medida en radianes, toma valores entre $[0, 2\pi)$ o $[-\pi, \pi)$. De la misma manera que las distribuciones clásicas de probabilidad, las distribuciones circulares son escencialmente de dos tipos: Discretas y Continuas. En cualquier caso, una función de densidad de probabilidad (pdf) $f(\theta)$ debe tener las siguientes propiedades:

- $f(\theta) \geq 0$;
- $\int_0^{2\pi} f(\theta) d\theta = 1$
- $f(\theta) = f(\theta + k \cdot 2\pi)$ para cualquier entero k (es decir f es periódica)

2.3.2. Algunos Métodos para obtener Distribuciones Circulares

En la secciones anteriores se ha mostrado que una variable aleatoria circular puede ser representada en términos del ángulo $\theta, (0 \leq \theta \leq 2\pi)$ o como un vector unitario bidimensional $(X = \cos\theta, Y = \sin\theta)'$.

Algunas distribuciones circulares pueden ser generadas a partir de distribuciones de probabilidad conocidas sobre la recta real o sobre el plano, por medio de diferentes mecanismos. A continuación se describen algunos de ellos:

(1) Por envoltura ("wrapping") de una distribución lineal alrededor del círculo unitario;

(2) Por medio de propiedades características tal como maximizar la entropía;

(3) Transformando una variable aleatoria lineal bivariada a sus componentes direccionales, las cuales son llamadas distribuciones de desplazamiento ("offset");

(4) Iniciando con una distribución sobre la recta real $\mathbb{R}$, se aplica un proyección estereográfica que identifica cada punto x en $\mathbb{R}$ con algún punto sobre la circunferencia del círculo unitario, llamado θ. Esta correspondencia es uno a uno, excepto por el hecho que los valores $-\infty$ y ∞ son identificados con π.

2.3.3. Distribución Uniforme

Cuando el total de la probabilidad es extendida uniformemente sobre la circunferencia se obtiene la distribución uniforme circular con función de densidad constante:

$$f(\theta) = \frac{1}{2\pi} \qquad 0 \leq \theta < 2\pi \tag{2.7}$$

La distribución uniforme circular tiene un papel principal en el análisis de datos circulares porque representa el estado de ausencia de dirección promedio o ausencia de dirección preferida. Cuando un conjunto de datos circulares no se ajusta a una distribución uniforme se piensa entonces en la presencia de una o más direcciones preferidas.

2.3.4. Distribucion de von Mises

Una variable aleatoria θ se dice que sigue una distribución de von Mises o Normal Circular si tiene como función de densidad:

$$f(\theta; \mu, \kappa) = \frac{1}{2\pi I_0(\kappa)} e^{\kappa \cos(\theta - \mu)}, \qquad 0 \leq \theta \leq 2\pi, \tag{2.8}$$

donde $0 \leq \mu < 2\pi$ y $\kappa \geq 0$ son parámetros. El término $I_0(\kappa)$ de la constante de normalización es la función modificada de Bessel de primera clase de orden cero (para detalles de la misma ver [Fisher, 1993]) y está dada por:

$$I_0(\kappa) = \frac{1}{2\pi} \int_0^{2\pi} \exp(\kappa \cos\theta) d\theta = \sum_{r=0}^{\infty} \left(\frac{\kappa}{2}\right)^{2r} \left(\frac{1}{r!}\right)^2 \tag{2.9}$$

Esta distribución es discutida originalmente por Langevin en 1905 en el contexto de la física y luego en 1916 utilizadas como modelo estadístico por von Mises.

La función de densidad de von Mises tiene las siguientes propiedades:

1. Simetría: Debido a la simetría de la función coseno, la densidad es simétrica alrededor de la dirección μ.

2. Moda en μ: Dado que la función coseno tiene máximo en cero, la densidad de von Mises tiene máximo en $\theta = \mu$, es decir que μ es moda direccional cuyo valor máximo es:

$$f(\mu) = \frac{e^{\kappa}}{2\pi I_0(\kappa)}. \tag{2.10}$$

3. Antimoda en $\mu \pm \pi$: Ya que $\cos \pi = -1$ es el valor mínimo, entonces $\theta = \mu \pm \pi$, da la densidad mínima en:

$$f(\mu \pm \pi) = \frac{e^{-\kappa}}{2\pi I_0(\kappa)}. \tag{2.11}$$

Así, $\mu \pm \pi$ es la dirección antimodal.

4. Parámetro de concentración κ: De (2.10) y (2.11), se tiene que:

$$\frac{f(\mu)}{f(\mu \pm \pi)} = e^{2\kappa}.$$

Así que, medida que κ aumenta, la razón de $f(\mu)$ y $f(\mu \pm \pi)$ es más grande; indicando mayor concentración alrededor de la media direccional poblacional μ. Por tal motivo, κ es conocido como el parámetro de concentración alrededor de la media direccional.

En la Figura 2.7 se muestra la función de densidad de von Mises para diferentes valores de κ y para $\mu = \pi$.

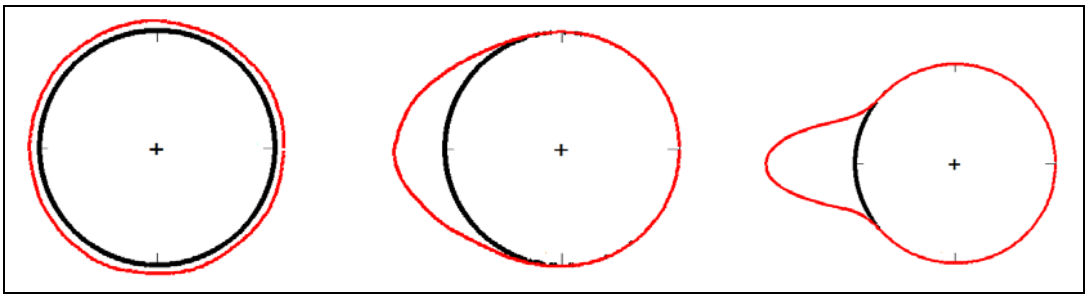

Figura 2.7: Densidades de von Mises con $(\kappa = 1/100, 5, 25)$, $(\mu = \pi)$ en los tres casos

2.4. Datos Direccionales Multidimensionales

2.4.1. Introducción

La generalización de datos circulares sobre la hiperesfera unitaria de dimensión p, S_p, es conocido por datos direccionales.

Usualmente cuando se tiene un conjuntos de datos es de interés estudiar la dirección y la magnitud del vector $x = (x_1, \ldots, x_p)^T$, pero en algunas ocaciones es deseable estudiar únicamente en la dirección de x. Entonces, dichos puntos pueden ser ubicados en la circuferencia del círculo en dos dimensiones o en la superficie de la esfera en tres dimensiones. En general, las direcciones pueden ser visualizadas como puntos en la superficie de la hiperesfera.

Se denota al vector direccional aleatorio p-dimensional por θ, donde $\theta^T\theta = 1$. El vector unitario θ toma valores sobre la superficie de la hiperesfera $p-1$-dimensional S_{p-1}, que tiene radio unitario y centro en el origen.

En general, es conveniente considerar la densidad de x en términos de las coorde-

nadas polares esféricas $x = ru(\theta)$ con $\theta = (\theta_1, \ldots, \theta_{p-1})^T$ donde:

$$u_i(\theta) = \cos\theta_i \prod_{j=0}^{i-1} \sin\theta_j, \qquad i = 1, \ldots, p, \qquad \sin\theta_0 = \cos\theta_p = 1 \tag{2.12}$$

y

$$0 \leq \theta_j \leq \pi, \qquad 0 \leq \theta_{p-1} < 2\pi, \qquad r > 0$$

El jacobiano de la transformación desde (r, θ) a x está dado en [Mardia y Jupp, 1999] y es:

$$J_p = r^{p-1} \prod_{i=1}^{p-1} sin^{p-i}\theta_{i-1}, \qquad J_2 = r \tag{2.13}$$

Usualmente se utiliza esta transformación sobre la hiperesfera de radio uno. Es importante notar que para $p = 2$, se tiene que es igual a (2.1).

2.4.2. Distribución de von Mises-Fisher

[Jammalamadaka y SenGupta, 2001] sugiere como una generalización adecuada de la distribución de von Mises (2.8) sobre esferas $p-1$ dimensionales, a distribuciones cuya log-densidad sea lineal en x, es decir, cuya densidad $f(x; \mu, \kappa)$ satisfacen:

$$\log f(x; \mu, \kappa) = \kappa \mu^T x \ + \ \text{constante}, \tag{2.14}$$

éstas son llamadas las distribuciones von Mises-Fisher.

Un vector aleatorio x sigue una distribución $(p-1)$ dimensional von Mises Fisher (o Langevin en la terminología de [Watson, 1974]), si la función de densidad de probabilidad es:

$$f(x;\mu,\kappa) = \left(\frac{\kappa}{2}\right)^{p/2-1} \frac{1}{\Gamma(p/2) I_{p/2-1}(\kappa)} \exp\{\kappa \mu^T x\}, \tag{2.15}$$

donde $\kappa \geq 0, \|\mu\| = 1$, y I_v denota la función modificada de Bessel de primer tipo de orden v. Los parámetros μ y κ son llamados *media direccional* y *parámetro de concentración*, respectivamente.

En el caso particular de p=2, de (2.15) se obtiene:

$$f(x;\mu,\kappa) = \frac{1}{I_0(\kappa)} e^{\kappa \mu^T u(x)}$$

y aplicando la transformada polar (2.12) a x:

$$u(\theta) = \begin{pmatrix} u_1(\theta_1) \\ u_2(\theta_1) \end{pmatrix} = \begin{pmatrix} \cos\theta_1 \\ \sin\theta_1 \end{pmatrix}$$

se obtiene

$$\begin{aligned} f(\theta;\mu,\kappa) &= \frac{1}{I_0(\kappa)} e^{\kappa(\cos\mu_1 \cos\theta_1 + \sin\mu_1 \sin\theta_1)} \\ &= \frac{1}{I_0(\kappa)} e^{\kappa \cos(\theta_1 - \mu_1)} \end{aligned}$$

la cual tiene el mismo comportamiento de la distribución circular de von-Mises (2.8), a excepción del parámetro de normalización $1/(2\pi)$.

Para el caso de p=3, la distribución de von Mises-Fisher es llamada la distribución Fisher, y está dada por la expresión:

$$f(x;\mu,\kappa) = \frac{\kappa}{2\sinh\kappa} e^{\mu^T x} \tag{2.16}$$

Escribiendo a x y μ en coordenadas polares esféricas:

$x = (\cos(\theta), \sin(\theta)\cos(\phi), \sin(\theta)\sin(\phi))^T, \mu = (\cos(\alpha), \sin(\alpha)\cos(\beta), \sin(\alpha)\sin(\beta))^T,$

se obtiene:

$$f(x;\mu,\kappa) = \frac{\kappa}{4\pi \sinh \kappa} e^{\kappa[\cos\theta\cos\alpha + \sin\theta\sin\alpha\cos(\phi-\beta)]}\sin\theta \tag{2.17}$$

Capítulo 3

Métodos de Clasificación

3.1. Formulación

El problema de clasificación de patrones es formulado en [Kulkarni et al., 1998] y [Devroye et al., 1996] de la siguiente manera. Hay J clases de objetos (estados en la naturaleza) de interés, los cuales utilizarán el subíndice j con $j = 1, \ldots, J$, para cada estado respectivamente. La información que poseemos sobre los objetos es resumida en un número finito, p, de medidas de valor real denominadas *características* ("features"). Todas juntas forman un *vector de características* $x \in R^p$. En el modelo de incertidumbre que analizaremos a continuación se supone que hay probabilidades *a priori* $\Pi_1, \Pi_2, \ldots, \Pi_J$ para cada una de las clases. Para modelar la relación entre el vector de características (incluyendo el ruido natural en los procesos de medición y ocasionales de la naturaleza misma), se asume que un objeto de la clase $y \in \{1, 2, \ldots, J\}$ es una realización del vector de variables aleatorias con función de distribución condicional a la clase $F_y(x)$. Vectores de características aleatorias X (los observados en el proceso) son generados de acuerdo al siguiente

proceso de J estados: la clase aleatoria $Y \in \{1, 2, \ldots, J\}$ es seleccionada de acuerdo a las probabilidades a priori $\{\Pi_1, \Pi_2, \ldots, \Pi_J\}$; el vector de características observadas X, es una selección de la distribución condicional a la clase, F_Y. Dada una realización del vector de características X, denotado por x, el problema al que se enfrenta el clasificador es decidir a cuál clase pertence el objeto x. Así, un clasificador o regla de decisión, es simplemente una función $g : R^p \to \{1, 2, \ldots, J\}$, donde $g(x)$ denota la clase asignada al vector de características x por el clasificador g. Dado un clasificador g, el rendimiento de g puede ser medido por la probabilidad de error, dada por:

$$L(g) = P\{g(x) \neq Y\}. \tag{3.1}$$

Si las probabilidades a priori y las distribuciones condicionales son conocidas, entonces el problema de clasificar se convierte en un problema de regla óptima de decisión en el sentido minimizar la probabilidad de error, o más general en minimizar el riesgo si se asignan diferentes costos a los diferentes tipos de error. La regla que minimiza esta probabilidad de error es denominada Regla de Decisión de Bayes, denotada por g^*. Esta regla de decisión usa las distribuciones conocidas

y la observación $X = x$ para calcular las probabilidades a posteriori

$$\eta_1(x) = P\{Y = 1|X = x\}$$

$$\eta_2(x) = P\{Y = 2|X = x\}$$

$$\vdots$$

$$\eta_J(x) = P\{Y = J|X = x\}$$

de las J clases, y se selecciona la clase con mayor probabilidad a posteriori (equivalentemente menor riesgo), es decir,

$$g^*(x) = \underset{j\in\{1,2,\ldots,J\}}{\text{argmax}}\ \eta_j(x). \tag{3.2}$$

La tasa de error de Bayes, denotada L^*, es dada por:

$$L^* = L(g^*) = E[min\{\eta_1(X), \eta_2(X), \ldots, \eta_J(X)\}]. \tag{3.3}$$

Desde luego, en la mayoría de las aplicaciones se desconoce por lo menos parcialmente las distribuciones condicionales, o no se desea realizar supuestos sobre ellas. En este caso, generalmente se asume que tenemos realizaciones previas, denominada *muestra de entrenamiento* del vector de características X para las diferentes clases. Es decir, se tiene

$$D_n = (X_1, Y_1), \ldots, (X_n, Y_n), \tag{3.4}$$

donde $Y_k \in \{1, 2, \ldots, J\}$ corresponde a la clase de los objetos, la cual es asumida idéntica e independientemente distribuida con $\{\Pi_1, \Pi_2, \ldots, \Pi_J\}$ para cada una de las clases, y X_k es un vector de características proveniente de la distribución de clase condicional $F_{Y_k}(x)$. Así, la pareja (X_k, Y_k) es asumida idéntica e independientemente distribuida de acuerdo a las distribuciones (desconocidas) P_y y $F_y(x)$, las cuales caracterizan el problema. Intuitivamente, los datos D_n brindan información parcial de las distribuciones desconocidas, y es de interés usar dichos datos para encontrar buenos clasificadores. Más formalmente, una regla de clasificación o *clasificador* es una función $g_n(x) = g_n(x, D_n)$, construida a partir del conjunto o muestra de entrenamiento D_n, la cual asigna una etiqueta $(1, 2, \ldots, J)$ a cada punto $X \in R^p$. Para un conjunto de datos de entrenamiento fijo D_n, la probabilidad condicional del error de un clasificador es:

$$L(g_n) = P\{g_n(X) \neq Y | D_n\} \tag{3.5}$$

donde la pareja (X, Y) es independiente de D_n. Es importante notar que este error depende de la muestra de entrenamiento, por lo tanto, la probabilidad de error $L(g_n)$ es una variable aleatoria que depende de D_n. Luego, se define la probabilidad esperada de error $\overline{L(g_n)} = E[L(g_n)] = P\{g_n(X) \neq Y\}$ tomando este valor esperado con respecto a la muestra de entrenamiento aleatoria. Teóricamente, cuando se desea evaluar el rendimiento se usa como cota la tasa de error de Bayes L^* la cual es óptima cuando se considera completamente conocidas las distribuciones.

3.2. Estimación de Densidad Noparamétrica

3.2.1. Introducción

A primera vista, se puede pensar que diseñar buenas reglas de clasificación es asumir correctamente distribuciones para $F_y(x)$ en cada clase. Estos métodos paramétricos han sido ampliamente estudiados, los cuales solucionan el problema al suponer distribuciones ampliamente conocidas (Sección 3.3) y que se ajusten a la naturaleza de los problemas particulares. Desafortunadamente en general, es difícil sustentar los supuestos paramétricos. Así, el estudio de métodos no paramétricos y reglas consistentes universalmente han recibido gran atención. El problema se convierte entonces en realizar estimaciones eficientes de las densidad $F_y(x)$ para cada clase, a partir de la muestra de entrenamiento D_n. A continuación se presenta la estimación de densidad por kernel y la estimación por k-vecinos más cercanos.

3.2.2. Estimación de Densidad por Kernel

El método de estimación de densidad más básico es el histograma, en el cual la funcion de densidad se construye a partir de los datos de la muestra: el espacio R^p es particionado en K p-celdas, cada una de ellas con n_j observaciones con $j = 1, 2, \ldots, K$ y la estimación de la densidad en x se halla por:

$$\widehat{f}(x) = \frac{n_j}{\sum_{j=1}^{k} n_j dV} \tag{3.6}$$

donde n_j es el número de observaciones en la j-ésima celda y dV es el volumen

de la j-ésima celda.

A pesar de ser un concepto sencillo y fácil de implementar, se presentan problemas prácticos principalmente en dimensiones altas (cuando p es grande). Si suponemos que tenemos una muestra de tamaño n, de vectores p dimensionales, se tendría n^p celdas en las que pueden localizarse la observaciones, por lo tanto en dimensiones altas es necesario una gran cantidad de observaciones para evitar zonas cero amplias en la estimación de la densidad (en las cuales no se tendría certeza en la clasificación). Un problema adicional es que la estimación de densidad construida por el histograma es discontinua en un número finito de puntos, y toma el valor cero en las fronteras de la region de manera abrupta. Los métodos de estimación de densidad por kernel buscan solucionar el problema de discontinuidad.

El método de kernel [Webb, 1999](también denominado como método de Parzen), fija el volumen de la celda, encuentra el número de muestras en la celda y usa esto para estimar la densidad. Inicialmente para el caso univariado, sea $x_1, x_2, \ldots, x_n$ el conjunto de observaciones univariadas que serán usadas para la estimación de la densidad. Para este conjunto la estimación de la función de distribución empírica es:

$$\widehat{F}(x) = \frac{n\acute{u}mero \quad de \quad observaciones \leq x}{n} \tag{3.7}$$

La función de densidad, $\widehat{f}(x)$ es la derivada de la función de distribución $\widehat{F}(x)$, pero en este caso es discontinua y su derivada resulta ser un conjunto de picos en

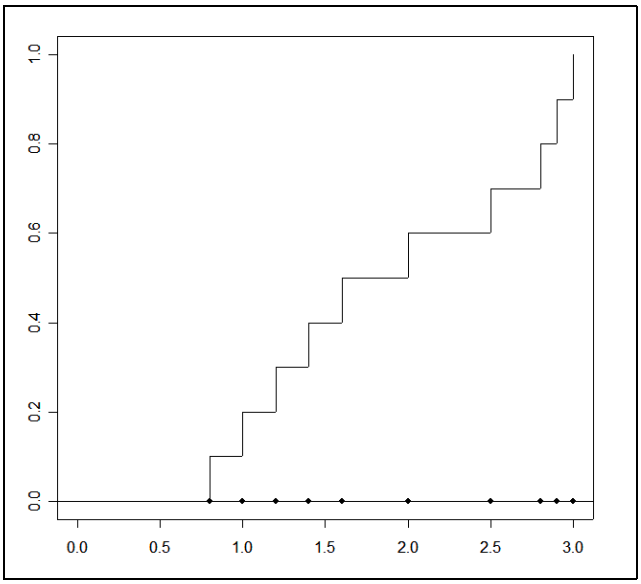

Figura 3.1: Función de distribución empírica

los punto de la muestra, x_i, y cero en el resto.

Sin embargo, se puede definir la estimación de densidad como:

$$\widehat{f}(x) = \frac{\widehat{F}(x+h) - \widehat{F}(x-h)}{2h} \tag{3.8}$$

con $h > 0$ y relativamente pequeño. De este modo, $\widehat{f}(x)$ es la proporción de observaciones que caen en un intervalo $(x-h, x+h)$ dividido por $2h$. Reescribiendo se tiene:

$$\widehat{f}(x) = \frac{1}{hn} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{3.9}$$

donde

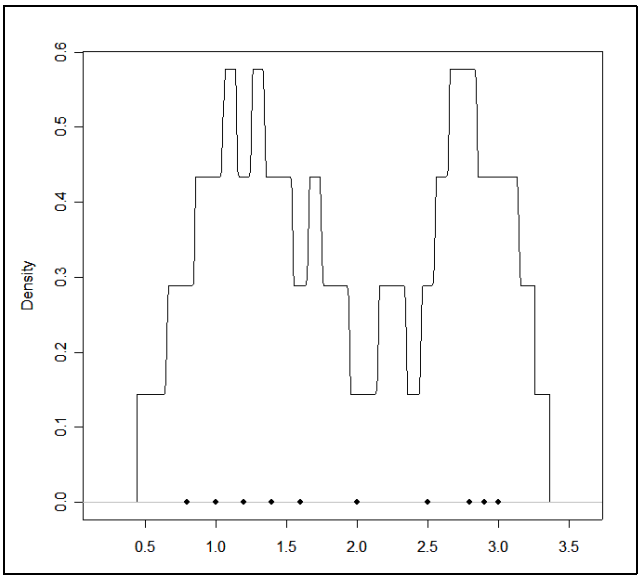

Figura 3.2: Densidad de probabilidad estimada por kernel rectangular

$$K(z) = \begin{cases} 0 & |z| > 1, \\ 1/2 & |z| \leq 1 \end{cases}$$

así, para los puntos muestrales x_i, dentro del intervalo con centro en x y radio h, la sumatoria (3.9) da un valor de $\frac{1}{2}$(Número de observaciones en el intervalo). Así cada punto dentro del intervalo contribuye igualmente a la sumatoria. La función $K(z)$ definida anteriormente es denominada, kernel rectangular.

La Figura 3.2 muestra que la estimación de densidad dada por (3.9) es discontinua. Esto se debe al hecho que los puntos a distancia menor o igual a h del punto x contribuyen un valor de $\frac{1}{2hn}$ a la densidad y los demás contribuyen con valor de cero. Es precísamente el salto entre $\frac{1}{2hn}$ y cero lo que causa esta discontinuidad. Se

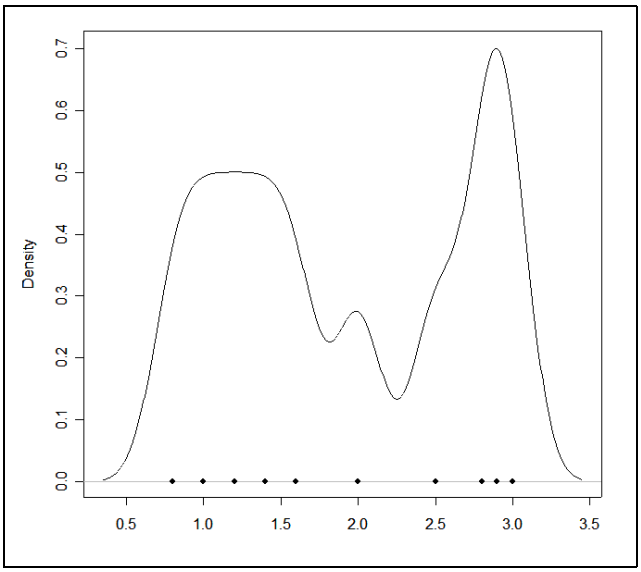

Figura 3.3: Densidad de probabilidad estimada por kernel gausiano

puede solucionar este problema y generalizar los estimadores usando otra función de pesos $K(z)$. Por ejemplo, se puede pensar en una ponderación por una función del tipo gausiano $K_1(z)$ (también con la propiedad que la integral sobre la recta real sea la unidad). La Figura 3.3, muestra la estimación de densidad utilizando la función de ponderación dada por:

$$K_1(z) = \frac{1}{\sqrt{(2\pi)}} \exp\left(\frac{-z^2}{2}\right)$$

para un valor particular de h=0.2.

Al comparar las estimaciones anteriores, Figuras 3.2 y 3.3, se observa que las funciones $K()$ realizan un ajuste sobre los datos originales. Por esta razón, surge la formulación general.

Función de Kernel	$K(x)$
Rectangular	$\frac{1}{2}$ para $\|x\| < 1$, 0 en otro caso
Triangular	$1 - \|x\|$ para $\|x\| < 1$, 0 en otro caso
Biweight	$\frac{15}{16}(1-x^2)^2$ para $\|x\| < 1$, 0 en otro caso
Normal	$\frac{1}{\sqrt{2\pi}}e^{-x^2/2}$
Epanechnikov	$\frac{3}{4}(1-x^2/5)/\sqrt{5}$ para $\|x\| < \sqrt{5}$, 0 en otro caso

Tabla 3.1: Funciones Kernel

Dado un conjunto de observaciones $x_1, x_2, \ldots, x_n$, una estimación de una función de densidad, unidimensional, es dada por:

$$\widehat{p}(t) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{t-x_i}{h}\right) \tag{3.10}$$

donde $K(z)$ es denominada *función de kernel* y h es el *parámetro de suavizamiento* (algunas veces llamado *ancho de banda*). Las funciones de kernel univariadas más populares se encuentran en la tabla 3.1.

Los kernel multivariados (p variables), son usualmente extenciones de kernel univariados simétricos tales como el kernel gausiano,

$$K(x) = (2\pi)^{-p/2}\exp\{-x^T x/2\}$$

y el kernel de Epanechnikov

$$K(x) = (1-x^T x)(p+2)/(2C_p) \quad \text{si} \quad |x| < 1 \quad \text{y} \quad 0 \quad \text{en otro caso}$$

donde $c_p = \pi^{p/2}/\gamma((p/2)+1)$ es el volumen de la esfera unitaria p dimensional.

Si se imponen las condiciones $K(z) \geq 0$ y $\int_{R^p} K(z)dz = 1$, entonces la función de densidad estimada satisface las condiciones de función de densidad de probabilidad [Webb, 1999].

La extensión de la estimación para una función de densidad multivariada f en $t = (t_1, \ldots, t_p)$, dada una muestra $X = (X_1, X_2, \ldots, X_n)$, con $X_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ por el método de kernel está dada por:

$$\widehat{f}(t) = \frac{1}{nh_1h_2 \ldots h_p} \sum_{i=1}^{n} K_p \left(\frac{t_1 - x_{i1}}{h_1}, \frac{t_2 - x_{i2}}{h_2}, \ldots, \frac{t_p - x_{ip}}{h_p} \right) \tag{3.11}$$

donde $h' = (h_1, h_2, \ldots, h_p)$ es el vector de anchos de banda y $K_p(x)$ es la función de densidad multivariada definida para x p-dimensional y

$$\int_{\mathbb{R}^p} K(x)dx = 1$$

Otra manera usual de estimar la función de densidad de probabilidad es usando la suma de *kernels producto*, es decir:

$$\widehat{f}(t) = \frac{1}{n} \frac{1}{h_1 \cdots h_p} \sum_{i=1}^{n} \prod_{j=1}^{p} K \left(\frac{(t_j - x_{ij})}{h_j} \right)$$

donde hay diferentes parámetros de suavizamientos asociados con cada variable. La $K_j(x)$ puede ser cualquier kernel univariado de la tabla 3.1.

En términos más generales, la función kernel que estima a x, basada en una muestra aleatoria de x_i con $i = 1, \ldots, n$, denominada $\mathbf{X}$ puede ser representada de la siguiente forma [Friedman, 2003]:

$$K(x, \mathbf{X}) = g(d(x, \mathbf{X})/h) \tag{3.12}$$

donde $d(x, \mathbf{X})$ es un "distancia"definida entre x y $\mathbf{X}$, h es un parámetro de escala (suavizamiento), y $g(z)$ es un función usualmente simétrica y que disminuye a medida que $|z|$ aumenta.

3.2.3. Estimación de Densidad por K-Vecinos más Cercanos

El método de K-NN (K Nearest Neighbor) [Fix y Hodges, 1951], es otro método de estimación noparamétrica. Sea $x_1, x_2, \ldots, x_n$ una muestra con función de densidad desconocida $f(x)$. Se estima $f(x)$ a partir de una celda de centro en x y que crece hasta capturar k elementos, donde k es definido arbitrariamente o como función de n. Estas muestras son las k vecinos más cercanos a x. Se tiene entonces:

$$\widehat{f}(x) = \frac{k/n}{V_k(x)}, \tag{3.13}$$

donde $V_k(x)$ es el volumen de un elipsoide centrado en x y de radio la distancia de x al k-ésimo vecino más cercano.

3.3. Clasificación por el método de K vecinos más cercanos (KNN)

Si la función de clase condicional $F_{y_k}(x)$ es estimada por 3.13 para cada una de las clases, la regla de decisión toma un aspecto más simple, clasificar a la clase

i tal que:

$$\frac{k_i P_i/n_i}{V_k(x)} > \frac{k_j P_j/n_j}{V_k(x)} \quad \forall j \neq i \tag{3.14}$$

Usualmente las probabilidades a priori P_i son estimadas como la proporción de n_i/n por tanto (3.14) se reduce a:

$$k_i > k_j \quad \forall j \neq i \tag{3.15}$$

Es decir $g_{KNN}(x)$ asigna x a la clase mayoría entre los k-vecinos más cercanos. En caso de empate la asignación es aleatoria.

3.4. Análisis Discriminante

El problema de clasificación paramétrica es formulado por diversos autores, por ejemplo [Mardia y Jupp, 1999] y [Anderson, 1984] de manera similar al siguiente desarrollo. Considere J poblaciones o grupos $\Pi_1, \ldots, \Pi_J$ con $J \geq 2$. Suponga que asociada a cada población Π_i hay una densidad de probabilidad $f_i(x)$ en R^p. Entonces el objetivo del análisis de discriminante es clasificar cada individuo x en uno de los g grupos, basándose en las mediciones de x.

En ciertas situaciones se tiene que las poblaciones poseen probabilidades a prior. Esto es información que se tiene de antemano, la cual da una idea del resultado global de la clasificación. Esta información debe incorporarse en el análisis por

medio de la denominada *regla de discriminante bayesiano* [Mardia y Jupp, 1999]. Dichas probabilidad a priori, denotadas por π_i, son estríctamente positivas para $j = 1, ..., J$.

3.4.1. Discriminante Bayesiano

[Mardia y Jupp, 1999] define la regla de discriminación bayesiana por: Si las poblaciones $\Pi_1, \ldots, \Pi_J$ tienen probabilidades a priori $(\pi_1, \ldots, \pi_J) = \pi$, entonces la regla de discriminación bayesiana (con respecto a π) asigna una observación x a la población en la cual:

$$\pi_j L_j(x) \tag{3.16}$$

es máxima.

Donde $L_j(x) = f_{y_j}(x)$ es la j-ésima función de densidad de probabilidad. Se escribe de esta forma para enfatizar que se está pensando en la verosimilitud del registro x como función del parámetro j.

El resultados clásico de clasificar el individuo x a la población Π_i se obtiene al suponer distribuciones normales multivariadas en cada clase, con igualdad en las matrices de covarianza Σ.

Capítulo 4

Métodos de clasificación en datos direccionales

Un problema específico en datos direccionales es la clasificación. Ha sido tratado anteriormente solamente en [Morris y Laycock, 1974] y [Ackermann, 1997]. En [Morris y Laycock, 1974] se presenta la función de discriminación cuando los datos provienen de distribuciones direccionales univariadas (circulares) y bivariadas (esféricas); adicionalmente calcula expresiones para el error de clasificación teórico. [Ackermann, 1997] analiza también el discriminante circular e introduce una versión de algoritmos de "cluster" basados en distancias circulares.

Un ejemplo concreto de clasificación para datos circulares es presentado a continuación. Suponga que un biólogo registra la dirección de tortugas luego de dejar el cascarón. Adicionalmente se conoce la especie (Clase). La Figura 4.1 es la representación de estos datos.

Es importante hacer notar que el problema de clasificación puede ser resuelto estadísticamente desde dos puntos de vista: paramétricamente, suponiendo alguna

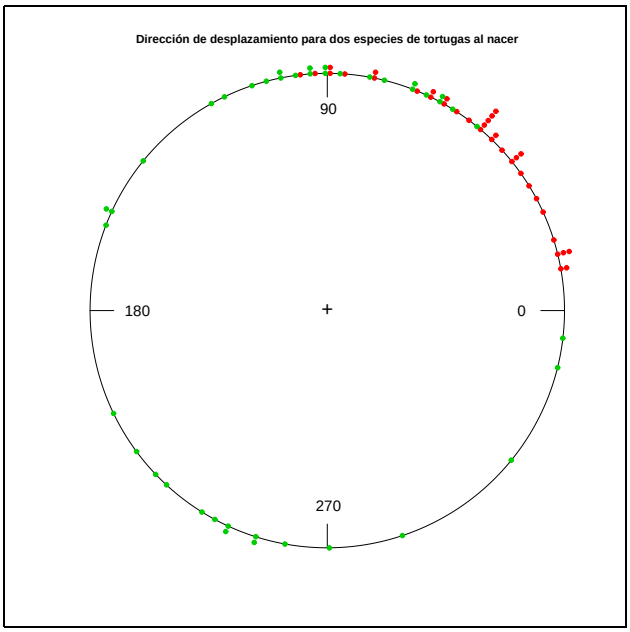

Figura 4.1: Ejemplo de datos circulares etiquetados

distribución conocida para la función de densidad de probabilidad de cada clase (Sección 4.1) o noparamétricamente al estimar la función de densidad de probabilidad de cada clase (Sección 4.2).

4.1. Clasificación por Discriminante direccional

De manera análoga al desarrollo del discriminante en espacios euclideanos, el discriminante direccional supone la distribución de von Mises-Fisher $p-1$ dimensional. Es decir, suponga que el vector direccional x sigue la distribución de von Mises-Fisher (2.15). Se tiene entonces que si las poblaciones $\Pi_1, \ldots, \Pi_J$ tiene probabilidades a priori $(\pi_1, \ldots, \pi_J) = \pi$, entonces la regla de discriminación bayesiana (con respecto a π) asigna una observación x a la población en la cual $\pi_j L_j(x)$ es

máxima. Donde

$$L_j(x) = f_j(x) = \left(\frac{\kappa_j}{2}\right)^{p/2-1} \frac{1}{\Gamma(p/2) I_{p/2-1}(\kappa_j)} \exp\{\kappa_j \mu_j^T x\} \tag{4.1}$$

y κ_j y μ_j son parámetros para cada clase j con $j = 1, \ldots, J$.

Proposición 3 (Función de Discriminante Direccional). *Bajo el supuesto de distribución de von Mises-Fisher en cada clase, la regla de clasificación η para un registro x es dada por:*

$$\eta(x) = \underset{j}{\operatorname{argmax}} \left\{ \pi_j \frac{\kappa_j^{p/2-1}}{I_{p/2-1}(\kappa_j)} \exp\{\kappa_j \sum_{i=1}^{p} (\cos(\mu_{ij}) \cos(\theta_i) \prod_{m=0}^{i-1} \sin \mu_{mj} \sin \theta_m)\} \right\} \tag{4.2}$$

Demostración. La regla es clasificar a la clase que maximice (4.1):

$$\begin{aligned}
\eta(x) &= \underset{j}{\operatorname{argmax}} \{\pi_j f_j(x)\} \\
&= \underset{j}{\operatorname{argmax}} \left\{ \pi_j \left(\frac{\kappa_j}{2}\right)^{p/2-1} \frac{1}{\Gamma(p/2) I_{p/2-1}(\kappa_j)} \exp\{\kappa_j \mu_j^T x\} \right\} \\
&= \underset{j}{\operatorname{argmax}} \left\{ \pi_j \frac{\kappa_j^{p/2-1}}{I_{p/2-1}(\kappa_j)} \exp\{\kappa_j \mu_j^T x\} \right\} \\
&= \underset{j}{\operatorname{argmax}} \left\{ \pi_j \frac{\kappa_j^{p/2-1}}{I_{p/2-1}(\kappa_j)} \exp\{\kappa_j u(\mu_j)^T u(x)\} \right\} \\
&= \underset{j}{\operatorname{argmax}} \left\{ \pi_j \frac{\kappa_j^{p/2-1}}{I_{p/2-1}(\kappa_j)} \exp\{\kappa_j \sum_{i=1}^{p} (\cos(\mu_{ij}) \cos(\theta_i) \prod_{m=0}^{i-1} \sin \mu_{mj} \sin \theta_m)\} \right\}
\end{aligned}$$

donde μ_{mj} y θ_m denotan la m-ésima coordenada de la transformada polar para la media direccional de la clase j y para el vector x respectivamente. □

Ahora, suponiendo inicialmente igualdad en las probabilidades a priori π_j y del parámetro de centralidad κ_j en cada clase (homocedasticidad), se tiene:

$$\eta(x) = \underset{j}{\operatorname{argmax}} \left\{ \sum_{i=1}^{p} (\cos(\mu_{ij}) \cos(\theta_{ij})) \prod_{m=0}^{i-1} \sin \mu_{mj} \sin \theta_{mj} \right\} \tag{4.3}$$

denominada *función de discriminación direccional* en el caso de igualdad de parámetros de dispersión.

Proposición 4 (Discriminante Circular con igualdad de κ_j y Π_j). *La regla de discriminación suponiendo distribución de von Mises-Fisher circular con igual parámetro de concentración κ y probabilidades a priori cada clase es:*

$$\eta(x) = \underset{j}{\operatorname{argmax}} \left(\cos(\theta - \mu_j)\right) \tag{4.4}$$

Demostración. Para el caso circular, $p = 2$, utilizando el hecho que en la transformación polar $u(\alpha)$ (2.12), $\sin(\alpha_0) = \cos(\alpha_p) = 1$ se tiene:

$$\begin{aligned} \eta(x) &= \underset{j}{\operatorname{argmax}} \left[\exp\{(\cos(\mu_j) \cos(\theta) + \sin(\mu_j) \sin(\theta))\}\right] \\ &= \underset{j}{\operatorname{argmax}} \left[\exp\{(\cos(\theta - \mu_j))\}\right] \\ &= \underset{j}{\operatorname{argmax}} \left(\cos(\theta - \mu_j)\right) \end{aligned}$$

□

Esta función de discriminación es máxima para la clase j donde θ esté más cercano a μ_j circularmente. Un esquema de esta clasificación se observa en la

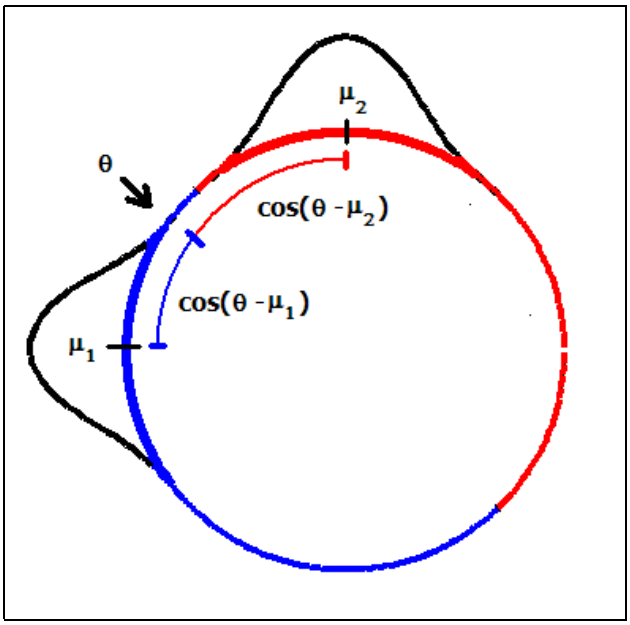

Figura 4.2: Clasificación por discriminante circular

Figura 4.2. Es importante recalcar que la función de discriminación circular (4.4) es similar a la distancia circular (2.5), es decir, mide la menor longitud de arco entre θ y cada μ_j, pero en (4.4) es máxima para la clase donde $\theta = \mu_j$.

Sin el supuesto de igualdad en el parámetro de concentración se obtiene el siguiente resultado.

Proposición 5 (Discriminante Circular). *La regla de discriminación suponiendo distribución circular de von Mises-Fisher es:*

$$\eta(x) = \underset{j}{\operatorname{argmax}} \left\{ \ln(\Pi_j) + \ln\left(\frac{1}{I_0(\kappa_j)}\right) + \kappa_j \cos(\theta - \mu_j) \right\} \tag{4.5}$$

Demostración. Para el caso de $p = 2$ sin el supuesto de igualdad del parámetro de

dispersión se tiene:

$$\begin{aligned}\eta(x) &= \underset{j}{\operatorname{argmax}} \left\{ \frac{\Pi_j}{I_0(\kappa_j)} \exp\{\kappa_j \cos(\mu_j)\cos(\theta) + \sin(\mu_j)\sin(\theta)\} \right\} \\ &= \underset{j}{\operatorname{argmax}} \left\{ \frac{\Pi_j}{I_0(\kappa_j)} \exp\{\kappa_j \cos(\theta - \mu_j)\} \right\} \\ &= \underset{j}{\operatorname{argmax}} \left\{ \ln(\Pi_j) + \ln\left(\frac{1}{I_0(\kappa_j)}\right) + \kappa_j \cos(\theta - \mu_j) \right\}\end{aligned}$$

□

El resultado anterior es equivalente al encontrado por [Morris y Laycock, 1974].

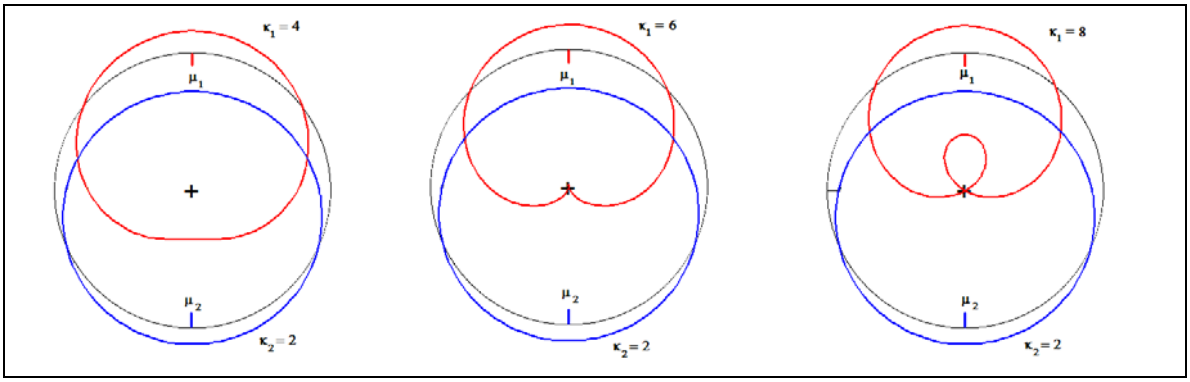

Figura 4.3: Funciones de discriminación para ($\kappa_1 = 4, 6, 8$,$\mu_1 = \pi/2$) y ($\kappa_2 = 2$,$\mu_1 = 3\pi/2$)

La función de discriminación tiene una figura estilo cardioide para cada clase j, centrado en μ_j con aplanamiento para valores pequeños de κ_j y tiene un efecto de apuntamiento para valores grandes de κ_j. La Figura 4.3 muestra la funciones de discriminación en caso de dos clases, para diferentes valores de κ_j. El valor de la función de discriminación para cada θ está determinado en la Figura 4.3 por la distancia del centro del circulo al punto de la función hacia la dirección θ. Los

valores negativos de la función de discriminación, se ubican en sentido opuesto a θ.

El caso esférico $p = 3$ es de interés principalmente en climatología, astronomía y geología, ya que el problema de clasificación con mediciones en la esfera celeste o en coordenadas sobre el globo terreste puede ser solucionados con clasificadores direccionales esféricos.

Proposición 6 (Función de Discriminante Esférico). *Bajo el supuesto de distribución esférica de von Mises-Fisher en cada clase, la regla de clasificación es:*

$$\eta(x) = \underset{j}{\operatorname{argmax}} \quad [\ln(\Pi_j) + \ln\left(\frac{\sqrt{\kappa_j}}{I_{1/2}(\kappa_j)}\right) + \kappa_j \left\{\cos(\mu_{1j})\cos(\theta_1) + \sin(\theta_1)\sin(\mu_{1j})\cos(\theta_2 - \mu_{2j})\right\}] \quad (4.6)$$

Demostración. Para el caso de $p = 3$ se tiene que:

$$\begin{aligned}
\eta(x) &= \underset{j}{\operatorname{argmax}} \left[\frac{\pi_j \kappa_j^{3/2-1}}{I_{3/2-1}(\kappa_j)} \exp\Big[\kappa_j \Big[\cos(\mu_{1j})\cos(\theta_1) \right. \\
&\qquad \left. + \cos(\mu_{2j})\sin(\mu_{1j})\cos(\theta_2)\sin(\theta_1) + \sin(\mu_{1j})\sin(\mu_{2j})\sin(\theta_1)\sin(\theta_2)\Big]\Big]\right] \\
&= \underset{j}{\operatorname{argmax}} \left[\frac{\pi_j \kappa_j^{1/2}}{I_{1/2}(\kappa_j)} \exp\Big[\kappa_j \Big[\cos(\mu_{1j})\cos(\theta_1) \right. \\
&\qquad \left. + \sin(\mu_{1j})\sin(\theta_1)[\cos(\mu_{2j})\cos(\theta_2) + \sin(\mu_{2j})\sin(\theta_2)]\Big]\Big]\right] \\
&= \underset{j}{\operatorname{argmax}} \left[\frac{\pi_j \sqrt{\kappa_j}}{I_{1/2}(\kappa_j)} \exp\Big[\kappa_j \Big[\cos(\mu_{1j})\cos(\theta_1) \right. \\
&\qquad \left. + \sin(\mu_{1j})\sin(\theta_1)\cos(\theta_2 - \mu_{2j})\Big]\Big]\right] \\
&= \underset{j}{\operatorname{argmax}} \left[\ln(\Pi_j) + \ln\left(\frac{\sqrt{\kappa_j}}{I_{1/2}(\kappa_j)}\right) \right. \\
&\qquad \left. + \kappa_j \left\{\cos(\mu_{1j})\cos(\theta_1) + \sin(\theta_1)\sin(\mu_{1j})\cos(\theta_2 - \mu_{2j})\right\}\right]
\end{aligned}$$

donde el vector θ es la transformación a coordenadas esféricas del vector x y el vector μ_{ij} es la i-ésima coordenada del parámetro de centralidad de la j-ésima clase. □

Proposición 7 (Discriminante Esférico con igualdad de κ_j y Π_j). *La regla de discriminación suponiendo distribución de von Mises Fisher esférico con igual parámetro de concentración κ_j y probabilidad a priori en cada clase es:*

$$\eta(x) = \underset{j}{\operatorname{argmax}} \left\{\cos(\mu_{1j})\cos(\theta_1) + \sin(\theta_1)\sin(\mu_{1j})\cos(\theta_2 - \mu_{2j})\right\} \tag{4.7}$$

4.2. Clasificación con estimador de densidad direccional

En la estimación de densidad por k vecinos más cercanos para datos circulares es necesario cambiar la forma de medir distancia utilizando medidas direccionales, tales como (2.5) ó (2.6). La regla de clasificar a x sería asignar a la clase más presente entre los vecinos más cercanos circularmente. En otras palabras, examinar las etiquetas de los k vecinos más cercanos y escoger la clase de mayor frecuencia.

Actualmente hay un gran número de artículos sobre estimación no paramétrica de funciones de densidad de probabilidad de variables aleatorias que toman valores en R^k por medio de funciones kernel. En esta sección se describe la estimación de densidad por kernel para variables aleatorias que toman valores en esferas unitarias q dimensionales Ω_q, con $q \geq 2$. Teóricamente, [Bai et al., 1988], resume el proceso de estimar la densidad $f(x)$ sobre Ω_q, de la siguiente manera. Primero seleccione una función uno a uno, ϕ, desde Ω_q a R^{k-1}. Entonces se puede aplicar un kernel usual sobre los datos transformados $\phi(X_1), \ldots, \phi(X_n)$, y encontrar la estimación de los datos transformados. Finalmente, utilizando la tranformación inversa, se obtiene la estimación de $f(X)$. En [Bai et al., 1988] se muestra que este proceso tiene dos dificultades en la práctica. Primero, la transformación inversa puede ser computacionalmente difícil de calcular, especialmente para valores grandes de k. Segundo, cualquier transformación que sea usada deja al menos un punto de la densidad sin estimar. Un desarrollo natural de los estimadores de densidad por kernel

en datos direccionales es el siguiente. Si dos vectores x y y son unitarios entonces el ángulo entre ellos tiene un coseno igual a $x^T y$. Por esta razón [Hall et al., 1987] sugiere reemplazar la cantidad $t - x_i$ en la ecuación (3.11), por su equivalente esférico $t^T x_i$ si t y x_i están sobre la superficie de la esfera. La interpretación literal de esta proposición contruye el estimador:

$$\widehat{f}(t) = \frac{1}{n} c_0(\kappa) \sum_{i=1}^{n} K\left(\kappa t^T x_i\right) \tag{4.8}$$

donde el nuevo parámetro de suavizamiento (o ancho de banda) κ reemplaza el h de (3.15) y donde $c_0(\kappa)$ es seleccionado tal que $\widehat{f}(t)$ integre a uno. [Hall et al., 1987] muestra que funciones del tipo $K(t) = e^t$ son buenas funciones kernel.

Teniendo en mente que cuando t es cercano a x_i, $t^T x_i$ es cercano a uno, pero $t - x_i$ es cercano a cero, se propone otro tipo de estimador de kernel similar en espíritu a los kernel en espacios euclidianos:

$$\widehat{f}(t) = \frac{1}{n} d_0(\lambda) \sum_{i=1}^{n} K\left(\lambda(1 - t^T x_i)\right) \tag{4.9}$$

donde el nuevo parámetro de suavizamiento es λ, nuevamente, $d_0(\lambda)$ es seleccionado para que la integral sea uno. [Hall et al., 1987] muestra que funciones kernel del tipo $K(t) = e^t$ en 4.8 y $K(t) = e^{-t}$ son equivalentes a medida que n aumenta.

Una propuesta natural para realizar la estimación por densidad de kernel en conjuntos direccionales surge de la ecuación (3.12) al utilizar la distancia circular definida en (1) y utilizar como función $g(z)$ alguna utilizada en el kernel de Hall,

es decir, $g(z) = e^{-t}$.

Proposición 8 (Kernel Direccional Coseno). *Se define el kernel direccional basado en la distancia circular por la expresión:*

$$\widehat{f}(t) = \frac{1}{n} C_\psi \sum_{i=1}^{n} e^{-\kappa \rho_1(t, x_i)}$$
$$\widehat{f}(t) = \frac{1}{n} C_\psi \sum_{i=1}^{n} e^{-\kappa \sum_{j=1}^{p} (1 - \cos(t_j - x_{ij}))}$$

donde κ es un parámetro de suavizamiento y C_ψ es la constante de integración (para que $\widehat{f}(t)$ integre a uno).

4.3. Extensión de la clasificación direccional a conjuntos estándares

El desarrollo de clasificadores para datos direccionales tiene aplicaciones limitadas ya que requiere que la matriz de características contenga registros angulares. El procedimiento que se describe en esta sección, implementa los clasificadores direccionales a conjuntos de datos con observaciones en $\mathbb{R}^p$. Suponga que p variables son medidas en n casos. Se organizan estas mediciones en una matriz X de nxp, así las columnas representan las variables. La distinción entre variables y casos no es fija a priori, ya que en muchas ocasiones la prioridad principal la tiene el análisis de relación entre variables y no entre individuos [Mukherjee et al., 1999]. Sea $R = (r_{jk})$ denotando la matriz de coeficientes de correlación nxn calculada

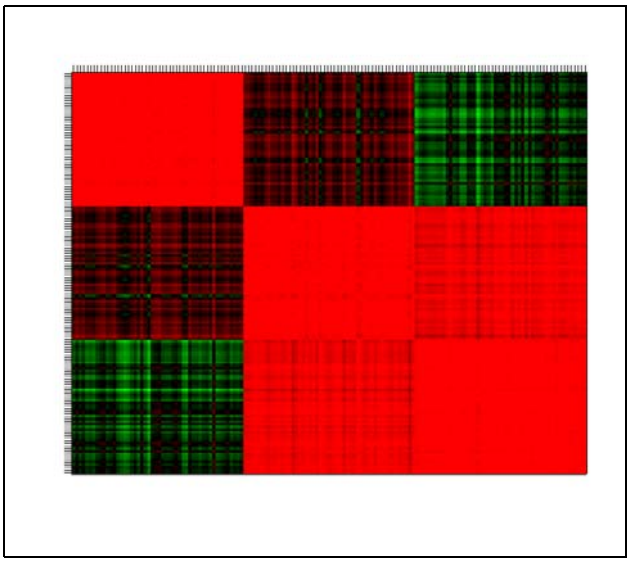

Figura 4.4: Matriz de correlación para Iris

apartir de las variables de X^T. Recientemente ha sido de interés la clasificación o construcción de cluster basado en la matriz de correlación [Trosset, 2002].

La Figura 4.4 muestra la matriz de correlación entre individuos para el clásico conjunto *Iris* [Fisher, 1936], donde las correlaciones son representadas en tonalidades rojas o verdes, de acuerdo a si son cercanas a 1 ó -1 respectivamente. Dicho conjunto de datos tiene 150 individuos pertenecientes a tres clases. La figura muestra una alta correlación entre individuos de la misma clase.

Una de las ventajas estadísticas de trabajar con la distancia euclideana es la existencia de variedad de técnicas multivariadas que operan directamente sobre X. Si se reemplaza la distancia euclidiana por alguna otra medida de disimilaridad, entonces se tienen dos caminos a seguir. Primero, limitarse a métodos que

operan directamente con similaridad. Segundo, utilizar escalamiento multidimensional para proyectar los objetos a espacios euclidianos. Similaridades son fáciles de convertir en disimilaridad y viceversa. Pero la correlación varía entre $[-1, 1]$, por tanto no es una medida de similaridad. La clave de la implementación de técnicas de clasificación circular a conjuntos de cualquier tipo está en el hecho que el coeficiente de correlación de producto momento de Pearson entre la variables x_j y x_k es:

$$\rho(x_j, x_k) = \left(\frac{x_j^T x_k}{\|x_j\| \|x_k\|} \right) = \cos(\theta) \tag{4.10}$$

donde θ es el ángulo entre dos las dos variables centradas x_j y x_k . Así, esta correlación es conocida como un medida de separación angular. Luego a partir de la matriz R se puede obtener una matriz de vectores con medidas direccionales C, al aplicar $C = \arccos(R)$ donde la columna j-ésima denotada por c_j, tiene ángulo de separación entre x_j y x_i con $i = 1, \ldots, n$. Esta idea de representación de matrices a medidas direccionales es encontrada en [Trosset, 2002].

Otras medidas angulares puede ser utilizadas para el efecto de transformar X a medidas angulares, por ejemplo el coeficiente de alienación definido por [Basilevsky, 1983]:

$$\varphi(j, k) = \frac{\left[\|x_j\|^2 \|x_k\|^2 - (x_j^T x_k) \right]^{1/2}}{\|x_j\| \|x_k\|} = \sin(\theta^*) \tag{4.11}$$

toma valores entre 0 y 1, obteniendo valor máximo cuando los vectores son perpendiculares.

4.3.1. Algoritmo de clasificación por k vecinos más cercanos direccionales

Suponga que se tiene un conjunto de datos X de dimensión nxp y un vector de clases Y de dimensión nx1, donde $Y[i,1] = j$ indica que la fila i-ésima de X pertenece a la clase j, con $j = 1, \ldots, J$. El algoritmo de clasificación por k vecinos más cercanos (KNN) direccionales para un nuevo registro x^{NEW} de dimensión 1xp se describe a continuación:

Conversión a datos direccionales: Sea $Z = [x^{NEW}|X]$ con $n+1$ filas y p columnas. Se calcula $R = \rho(Z^T)$, la matriz de correlación entre individuos y luego la función inversa $\Theta = \arccos\{R\}$. Es importante notar que la matriz Θ es cuadrada con $n+1$ filas, donde Θ_{ij} contiene el ángulo entre los vectores x_{i-1}^T y x_{j-1}^T. Adicionalmente, la primera fila de Θ contiene el ángulo entre el vector nuevo para predecir y los vectores de la matriz X.

Aplicación de clasificadores direccionales: Se calcula la matriz de distancia $D_{1\text{x}n}(1,j)$ entre los vectores fila Θ_1 y Θ_j para $j = 2, \ldots, n+1$ donde la distancia es el promedio de la distancia circular (2.5) coordenada a coordenada.

Selección de la Clase El vector x^{NEW} es clasificado a la clase de mayor votación entre los k vecinos más cercanos en la matriz D

4.3.2. Algoritmo de clasificación con estimación de densidad por kernel direccional

Suponga que se tiene un conjunto de datos X de dimensión nxp y un vector de clases Y de dimensión nx1, donde $Y[i,1] = j$ indica que la fila i-ésima de X pertenece a la clase j, con $j = 1, \ldots, J$. Suponga que se tiene $n_1, n_2, \ldots, n_J$ registros de la clase $1, 2, \ldots, J$ respectivamente. El algoritmo de clasificación realizando la estimación de densidad direccional por kernel, para un nuevo registro x^{NEW} de dimensión 1xp se describe a continuación:

Conversión a datos direccionales: Suponga que la matriz X está ordenada de acuerdo a la clase a la que pertenece. Sea $Z = [x^{NEW}|X]$ con $n+1$ filas y p columnas. Se calcula $R = \rho(Z^T)$ la matriz de correlación entre individuos y luego la función inversa $\Theta = \arccos\{R\}$. Es importante notar que la primera fila Θ_1 contiene el ángulo entre x^{NEW} y los vectores del matriz X. Específicamente, las columnas $2, 3, \ldots, 1 + n_1$ contiene las correlaciones de x^{NEW} con los registros de la clase 1. Así mismo entre $1 + n_1$ y $1 + n_1 + n_2$ la correlación con la clase 2, y así hasta las columnas $n - n_J$ y n donde se encuentra la correlación con la clase J. Es decir: $\Theta_1 = (0, \underbrace{\theta_2, \ldots, \theta_{n_1+1}}_{Clase1}, \underbrace{\theta_{n_1+2}, \ldots, \theta_{n_1+n_2+1}}_{Clase2}, \ldots, \underbrace{\theta_{n-n_J}, \ldots, \theta_n}_{ClaseJ})$ y se denota $\Theta_1 = (\Theta_{11}, \Theta_{12}, \ldots, \Theta_{1J})$.

Aplicación de clasificadores direccionales: Se calcula kernel direccional para

cada clase, es decir:

$$\widehat{f}_j(\Theta_{1j}) = \frac{1}{n} d_0(\lambda) \sum_{i \in Clase_j} K\left(\lambda(1 - (\Theta_{1j})^T \Theta_{ij})\right)$$

Con lo que se obtiene una estimación de densidad para cada clase.

Selección de la Clase El vector x^{NEW} es clasificado a la clase donde la estimación de densidad por kernel direccional sea mayor.

4.3.3. Algoritmo de clasificación por discriminante direccional

Suponga que se tiene un conjunto de datos X de dimensión nxp y un vector de clases Y de dimensión nx1, donde $Y[i,1] = j$ indica que la fila i-ésima de X pertenece a la clase j, con $j = 1, \ldots, J$. Suponga que se tiene $n_1, n_2, \ldots, n_J$ registros de la clase $1, 2, \ldots, J$ respectivamente. El algoritmo de clasificación por discriminante direccional, para un nuevo registro x^{NEW} de dimensión 1xp se describe a continuación:

Conversión a datos direccionales: Suponga que la matriz X está ordenada de acuerdo a la clase que pertenece. Sea $Z = [x^{NEW}|X]$ con $n+1$ filas y p columnas. Se calcula $R = \rho(Z^T)$ la matriz de correlación entre individuos y luego la función inversa $\Theta = \arccos\{R\}$. Es importante notar que la primera fila Θ_1 contiene el ángulo entre x^{NEW} y los vectores del matriz X. Específicamente, las columnas $2, 3, \ldots, 1 + n_1$ contiene las correlaciones de x^{NEW} con los registros de la clase 1. Así mismo entre

$1 + n_1$ y $1 + n_1 + n_2$ la correlación con la clase 2, y así hasta las columnas $n - n_J$ y n donde se encuentra la correlación con la clase J. Es decir: $\Theta_1 = (0, \underbrace{\theta_2, \ldots, \theta_{n_1+1}}_{Clase1}, \underbrace{\theta_{n_1+2}, \ldots, \theta_{n_1+n_2+1}}_{Clase2}, \ldots, \underbrace{\theta_{n-n_J}, \ldots, \theta_n}_{ClaseJ})$ y se denota $\Theta_1 = (\Theta_{11}, \Theta_{12}, \ldots, \Theta_{1J})$.

Aplicación de clasificadores direccionales: Se calcula el discriminante direccional para cada clase, es decir:

$$\eta_j(\Theta_{1j}) = \pi_j \frac{\kappa_j^{n_j/2-1}}{I_{n_j/2-1}(\kappa_j)} \exp\{\kappa_j \sum_{i=1}^{n_j} (\cos(\mu_{ij}) \cos(\theta_{ij}) \prod_{m=0}^{i-1} \sin \mu_{mj} \sin \theta_{mj})\}$$

donde θ_i es la i-ésima coordenada de Θ_j, y κ_j y el vector μ_j son calculados a partir de la matriz Θ_{ij} la cual contiene los registro circulares para las i que pertenece a la clase j.

Selección de la Clase: El vector x^{NEW} es clasificado a la clase donde la función de discriminación sea mayor, es decir:

$$\eta(\Theta_{1j}) = \underset{j}{\text{argmax}}\{\pi_j \frac{\kappa_j^{n_j/2-1}}{I_{n_j/2-1}(\kappa_j)} \exp\{\kappa_j \sum_{i=1}^{n_j} (\cos(\mu_{ij}) \cos(\theta_{ij}) \prod_{m=0}^{i-1} \sin \mu_{mj} \sin \theta_{mj})\}\}$$

Capítulo 5

Metodología

Para llevar a cabo esta investigación se realizaron los siguientes pasos:

I.Revisión de Bibliografía: Esta investigación tiene su origen del estudio de textos clásicos de análisis de datos multivariados como [Mardia y Jupp, 1999] y [Anderson, 1984], luego al observar los principales libros de datos direccionales [Fisher, 1993], [Mardia, 1999] y [Jammalamadaka y SenGupta, 2001], es notoria la ausencia de aplicaciones de técnicas multivariadas de discriminación a datos direccionales. Un compendio de principales resultados se encuentran en el segundo y tercer capítulo de este documento.

II.Desarrollo de clasificadores direccionales: Se desarrolla tres tipo de clasificadores para vectores direccionales. El análisis de discriminante direccionales es un aplicación de la teoría de discriminación en el caso análogo de la distribución normal multivariada en datos direccionales es decir la distribución de von-Mises Fisher. Adicionalmente se desarrolla una aplicación muy natural del clásico clasificador de K-vecinos más cercano a vectores direccionales. Finalmente, se aplica la

teoría de clasificadores de kernel circulares para clasificar conjntos direccionales. El desarrollo detallado se encuentra en el cuarto capítulo .

III.Implementación de programas: Se construyeron funciones en el programa estadístico **R** para el desarrollo de clasificadores por discriminante, knn y kernel todos ellos direccionales. Adicionalmente para el proceso de estimación del error se desarrolla una función de estimación del error de clasificación por validación cruzada. El compendio de estos programas se encuentra en el apéndice.

IV.Estudio del rendimiendo en conjuntos simulados:

Bajo simulaciones se compara el rendimiento, en cuanto a porcentaje de error en la clasificación, de los clasificadores direccionales en conjuntos de datos que siguen la distribución de von Mises-Fisher, sexto capítulo.

V.Análisis del rendimiento en bases de datos de *Machine Learning*: Se utilizan cinco bases de datos reales, que han sido analizadas anteriormente por diversos autores en el marco de análisis de rendimiento de clasificadores. Estas se presentan en la tabla 5.1.

Nombre	Casos	Clases	Continuas	Discretas
Breastw	683	2	9	-
Diabetes	768	2	8	-
Glass	214	6	9	-
Ionosphere	351	2	32	-
Iris	150	3	4	-

Tabla 5.1: Descripción de bases de datos de Machine Learning

La comparación del rendimiento sobre estos conjuntos se encuentra en el capí-

tulo 6.

VI.Estudio del rendimiento en bases de datos de datos anchos Se analiza el rendimiento de clasificadores sobre bases de datos disponibles en internet, sobre tumores cancerígenos usando datos de expresión genética obtenidos mediante microarreglos. En la tabla 5.2 se presentan sus características principales:

Nombre	Casos	Clases	Variables
Golub	72	2	7129
Breastcancer	58	2	6728
Khan	83	4	6567
colonCA	62	2	2000

Tabla 5.2: Descripción de bases de datos de Microarreglos

Golub es la base de datos de expresión genética sobre cáncer de sangre o leucemia ([Golub et al., 1999]). *BreastCancer* es la base de datos de expresión genética para pacientes con cáncer de pecho ([Gruvberger et al., 2001]). *Khan* es la base de datos, para estudiar tumores cancerígenos pediátricos ([Khan et al., 2001]). *colonCA* es una base de datos de expresión genética en pacientes con cáncer de colon ([Alon et al., 1997]). La comparación del rendimiento sobre estos conjuntos se encuentra en el capítulo 6.

Capítulo 6

Aplicaciones

6.1. Resultados en datos simulaciones

6.1.1. Clasificación bajo distribución de von Mises-Fisher

Los clasificadores para datos direccionales desarrollados en el Capítulo 4 pueden ser aplicados a conjuntos de datos donde cada una de las variables predictoras contenga información en forma de ángulo. Suponga que se tiene una muestra aleatoria de 100 individuos, 50 por clase, que proviene de la distribución de von Mises-Fisher (vMF), con los siguientes parámetros:

$$\begin{cases} x_i \sim vMF_p(\mu = (\underbrace{4, 4, ..., 4}_{p-veces}), \kappa = 1) & \text{Si} \quad x_i \in \text{Clase 1}, \\ x_i \sim vMF_p(\mu = (\underbrace{0, 0, ..., 0}_{p-veces}), \kappa = 1) & \text{Si} \quad x_i \in \text{Clase 2}, \end{cases}$$

Se realiza la simulación 20 veces para cada p entre 1 y 100, y se calcula el error por validación cruzada 10, con 10 repeticiones. La Figura 6.1 compara el rendimento en cuando a tasa de error, para el clasificador de K-vecinos más cercanos direccionales y el clasificador de K-vecinos más cercanos clásico (utilizando

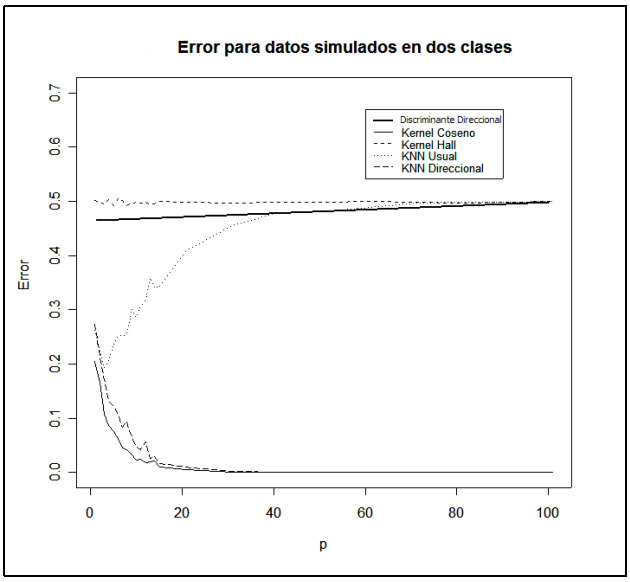

Figura 6.1: Comparación en tasa de error para simulación en dos clases

distancia euclideana).

Esta simulación ilustra claramente el efecto negativo que tiene sobre el rendimiento del clasificador por k-vecinos más cercanos, el *no* tomar en cuenta la naturaleza direccional del conjunto de datos. El clasificador por k-vecinos más cercanos clásico es afectado por la "maldición de la dimensionalidad" y tiene pésimo rendimientos a medida que el número de variable p aumenta.

Adicionalmente la Figura 6.1 muestra el rendimiento de los clasificadores basados en estimación de densidad por kernel de Hall (4.9) y el kernel utilizando la distancia circular coseno(8). Es claro que el clasificador de Hall tiene un pobre rendimiento. En general, se observa que los clasificadores basados en la distancia circular tanto en K-vecinos más cercanos y en estimación por Kernel producen, para este caso, clasificadores consistentes y con excelentes rendimientos.

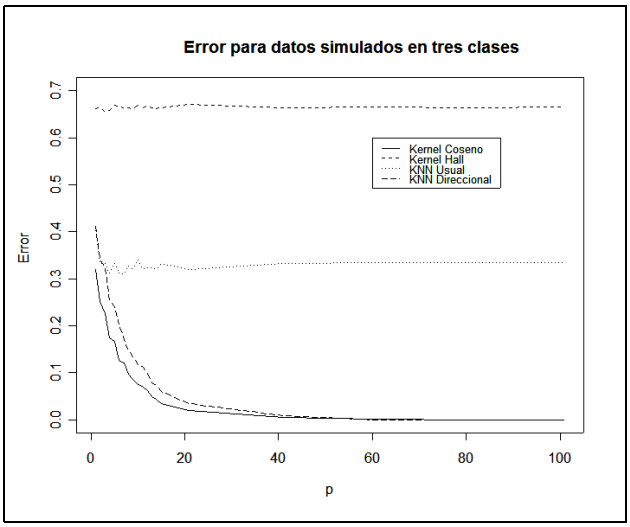

Figura 6.2: Comparación entre clasificadores en simulación para tres clases

Se observa que el clasificador por kernel basado en la distancia circular, tiene un rendimiento óptimo, ya que clasifica adecuadamente al crecer el número de variables, lo cual no sucede con el clasificador basado en el kernel de Hall.

Se puede pensar que el rendimiento de los clasificadores direccionales es favorecido en conjunto de datos donde el número de clases sea dos. Por tal motivo se presenta la siguiente simulación.

Suponga que se tiene una muestra aleatoria de 150 individuos, 50 por clase, que proviene de la distribución de von Mises-Fisher (vMF), con los siguiente parámetros:

$$\begin{cases} x_i \sim vMF_p(\mu = (\underbrace{4,4,...,4}_{p-veces}), \kappa = 1) & \text{Si} \quad x_i \in \text{Clase 1}, \\ x_i \sim vMF_p(\mu = (\underbrace{2,2,...,2}_{p-veces}), \kappa = 1) & \text{Si} \quad x_i \in \text{Clase 2}, \\ x_i \sim vMF_p(\mu = (\underbrace{0,0,...,0}_{p-veces}), \kappa = 1) & \text{Si} \quad x_i \in \text{Clase 3}, \end{cases}$$

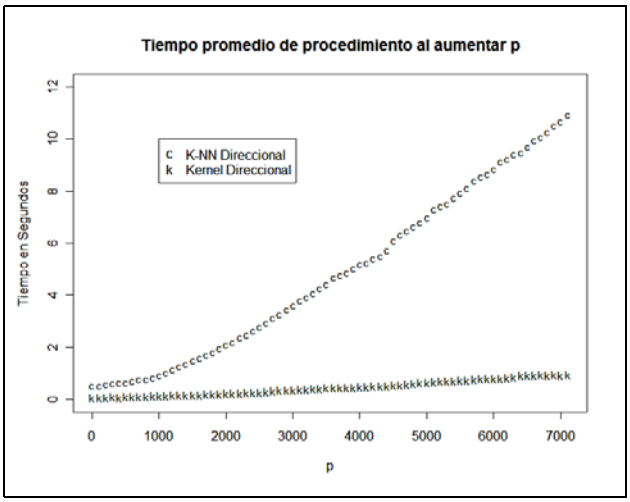

Figura 6.3: Rendimientos para diferentes p

6.1.2. Rendimiento con p mucho mayor de n

Para determinar el rendimiento de los clasificadores al aumentar la dimensión de los datos, se compara el tiempo promedio en 10 repeticiones, de la clasificación del conjunto de datos de expresión genética de Golub ([Golub et al., 1999]), para diferentes números de variables seleccionadas aleatoriamente.

La Figura 6.3, muestra que la complejidad con respecto a p es lineal para ambos algoritmos, con menor tiempo de cómputo para el kernel direccional.

6.2. Resultados en conjuntos de *Machine Learning*

Para comparar el rendimiento de los clasificadores desarrollados en esta tesis, es importante observar su comportamiento sobre conjuntos de datos que han sido analizados anteriormente en la literatura de *Machine Learning*. En todos los ca-

sos se presenta el error de clasificación estimado por validación cruzada 10 con 25 repeticiones, con su respectiva desviación estándar. Adicionalemnte, para el caso de los clasificadores con estimación de kernel, se reporta el parámetro de suavizamiento utilizado en cada caso, el cual fue encontrado como valor óptimo luego explorar diversos rangos de valores para cada conjunto. De manera análoga, para el caso del clasificador de k-vecinos más cercanos direccionales, se incluye el valor de k óptimo en cada caso.

Nombre	Kernel Hall	Kernel Cos	K-NN Direcc.	Discrim. Direcc.	Mejor Class
Breastw	20.36* (5.48)** [1]***	8.59 (2.95)	8.74 (0.43) [5]	34.5 (5.88)	3.3 (NN)♯
Diabetes	37.53 (5.97) [9]	38.01 (5.14)	31.92 (0.91) [7]	34.9 (5.35)	22.3 (Log)♯
Glass	60.50 (4.5) [0.8]	39.89 (1.49) [1000]	38.56 (2.51) [1]	65.35 (9.66)	23.8 (KNN)
Ionosphere	27.93 (0.21) [1]	5.49 (0.51) [1000]	4.80 (0.45) [5]	35.91 (7.50)	8.10 (C4.5)♯
Iris	3.33 (4.67) [3]	3.40 (0.30) [100]	3.20 (0.63) [7]	9.13 (8.72)	2.00 (LDA)♯

Tabla 6.1: Resultados de Clasificación en Bases de *Machine Learning*

*Error por VC-10 50 en repeticiones
**(Desviación Estándar del Error)
***[Parámetro del Modelo]
♯ NN=Redes Neurales ("Neural Network"), Log=Regresión Logística, C4.5= Clasificador basado en árboles C4.5, LDA=Discriminante Lineal

En la tabla 6.1 se observa que en el caso de *Ionosphere*, los clasificadores basados

en datos direccionales mejoran el *mejor* resultado reportado en la literatura hasta el momento. Los resultados en los otros conjuntos de datos pueden denominarse aceptables en comparación con el algoritmo de mejor rendimiento, reportando en la literatura. Dado que los clasificadores direccionales para conjuntos estándares se basan en la matriz de correlación entre individuos, es lógico pensar que si se tiene una estructura de correlación fuerte dentro de la clases y débil entre clases, su rendimiento será óptimo.

Para verificar esta hipótesis de manera exploratoria, se construye el diagrama de la matriz de correlación entre individuos, donde valores en rojo representan correlaciones cercanas a 1 y valores en verde representan correlaciones cercanas a -1. La líneas azules separan los registros de cada clase.

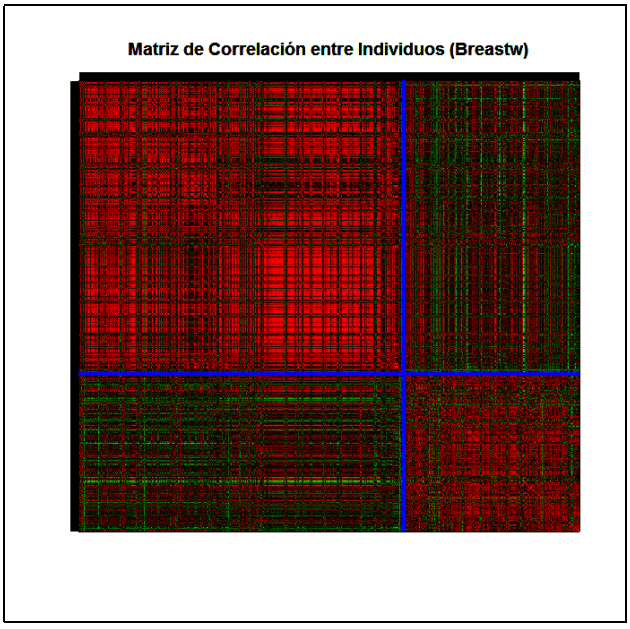

Figura 6.4: Matriz de Correlación entre individuos para el conjunto BreastW

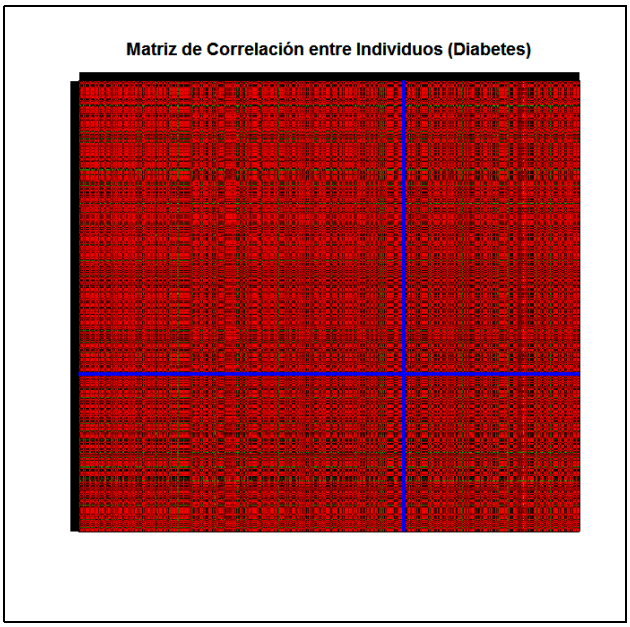

Figura 6.5: Matriz de Correlación entre individuos para el conjunto Diabetes

La Figura (6.7) muestra fuertes correlaciones dentro de cada clase y débiles entre clases; si se observa la tabla 6.1, en estas bases de datos el rendimiento de los clasificadores direccionales es excelente. Para los demás conjuntos, donde el rendimiento no es tan óptimo, la estructura de correlación entre individuos no tiene patrones relacionados con las clases (Ver Figuras 6.5 y 6.6). En las figura (6.8) muestra fuerte correlación, y por tal motivo se tiene una tasa de error baja, pero en comparación al algoritmo LDA (mejor resultado reportando en la literatura), el resultado no es tan eficiente.

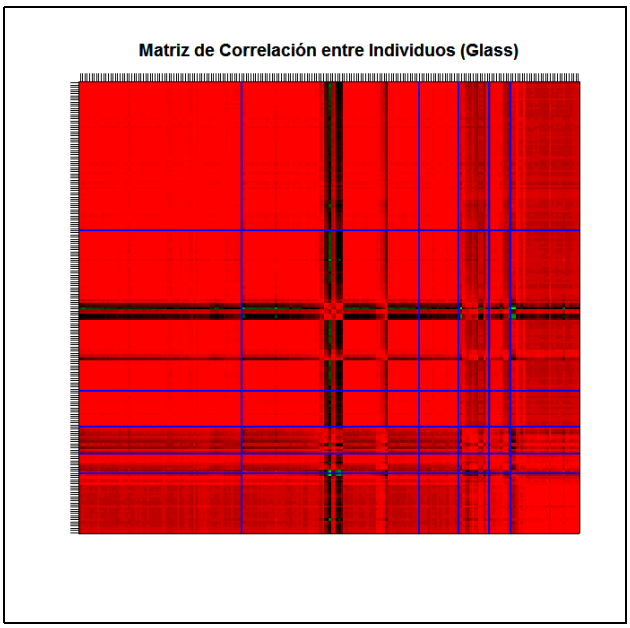

Figura 6.6: Matriz de Correlación entre individuos para el conjunto Glass

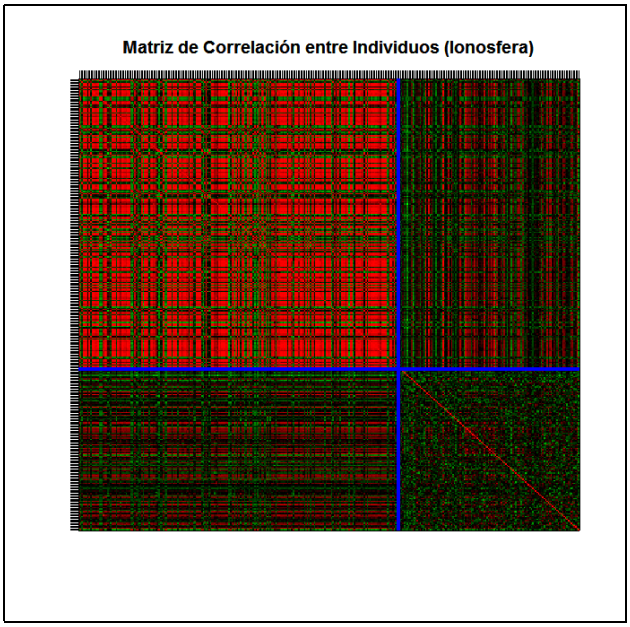

Figura 6.7: Matriz de Correlación entre individuos para el conjunto Ionosfera

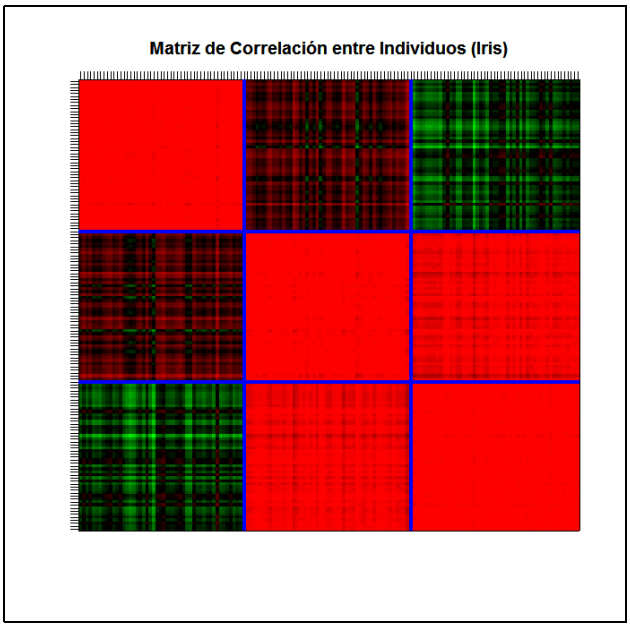

Figura 6.8: Matriz de Correlación entre individuos para el conjunto Iris

6.3. Resultados en conjuntos de Microarreglos

En las Secciones 6.1 y 6.1.2 se observa que el aumentar el número de variables con respecto al número de individuos, no causa efectos negativos al rendimientos de los clasificadores direccionales. Por tal motivo, se aplica estos clasificadores a conjuntos de datos de expresión genética, los cuales tiene la característica de p (número de variables) mucho mayor que n (número de individuos). En la tabla 6.2 se presentan los porcentajes de error por validación cruzada 10 en 50 repeticiones, entre paréntesis la correspondiente desviación estándar del error y en corchete el parámetro del modelo utilizado.

Se puede observar que el resultado de los clasificadores direccionales es bueno.

Base de Datos	Kernel Hall	Kernel Cos	K-NN Direcc.
Golub	8.75(2.70)[1]	7.30(2.45)[100]	11.59(2.9)[3]
ColonCA	12.83(2.90)[0.01]	12.10(2.21)[60]	14.54(2.64)[7]
Khan	29.20(8.3)[1.2]	19.55(3.04)[750]	24.20(3.12)[3]
Breastcancer	32.41(7.51[0.001]	26.23(3.74)[900]	33.43(0.05)[1]

Tabla 6.2: Resultados de Clasificación en Bases de Microarreglos

*Error por VC-10 50 en repeticiones (Desviación Estándar del Error) [Parámetro del Modelo]

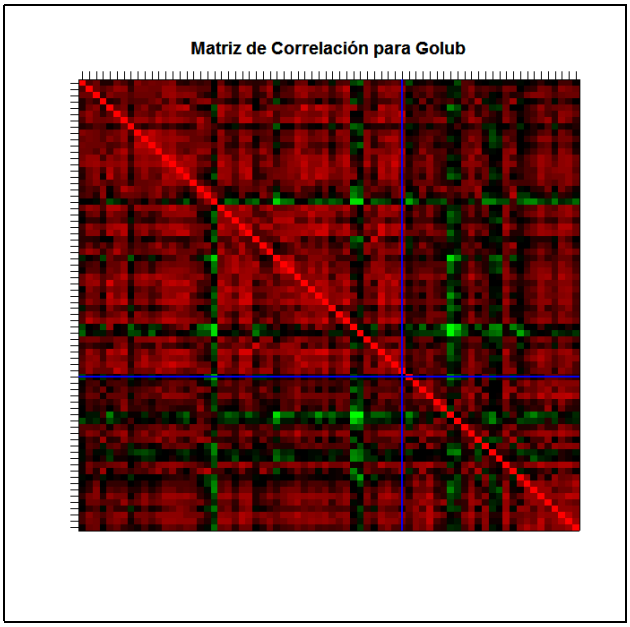

Figura 6.9: Matriz de Correlación entre individuos para el conjunto Golub

Es importante hacer énfasis que no se ha seleccionado variables, y el clasificador se enfrenta a problemas donde el número de variables en mucho mayor al número de individuos.

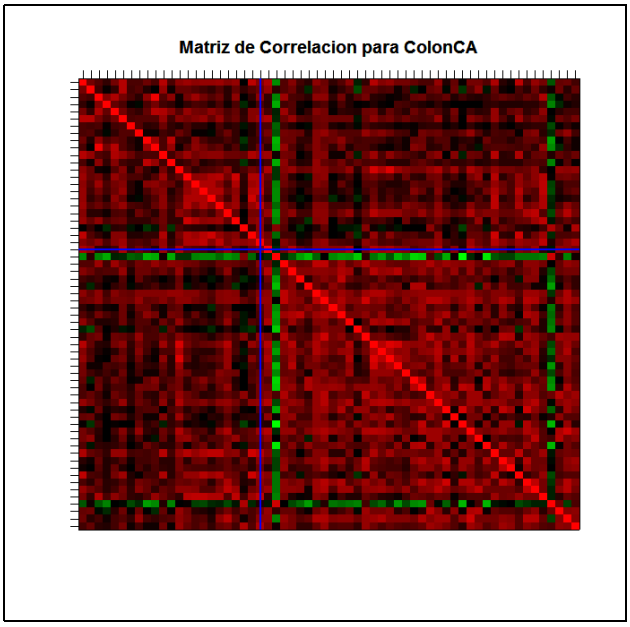

Figura 6.10: Matriz de Correlación entre individuos para el conjunto Glass

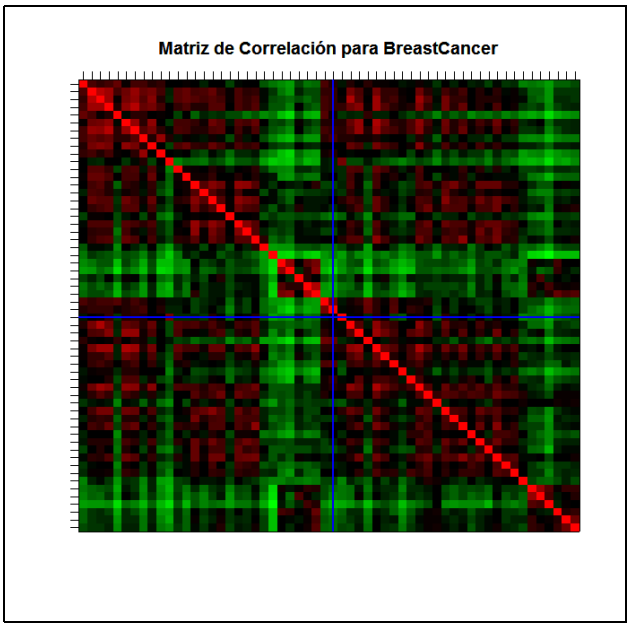

Figura 6.11: Matriz de Correlación entre individuos para el conjunto Breast-Cancer

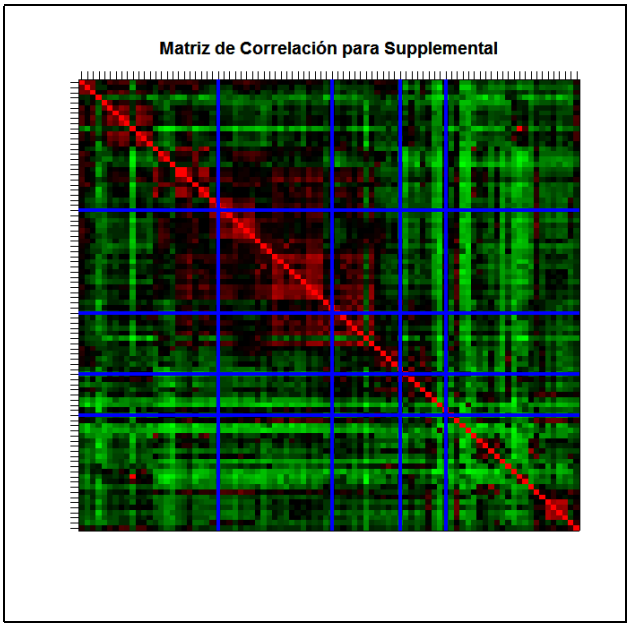

Figura 6.12: Matriz de Correlación entre individuos para el conjunto Supplemental

Capítulo 7

Conclusiones y Recomendaciones

En esta tesis se realizan los siguientes aportes basados en la clasificación de datos direccionales:

- Se desarrolla teóricamente una fórmula general para el discriminante direccional bajo distribución direccional de von Mises-Fisher.(Sección 5.1)
- Se realizan las modificaciones necesarias a los clasificadores de k-vecinos más cercanos y de estimación por kernel para ser aplicados a conjuntos direccionales, esto es utilizando la distancia direccional y el kernel de Hall. (Sección 5.2)
- Se desarrolla un nueva estimación de densidad para datos direccionales basado en la distancia direccional. (Sección 5.2)
- Se formula una metodología para aplicar los clasificadores direccionales a conjuntos de datos de cualquier tipo, basados en la correlación entre indi-

viduos. (Sección 5.3)

Adicionalmente se puede observar los siguientes aspectos al aplicar los clasificadores direccionales:

- El rendimiento de los algoritmos de clasificación para conjuntos con datos direccionales depende de la correlación que exista dentro y entre los individuos de las mismas clases.
- El kernel direccional basados en la distancia direccional tiene un mejor resultado en comparación con el kernel direccional de Hall.
- Los clasificadores direccionales no-paramétricos tienen un mejor rendimiento si se compara con los clasificadores paramétricos direccionales.
- Los clasificadores direccionales tienen buen rendimiento en conjuntos de datos donde el número de variables es mucho mayor al número de individuos.

Naturalmente se pueden recomendar investigaciones adicionales dirigidas a mejorar el desarrollo de clasificadores direccionales y estás son:

- Se debe explorar el efecto que tiene la selección de variables antes de utilizar los clasificadores direccionales, especialmente en conjuntos de datos donde n es muy pequeño.
- Es recomendable desarrollar clasificadores para datos direccionales siguiendo los principios de Regresión Logística y Árboles de Decisión.

- Es importante, especialmente cuando p es mucho mayor que n implementar algoritmos similares a Mínimos Cuadrados Parciales para conjuntos direccionales.

Apéndice A

Programas

A.1. k-vecinos más cercanos direccionales

A.1.1. Distancia Circular

```
dist.circ<-function(x,y,type)
{
type="c"
ny<-dim(y)[1]
dist<-rep(0,ny)
if (type =="p")
     {
     for (i in 1:ny)
        {
        dist[i]<-sum(pi- abs( pi - abs (x-y[i,])))
        }
     dist
     }
else
    {
    for (i in 1:ny)
        {
        dist[i]<-sum(1-cos(x-y[i,]))
        }
    dist
    }
}
```

A.1.2. k-vecinos más cercanos circular y direccional

Predice para un nuevo registro *"xnew"*, la clase basado en la matriz *"data"* relacionado con su vector de clases *"class"*. *"type"* puede ser c o p de acuerdo a la distancia direccional que se desee.

```
knn.circ<-function(xnew,data,class,knn,type)
{
mdist<-dist.circ(xnew,data,type)
classknn<-tabulate(class[rank(mdist)<knn+1])/knn
classpred<-which(classknn==max(classknn))
classpred
}
knn.direc<-function(data,class,k,type)
{
n<-dim(data)[1]
p<-dim(data)[2]
res<-rep(0,n)
if (type=="s")
    {
    data.cor<-cor(t(data))
    data<-acos(data.cor)
    }
for (i in 1:n)
    {
    res[i]<-knn.circ(data[i,],data,class,k,type)
    }
    res
}
```

A.1.3. k-vecinos más cercanos direccionales con conjunto de entrenamiento y de prueba

```
knn.direc.test<-function(data.train,data.test,class.train,k,typedata,typedist)
{
n<-dim(data.test)[1]
p<-dim(data.test)[2]
n2<-dim(data.train)[1]
p2<-dim(data.train)[2]
res<-rep(0,n)
if (typedata=="s")
    {
    data.cor<-cor(t(rbind(data.test,data.train)))
    data<-acos(data.cor)
    for (i in 1:n)
        {
        res[i]<-knn.circ(data[i,seq(n+1,n+n2)],
        data[seq(n+1,n+n2),seq(n+1,n+n2)],class.train,k,typedist)
        }
        }
else
    {
        for (i in 1:n)
        {
        res[i]<-knn.circ(data.test[i,],data.train,class.train,k,typedist)
        }
    }
res
}
```

A.1.4. Validación Cruzada 10 para 1-vecino más cercano clásico

```
cv10.knn1<-function(data,class)
{
exit<-rep(0,10)
n<-dim(data)[1]
p<-dim(data)[2]+1
dataclass<-cbind(data,class)
sort.data<-dataclass[rank(runif(n)),]
dim.segm<-floor(n/10)
for (j in 1:10)
    {
    d.s<-((j-1)*dim.segm+1):(j*dim.segm)
    if (j==10){d.s<-((j-1)*dim.segm+1):n}
    data.test<-sort.data[d.s,-p]
    data.train<-sort.data[-d.s,-p]
    class.test<-sort.data[d.s,p]
    class.train<-sort.data[-d.s,p]
    pred<-knn1(data.train,data.test,class.train)
    b<-table(pred,class.test)
    miss=sum(b)-sum(diag(b))
    exit[j]<-miss
    }
error<-sum(exit)/n
exit<-list(missclas=error)
}
```

A.1.5. Validación Cruzada 10 para k-vecinos más cercanos direccionales

```
cv10.knndirec<-function(data,class,k,typedata,typedist)
{
exit<-rep(0,10)
n<-dim(data)[1]
p<-dim(data)[2]+1
dataclass<-cbind(data,class)
sort.data<-dataclass[rank(runif(n)),]
dim.segm<-floor(n/10)
for (j in 1:10)
    {
    d.s<-((j-1)*dim.segm+1):(j*dim.segm)
    if (j==10){d.s<-((j-1)*dim.segm+1):n}
    data.test<-sort.data[d.s,-p]
    data.train<-sort.data[-d.s,-p]
    class.test<-sort.data[d.s,p]
    class.train<-sort.data[-d.s,p]
    b<-table(pred,class.test)
    miss=sum(b)-sum(diag(b))
    exit[j]<-miss
    }
error<-sum(exit)/n
exit<-list(missclas=error)
}
```

A.1.6. Validación cruzada 3 para k-vecinos más cercanos direccionales

```
cv3.knndirec<-function(data,class,k,typedata,typedist)
{
exit<-rep(0,3)
n<-dim(data)[1]
p<-dim(data)[2]+1
dataclass<-cbind(data,class)
sort.data<-dataclass[rank(runif(n)),]
dim.segm<-floor(n/3)
for (j in 1:3)
    {
    d.s<-((j-1)*dim.segm+1):(j*dim.segm)
    if (j==3){d.s<-((j-1)*dim.segm+1):n}
    data.test<-sort.data[d.s,-p]
    data.train<-sort.data[-d.s,-p]
    class.test<-sort.data[d.s,p]
    class.train<-sort.data[-d.s,p]
    pred<-knn.direc.test(data.train,data.test,class.train,k,typedata,typedist)
    b<-table(pred,class.test)
    miss=sum(b)-sum(diag(b))
    exit[j]<-miss
    print(paste(j,"-Fold with missclass=",miss))
    }
error<-sum(exit)/n
exit<-list(missclas=error)
}
```

A.2. Kernel direccional

A.2.1. Estimación de Kernel de Hall y Coseno para un vector

La función *"s.r"*, convierte a una fila en un vector de norma uno. La función *"kdirec.v"* calcula para un nuevo registro *"vector"*, la estimación de kernel sobre la matriz de datos *"x"* y el parámetro de suavizamiento *"kappa"*. *"type"* indica el tipo de kernel, es decir, si es el kernel de Hall *"hall"* o el kernel basado en la distancia

direccional *"cos"*.

```
s.r<-function(row)
{
row<-row/(sqrt(sum(row^2)))
row
}
kdirec.v<-function(vector,x,kappa,type)
{
n<-dim(x)[1]
if(type=="hall")
    {
    kc<-(1/n)*sum(exp((-kappa*((1-t(x))*vector))))
    }
if (type=="cos")
    {
    kc<-(1/n)*sum(exp(-kappa*(dist.circ(vector,x))))
    }
    kc
}
```

A.2.2. Predicción con estimación por densidad de kernel direccional

Calcula para el nuevo registro *"vector"* la clase que se predice, dado el conjunto de datos *"x"* con su respectivo vector de clase *"class"*, con un parámetro de suavizamiento *"kappa"* con respondiente al kernel direccional seleccionando *"hall"* para el kernel de Hall o *"cos"* para el kernel basado en la distancia direccional. *"typed"* especifica el tipo de datos que se tiene, es decir, si se tiene un conjunto de datos que ya es direccional se utiliza *"d"*, pero si el conjunto de datos no es direccional, use la opción *"s"*, y se realiza la clasificación basado en la correlación entre individuos

```
v.kdirec<-function(vector,x,class,kappa,typek,typed)
{
class<-as.numeric(class)
J<-length(tabulate(class))
kc<-rep(0,J)
if (typed=="s")
    {
    x<-acos(cor(t(rbind(vector,x))))
    x.v<-x[1,-1]
    x.v<-s.r(x.v)
    x<-x[-1,-1]
    x<-t(apply(x,1,s.r))
    for (j in 1:J)
        {
        class.j=which(class==j)
        x.j<-x[class.j,]
        kc[j]<-kdirec.v(x.v,x.j,kappa,typek)
        }
    }
if (typed=="d")
    {
    for (j in 1:J)
        {
        class.j=which(class==j)
        x.j<-x[class.j,]
        kc[j]<-kdirec.v(vector,x.j,kappa,typek)
        }
    }
p.c<-which(kc==max(kc))
list(kdirec=kc,pred=p.c)
}
```

A.2.3. Kernel direccional con conjunto de entrenamiento y de prueba

```
kdirec<-function(xtrain,classtrain,xtest,kappa,typek,typed)
{
n.test<-dim(xtest)[1]
pred<-rep(0,n.test)
class<-as.numeric(classtrain)
kc<-matrix(0,n.test,length(tabulate(class)))
for (i in 1:n.test)
    {
    a<-v.kdirec(xtest[i,],xtrain,classtrain,kappa,typek,typed)
    pred[i]<-a\$pred
    kc[i,]<-a\$kdirec
    }
list(kdirec=kc,pred=pred) }
```

A.2.4. Validación cruzada 10 para el kernel direccional

```
cv10.kdirec<-function(data,class,kappa,typek,typed)
{
exit<-rep(0,10)
n<-dim(data)[1]
p<-dim(data)[2]+1
dataclass<-cbind(data,class)
sort.data<-dataclass[rank(runif(n)),]
dim.segm<-floor(n/10)
for (j in 1:10)
    {
    d.s<-((j-1)*dim.segm+1):(j*dim.segm)
    if (j==10){d.s<-((j-1)*dim.segm+1):n}
    data.test<-sort.data[d.s,-p]
    data.train<-sort.data[-d.s,-p]
    class.test<-sort.data[d.s,p]
    class.train<-sort.data[-d.s,p]
    pred<-kdirec(data.train,class.train,data.test,kappa,typek,typed)$pred
    b<-table(pred,class.test)
    miss=sum(b)-sum(diag(b))
    exit[j]<-miss
    }
error<-sum(exit)/n
exit<-list(missclas=error)
}
```

A.2.5. Validación cruzada 3 para el kernel direccional

```
cv3.kdirec<-function(data,class,kappa,typek,typed)
{
exit<-rep(0,3)
n<-dim(data)[1]
p<-dim(data)[2]+1
dataclass<-cbind(data,class)
sort.data<-dataclass[rank(runif(n)),]
dim.segm<-floor(n/3)
for (j in 1:3)
    {
    d.s<-((j-1)*dim.segm+1):(j*dim.segm)
    if (j==3){d.s<-((j-1)*dim.segm+1):n}
    data.test<-sort.data[d.s,-p]
    data.train<-sort.data[-d.s,-p]
    class.test<-sort.data[d.s,p]
    class.train<-sort.data[-d.s,p]
    pred<-kdirec(data.train,class.train,data.test,kappa,typek,typed)\$pred
    b<-table(pred,class.test)
    miss=sum(b)-sum(diag(b))
    exit[j]<-miss
    }
error<-sum(exit)/n
exit<-list(missclas=error)
}
```

Bibliografía

[Ackermann, 1997] Ackermann, H. (1997). A note on circular nonparametrical classification. *Biometrical Journal*, 5:577–587.

[Alon et al., 1997] Alon U. et al. (1997). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissue probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. USA*, 96:6745-6750.

[Anderson, 1984] Anderson, T. (1984). *An Introduction to Multivariate Statistical Analysis*. Wiley-Interscience.

[Bai et al., 1988] Bai, Z., Radhakrisha, R., y Zhao, C. (1988). Kernel estimators of density function of directional data. *Journal of Multivariate Analysis*, 27.

[Baragona, 2003] Baragona, R. (2003). Further results on lund's statistic for identifying cluster in a circular data set with application to time series. *Communications in Statistics - Simulation and Computation*, 32(3).

[Basilevsky, 1983] Basilevsky, A. (1983). *Applied Matrix Algebra in the Statistical Sciences*. New Jersey North-Holland.

[Devroye et al., 1996] Devroye, L., Gyorfi, L., Krzyzak, A., y Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag.

[Fisher, 1989] Fisher, N. (1989). Smoothing a sample of circular data. *Journal of Structurak Geology*, 11(6).

[Fisher, 1993] Fisher, N. (1993). *Statistical Analysis of Circular Data*. Cambridge University Press.

[Fisher y Lee, 1992] Fisher, N. y Lee, A. (1992). Regression models for an angular response. *Biometrics*, 48.

[Fisher y Lee, 1994] Fisher, N. y Lee, A. (1994). Time series analysis of circular data. *Journal Royal Statistical Society B*, 56(2).

[Fisher, 1936] Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Ann. Eugen*, 7:179–188.

[Fix y Hodges, 1951] Fix, E. y Hodges, J. (1951). Discriminatory analysis nonparametric discrimination: Consistency properties. Technical Report 21-49-004, US Air Force.

[Friedman, 2003] Friedman, J. H. (2003). Recent advances in predictive (machine) learning. Technical report, Stanford Linear Accelerator Center.

[Fuller et al., 1996] Fuller, M., Laj, C., y Herrero-Bervera, E. (1996). The reversal of the earth's magnetic field. *Amer. Sci.*, 84.

[Golub et al., 1999] Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, C., Mesirov, M., Coller, J., Loh, H., Downing, J., Caligiuri, M., Bloomfield, C., y Lander, E. (1999). Molecular clasification of cancer: class discovery and class prediction by gene espression monitoring. *Science*, 286.

[Gruvberger et al., 2001] Gruvberger, S., Ringnir, M., Chen, Y., Panavally, S., Saal, L., Borg, E., Fernv, M., Peterson, C., y Meltzer, P. (2001). Estrogen receptos status in breast cancer is associatedqith remarkably distinct gene expression patterns. *Cancer Research*, 61:5979–5984.

[Hall et al., 1987] Hall, P., Watson, G., y Cabrera, J. (1987). Kernel density estimation with spherical data. *Biometrika*, 74:751–762.

[Jammalamadaka y SenGupta, 2001] Jammalamadaka, S. R. y SenGupta, A. (2001). *Topics in Circular Statistics*. World Scientific Publishing Company.

[Khan et al., 2001] Khan, J., Wei, J., Ringnir, M., Saal, L., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C., Peterson, C., y Meltzer, P. (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicine*, 7:673–679.

[Kulkarni et al., 1998] Kulkarni, S., Lugosi, G., y Venkatesh, S. (1998). Learning pattern classification - a survey. *IEEE Transaction on Information Theory*, 44(6):2178–2206.

[Lenth, 1981] Lenth, R. (1981). Robust measures of location for directional data. *Technometrics*, 23:77–81.

[Lund, 1999a] Lund, U. (1999a). Cluster analysis for directional data. *Communications in Statistics – Simulation and Computation*, 28.

[Lund, 1999b] Lund, U. (1999b). Least circular distance regression for directional data. *Journal of Applied Statistics*, 26(6).

[Lund, 2002] Lund, U. (2002). Tree-based regression for a circular response. *Communications in Statistics - Thoery and Methods*, 31(9).

[Mardia, 1999] Mardia, K. (1999). Directional statistics and shape analysis. *Journal of Applied Statistics*, 26(8):949–959.

[Mardia y Jupp, 1999] Mardia, K. y P.Jupp (1999). *Directional Statistics*. John Wiley & Sons.

[McLachlan, 1992] McLachlan, G. (1992). *Discriminant Analysis and Statistical Pattern Recognition*. Jhon Wiley and Sons.

[Morris y Laycock, 1974] Morris, J. y Laycock, P. (1974). Discriminant analysis of directional data. *Biometrika*, 61(2).

[Mukherjee et al., 1999] Mukherjee, S., Tamayo, P., Slonim, D., Verri, A., Golub, T., Mesirov, J., y Poggio, T. (1999). Support vector machine classification of microarray data.

[Sengupta y Rao, 1967] Sengupta, S. y Rao, J. (1967). Statistical analysis of croobedding azimuths from the kanthi formation around bheemaram. *Sankhya*, 28.

[Silverman, 1972] Silverman, B. (1972). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall.

[Trosset, 2002] Trosset, M. (2002). Visualizing correlation. Technical report, Department of Mathematics College of William and Mary.

[Watson, 1974] Watson, G. (1974). *Statistics in Sphere*. Springer-Verlag.

[Webb, 1999] Webb, A. (1999). *Statistical Pattern Recognition*. Arnold.