Clasificación

Una introducción a k-vecinos, LDA y kernels

Sergio Camelo
Quantil SAS - Minería de Datos

El Contexto

- Se tiene una población y un conjunto de grupos. Cada individuo pertenece a un único grupo
- En el caso general, la clasificación es aleatoria.
- Las características de cada individuo, sin embargo, determinan la probabilidad con la que cada individuo pertenece a un grupo determinado. A estas características las llamaremos covariantes.
- Como consecuencia, individuos con las mismas características tienen la misma distribución de probabilidad sobre el conjunto de grupos.
- En algunas aplicaciones puede que la división sea determinística, cómo se resuelve esta situación?

El Problema de Clasificación

- Se dispone de una base de datos marcada, es decir, de un conjunto de individuos a los que se les asocia un conjunto de covariantes y un label con el grupo al que el individuo pertenece.
- Se busca entrenar un algoritmo que estime la probabilidad con la que un individuo pertenece a un determinado grupo dado un conjunto de covariantes.
- En la mayoría de los casos, basta con hacer una predicción del grupo al que pertenecerá el individuo.

Algunas Aplicaciones

- Riesgo de crédito
- Reconocimiento de caracteres
- Reconocimiento de voz
- Buscadores en Internet
- Mercado financiero: Burbujas
- Diagnóstico Médico

Formulación Matemática

- Tenemos una muestra $((X_1, Y_1), \ldots, (X_n, Y_n))$ i.i.d donde los X_i son vectores aleatorios continuos (covariantes), y los Y_i son variables discretas que indican el grupo al que pertenece el individuo i. Denotemos por G al conjunto de grupos.
- Un clasificador es una función f que asocia a un vector de covariantes X, una predicción acerca del grupo al que pertenece un individuo con las características X
- Claramente f depende de la muestra aleatoria.

Clasificador de Bayes

- Todo el tiempo se asume una distribución de probabilidad para la pareja (X,Y).
- Si se conoce la distribución de probabilidad, se pueden calcular las probabilidades condicionales

$$P(Y=1|X)=p_1$$
$$P(Y=2|X)=p_2$$
$$P(Y=3|X)=p_3$$

- Cuál será el mejor clasificador si se quiere minimizar la probabilidad de equivocarse?

Clasificador de Bayes

- Al clasificador que minimiza la probabilidad de equivocarse, dado que se conoce la función de distribución de los datos, se le llama el clasificador de Bayes.
- El error de Bayes es la probabilidad de equivocarse cuando se usa el clasificador de Bayes.
- Ningún método puede superar al clasificador de Bayes. ¿Por qué?

Clasificador de Bayes

Se tiene una moneda donde la probabilidad de obtener cara es a

Se quiere determinar si la moneda está cargada hacia la cara o hacia el sello. ¿Cuál es el clasificador de Bayes? ¿Cuál es el error de Bayes?

Clasificadores Paramétricos

- Ya se vio que si se conoce la distribución de los datos, se puede obtener el estimador que minimiza el error de equivocarse.
- Cuando no se conoce la distribución, se puede asumir que la distribución pertenece a una familia de distribuciones determinadas por un conjunto de parámetros, como pueden ser la media, la varianza, etc.
- Estos parámetros son estimados a partir de la muestra. Después de identificar la distribución, se usa la metodología de Bayes asumiendo que la distribución estimada es la verdadera.

Linear Discriminant Analysis

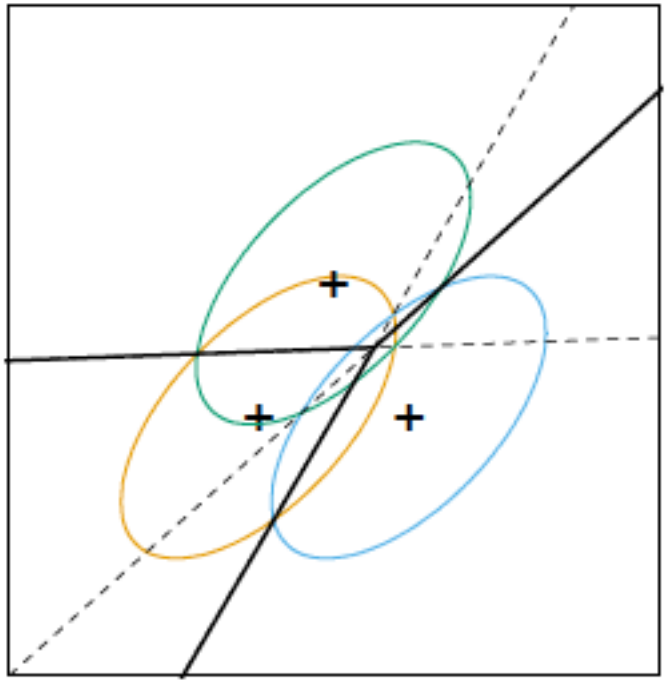
Se supone que los covariantes dentro de una clase determinada se distribuyen multinormales. Esto quiere decir que la densidad condicional

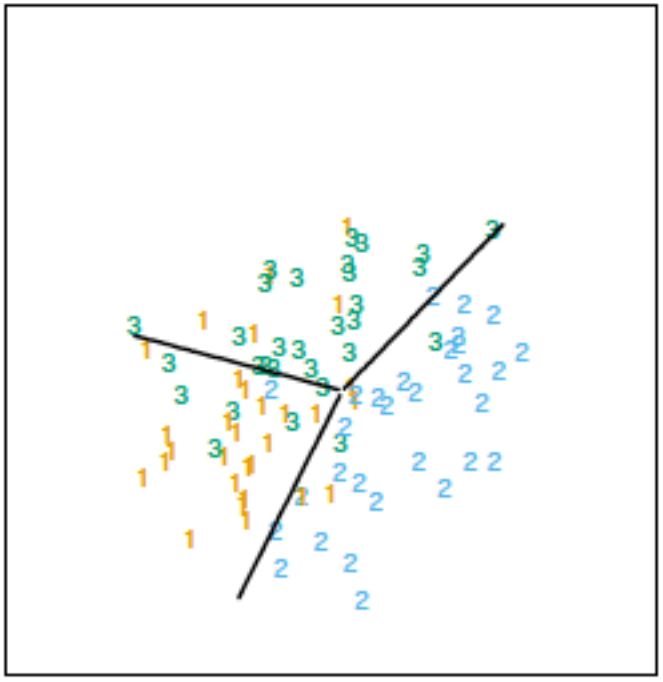
$$f_k(x) = f(x \mid Y = k)$$

es una densidad multinormal con media y matriz de var-cov desconocidas. Es decir,

$$f_k(x) = \frac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}_k|^{1/2}} e^{-\frac{1}{2}(x-\mu_k)^T \mathbf{\Sigma}_k^{-1}(x-\mu_k)}.$$

LDA





Linear Discriminant Analysis

Usando la regla de Bayes se tiene que

$$\Pr(G = k|X = x) = \frac{f_k(x)\pi_k}{\sum_{\ell=1}^{K} f_\ell(x)\pi_\ell}.$$

Así que el clasificador de Bayes será quien maximice esta probabilidad. Es fácil ver que esto es equivalente a maximizar

$$\log \frac{\Pr(G = k|X = x)}{\Pr(G = \ell|X = x)}$$

Linear Discriminant Analysis

Es decir, maximizar

$$\log \frac{\Pr(G=k|X=x)}{\Pr(G=\ell|X=x)} = \log \frac{f_k(x)}{f_\ell(x)} + \log \frac{\pi_k}{\pi_\ell}$$
$$= \log \frac{\pi_k}{\pi_\ell} - \frac{1}{2}(\mu_k+\mu_\ell)^T \mathbf{\Sigma}^{-1}(\mu_k-\mu_\ell)$$
$$+ x^T \mathbf{\Sigma}^{-1}(\mu_k-\mu_\ell),$$

donde se supone que la matriz de var-cov es la misma para todos los datos.

Linear Discriminant Analysis

Finalmente, esto equivale a maximizar

$$\delta_k(x) = x^T \mathbf{\Sigma}^{-1} \mu_k - \frac{1}{2} \mu_k^T \mathbf{\Sigma}^{-1} \mu_k + \log \pi_k$$

Como esta es una función lineal en los covariantes, si graficamos los covariantes y las regiones en las que se da la clasificación de cada grupo, estas regiones vendrán separadas por hiperplanos.

Linear Discriminant Analysis

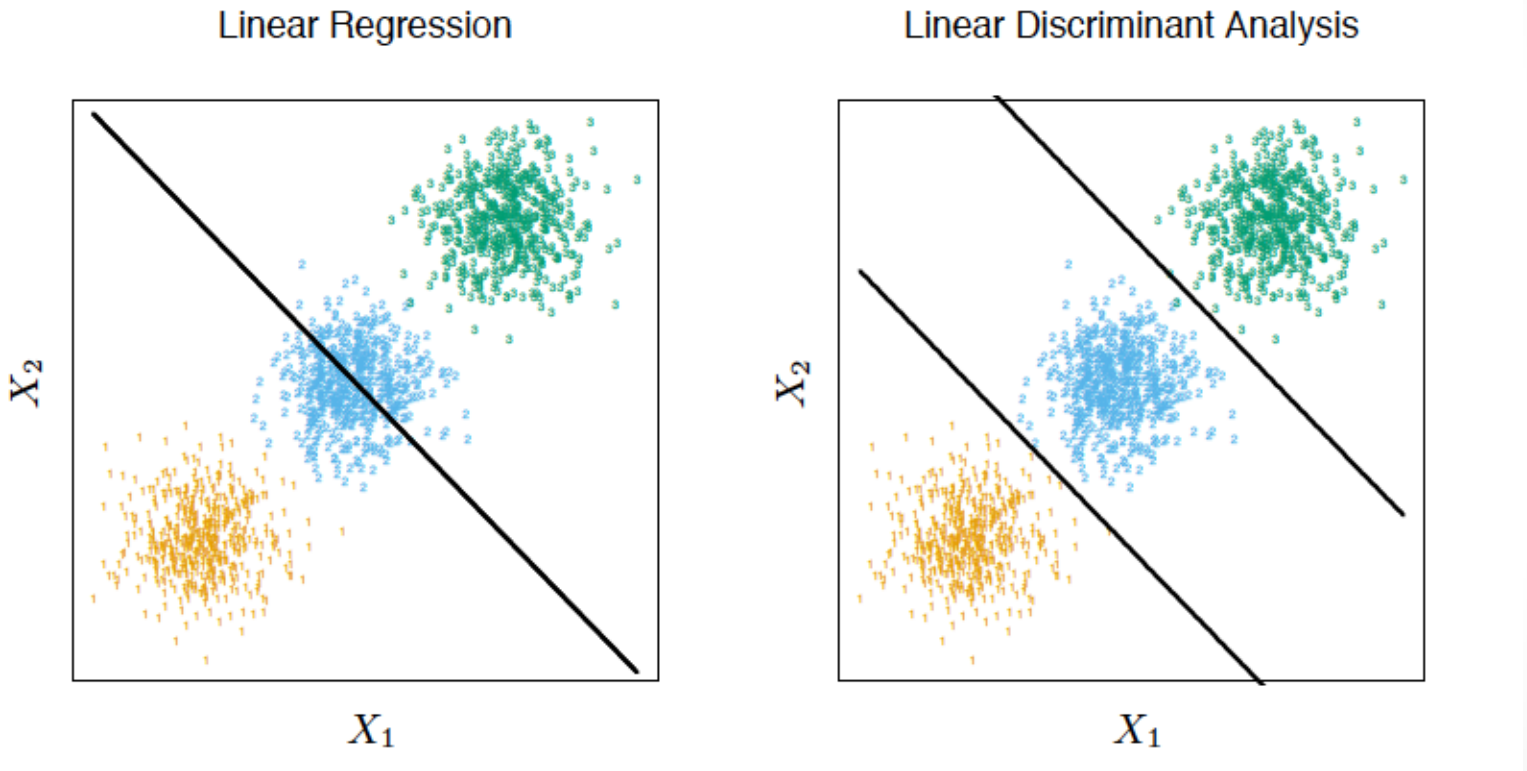
Pregunta: Cómo estimar la matriz de varianzas y covarianzas.

Linear Discriminant Analysis

Pregunta: Cómo estimar la matriz de varianzas y covarianzas.

$$\hat{\mathbf{\Sigma}} = \sum_{k=1}^{K} \sum_{g_i=k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T / (N - K).$$

LDA vs. Modelo de Prob. Lineal



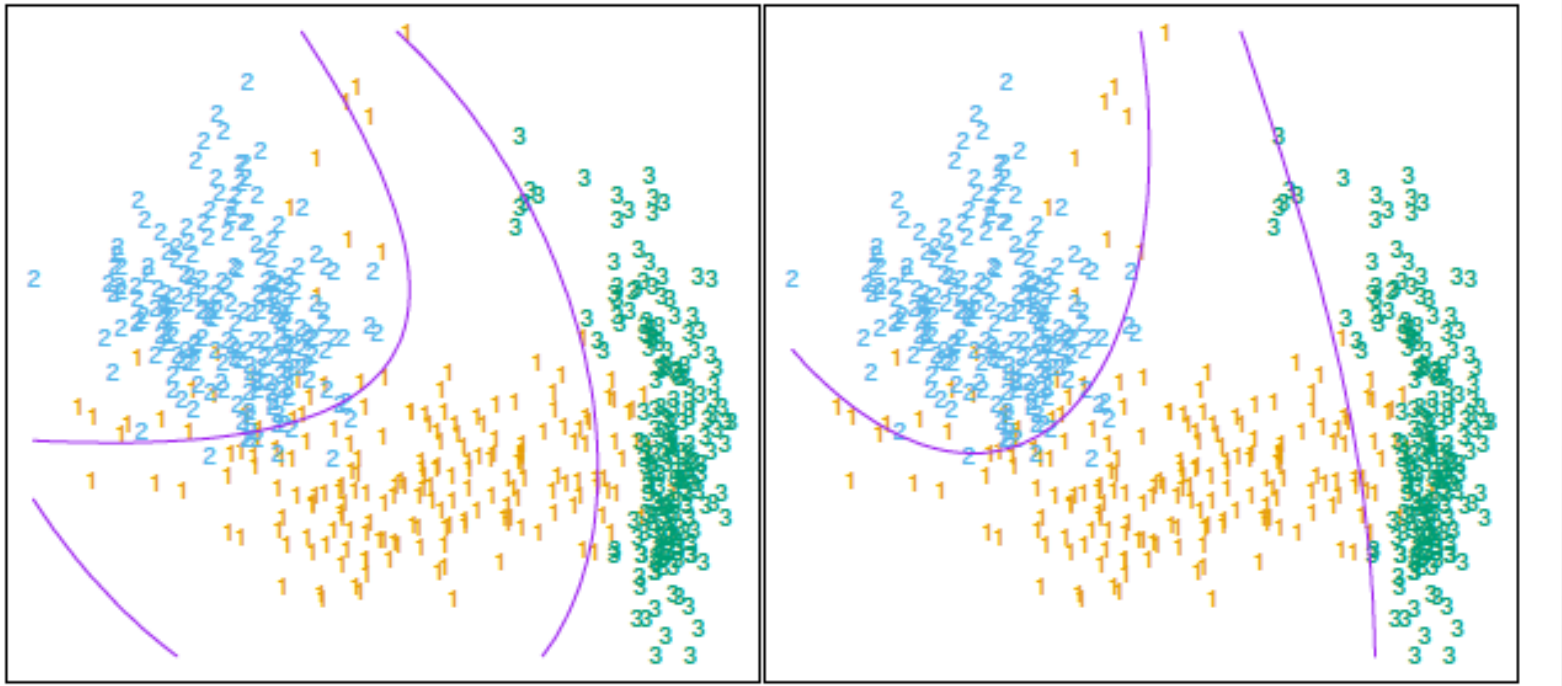
Quadratic Discriminant Analysis

Es posible suponer que las matrices de varianzas y covarianzas no son constantes entre grupos. Esto es recomendable cuando la cantidad de datos es alta.

Los bordes ya no serán hiperplanos, sino hiperboloides, elipsoides, esferas, etc.

También es posible hacer LDA, pero agregando interacciones entre covariantes. Ambas metodologías arrojan resultados similares.

LDA vs. QDA



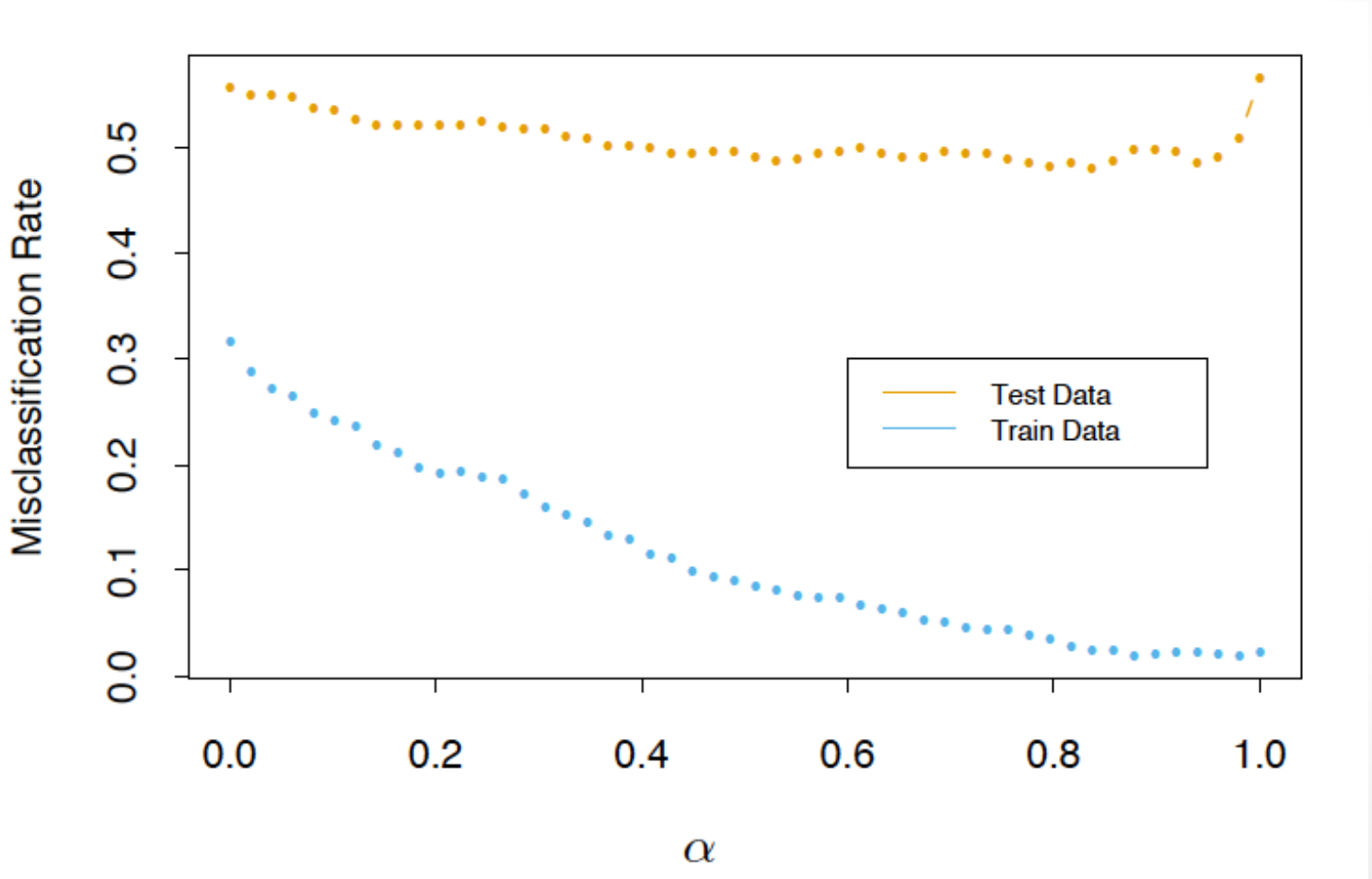
Regularized Discriminant Analysis

Se puede obtener una combinación de LDA y CDA. Para esto, se supone que la matriz de varianzas y covarianzas es una combinación lineal de la estimada en LDA y la estimada en QDA

$$\hat{\Sigma}_k(\alpha) = \alpha\hat{\Sigma}_k + (1-\alpha)\hat{\Sigma},$$

Pregunta: ¿Cómo determinar el valor de a ?

Regularized Discriminant Analysis



Clasificadores No Paramétricos

- Se llama clasificadores no paramétricos a aquellos que no asumen una forma funcional para los datos, sino que “permiten que los datos hablen por sí mismos”.
- En comparación con los clasificadores paramétricos, tienen una varianza mucho más alta, pues hay un mayor número de grados de libertad. Pueden entonces ser menos eficientes.
- Pregunta: ¿Por qué no usar siempre metodologías paramétricas?

Clasificadores No Paramétricos

- El más famoso: K-vecinos.
- Árboles de regresión, random forests.
- Redes neurales.
- Clasificadores de Kernel.

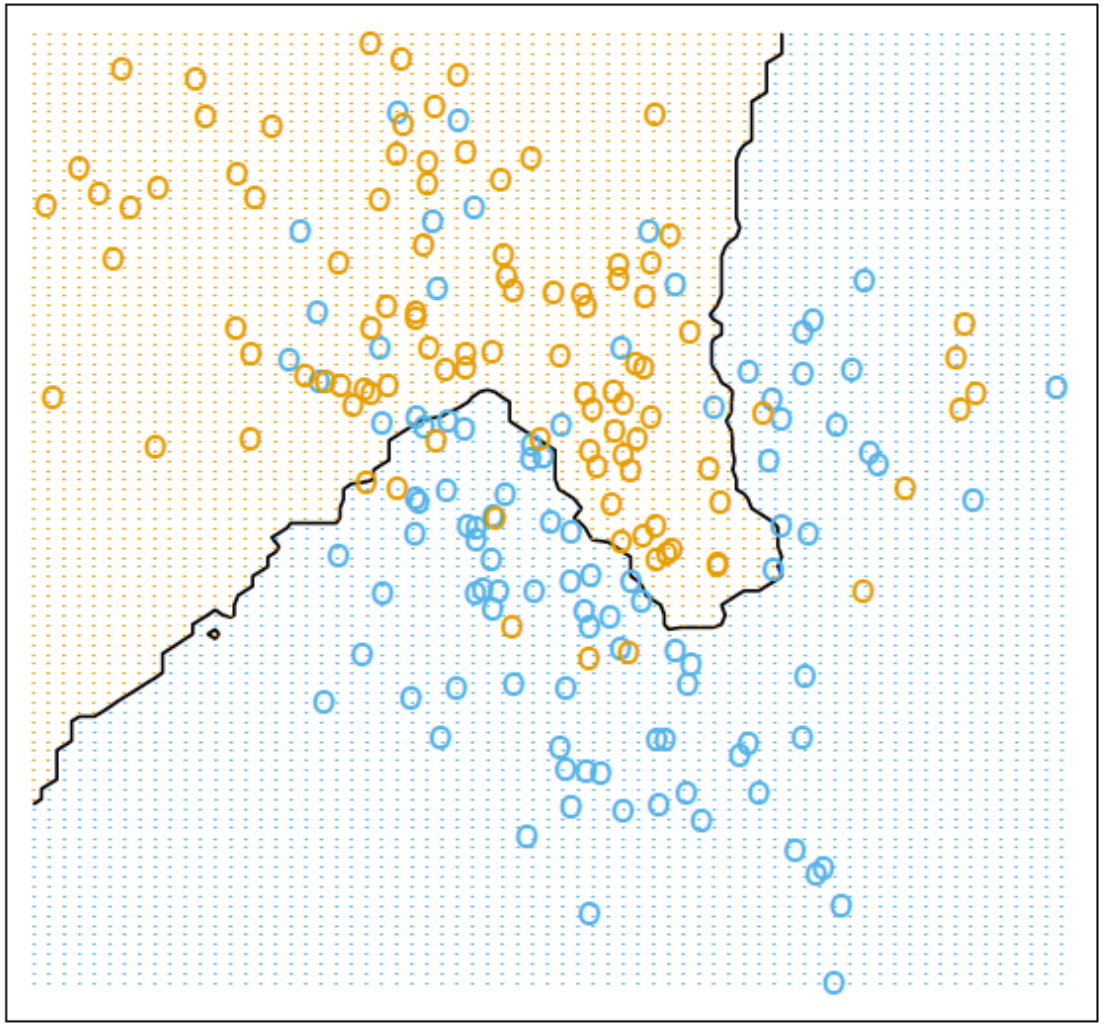
K-Vecinos

- Se define una distancia entre cualquier par de individuos. Esta distancia es una función que involucra únicamente los covariantes.
- Normalmente se usa la distancia euclidiana después de normalizar los datos si los covariantes son continuos.
- El clasificador de k-vecinos toma los k elementos de la base más cercanos a un individuo y usa los grupos a los que pertenecen para decidir el grupo al que pertenece el individuo. Generalmente se usa votación por mayoría.
- El algoritmo permite calcular probabilidades. Estas estarán dadas por

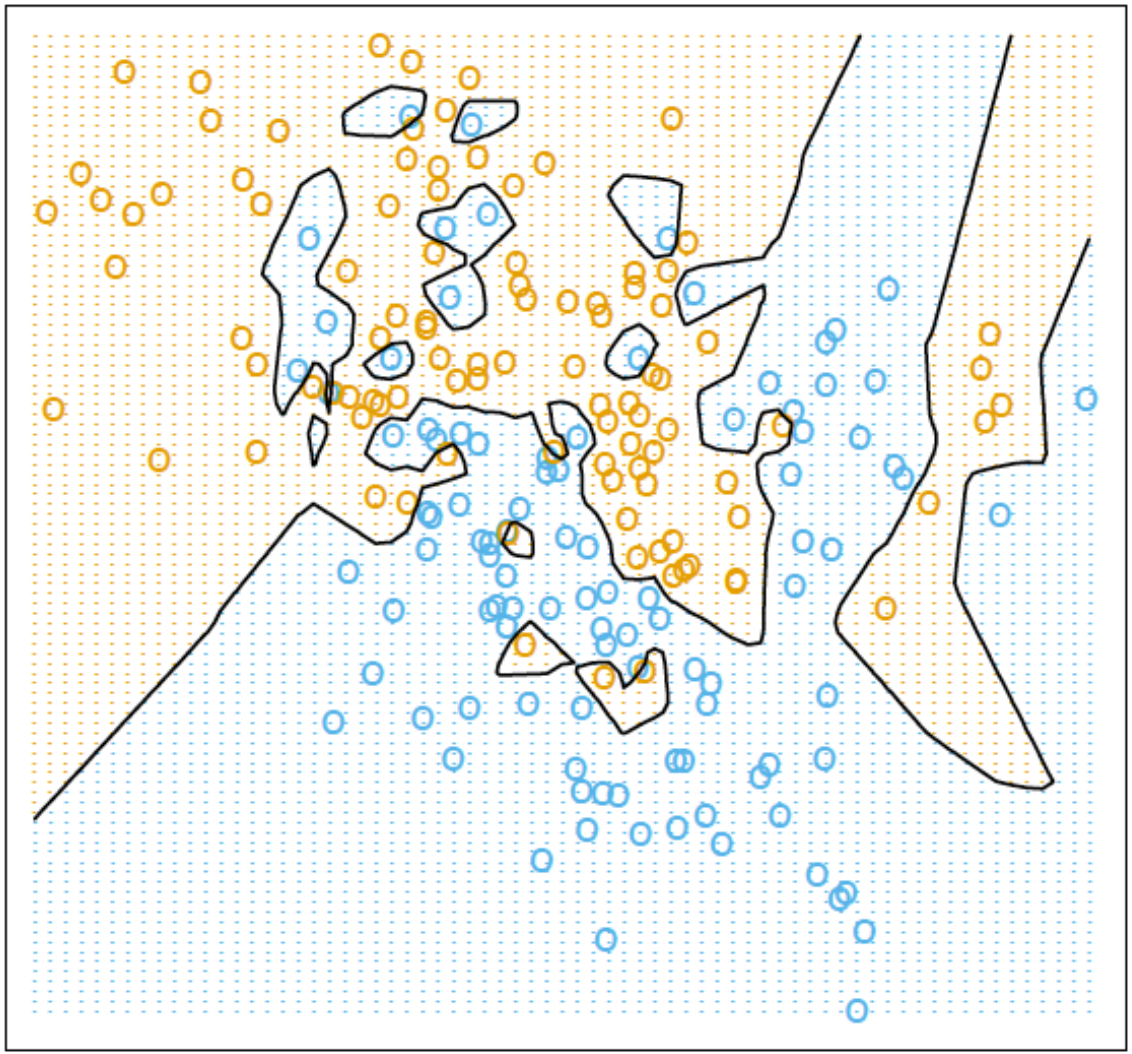
$$\hat{Y}(x) = \frac{1}{k} \sum_{x_i \in N_k(x)} y_i,$$

- ¿Qué pasa si hay un empate?

15-Nearest Neighbor Classifier

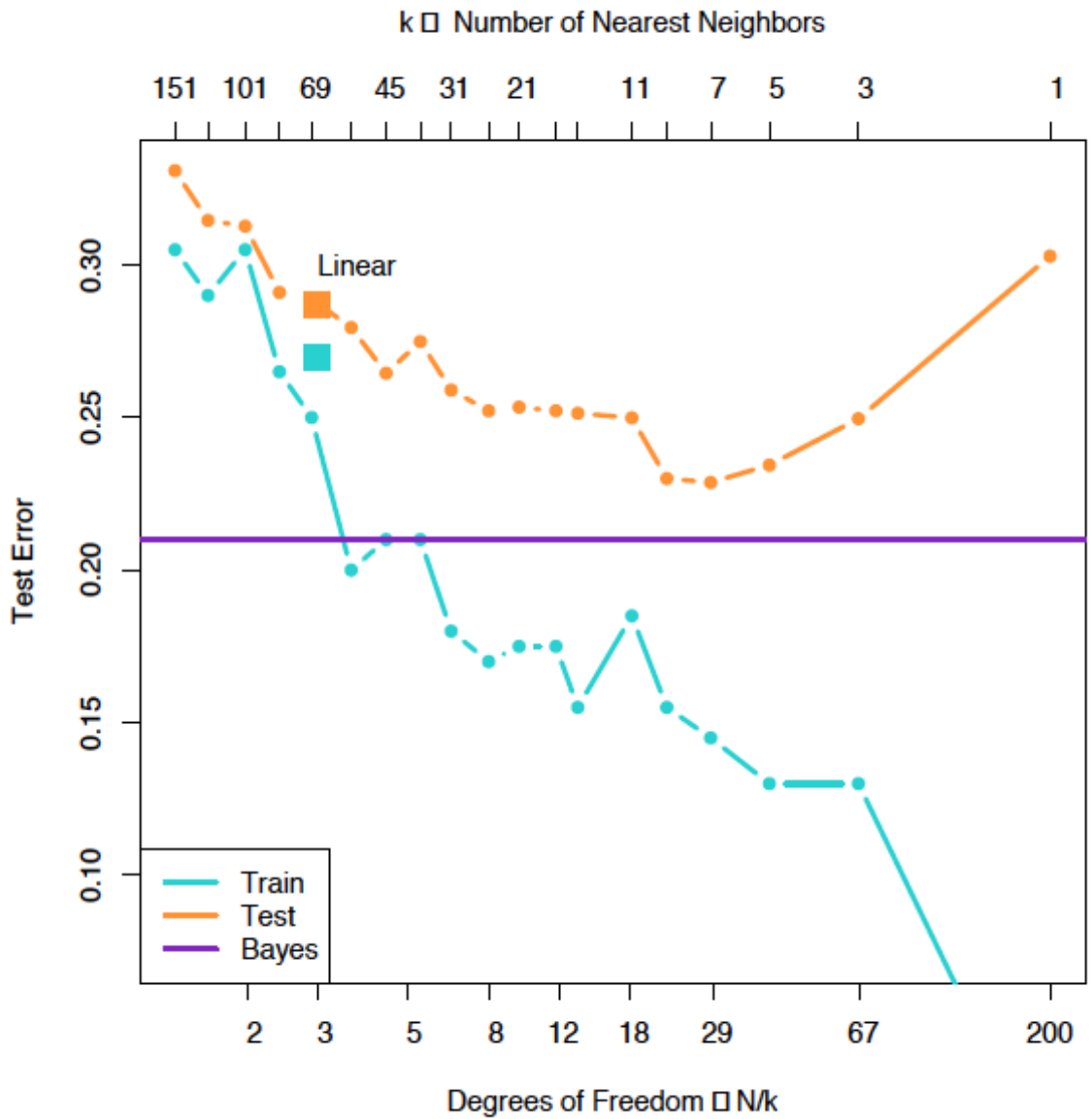


1-Nearest Neighbor Classifier

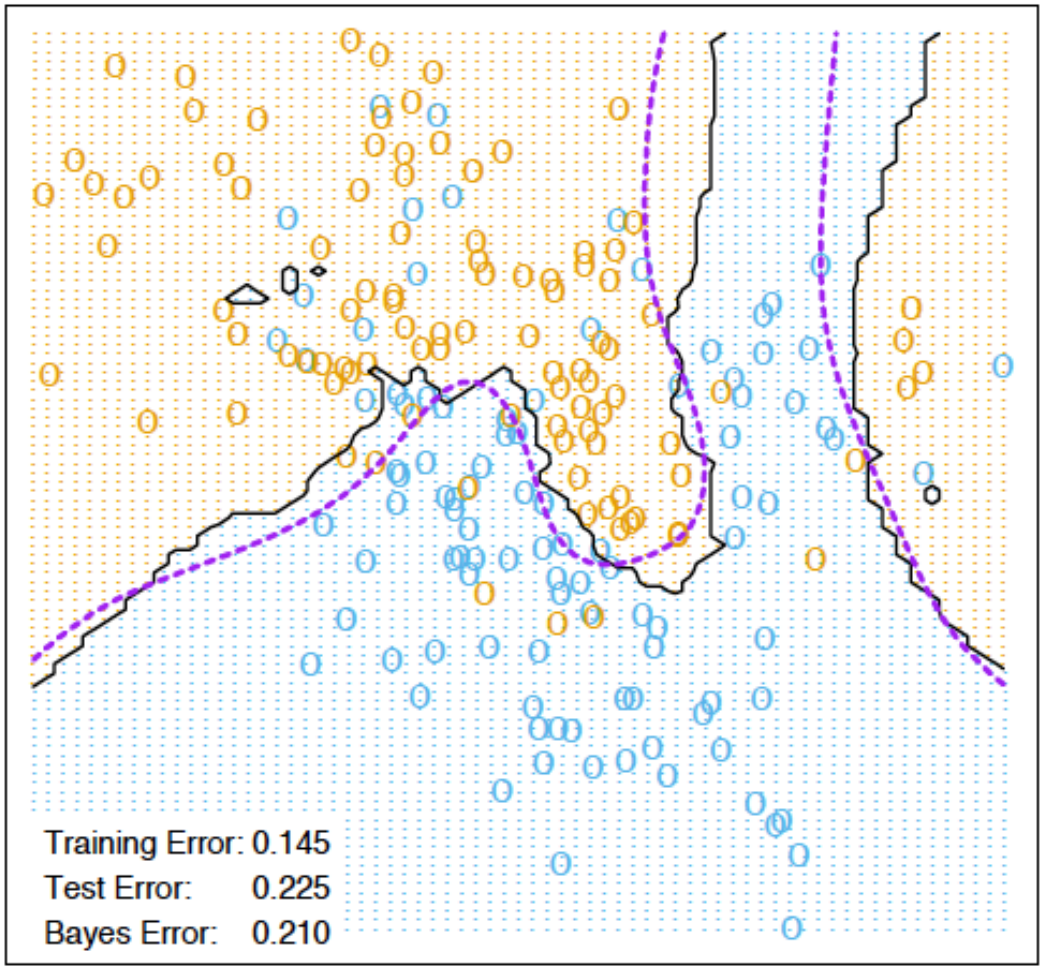


K-Vecinos

¿Cómo escoger el número de vecinos? Se da un conflicto entre sesgo y varianza.


k □ Number of Nearest Neighbors
151
101
69
45
31
21
11
7
5
3
1
Linear
Test Error
0.30
0.25
0.20
0.15
0.10
Train
Test
Bayes
2
3
5
8
12
18
29
67
200
Degrees of Freedom □ N/k

7-Nearest Neighbors



Propiedades K-Vecinos

- ¿Por qué funciona?
- Es consistente, es decir, converge al error de Bayes, cuando la muestra se va a infinito.
- Es necesario que el número de vecinos aumente. ¿A qué se debe esto?

Clasificación por Kernels

- Tenemos la regla de Bayes

$$\Pr(G = k|X = x) = \frac{f_k(x)\pi_k}{\sum_{\ell=1}^{K} f_\ell(x)\pi_\ell}.$$

- Así que si conocemos las prior $f_k(x) = f(x \mid Y = k)$

y las probabilidades no condicionales, $\rho_k = P(Y = k)$

es posible estimar la probabilidad de pertenecer a un grupo.

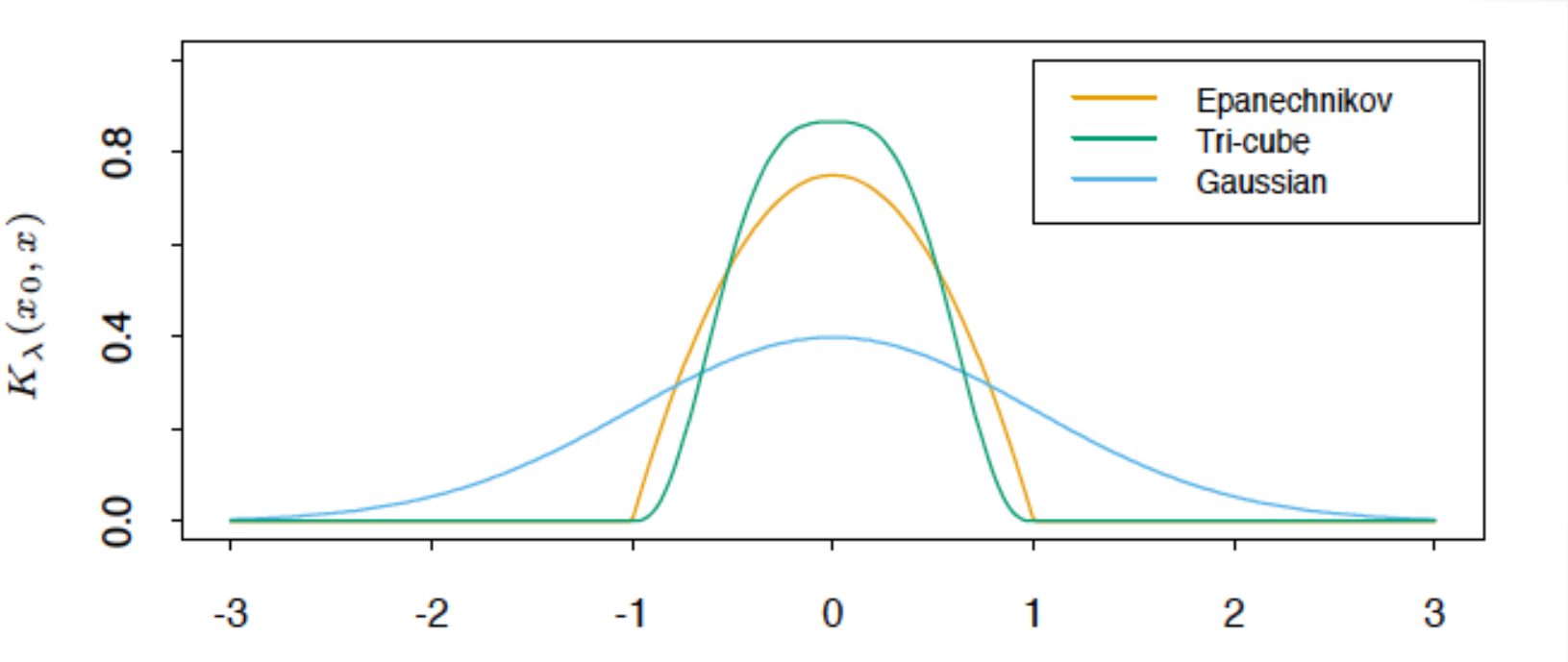
Clasificación por Kernels

Ideas para estimar $p_k = P(Y = k)$?

Clasificación por Kernels

- La estimación de las prior $f_k(x) = f(x \mid Y = k)$

se hace por Kernels.


$K_\lambda(x_0, x)$
0.0
0.4
0.8
-3
-2
-1
0
1
2
3
Epanechnikov
Tri-cube
Gaussian

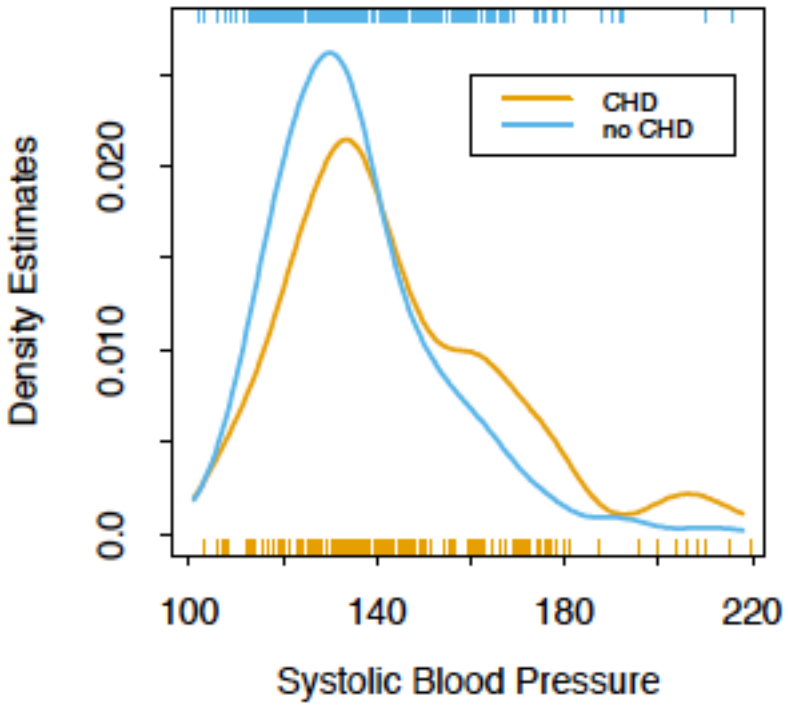
Clasificación por Kernels

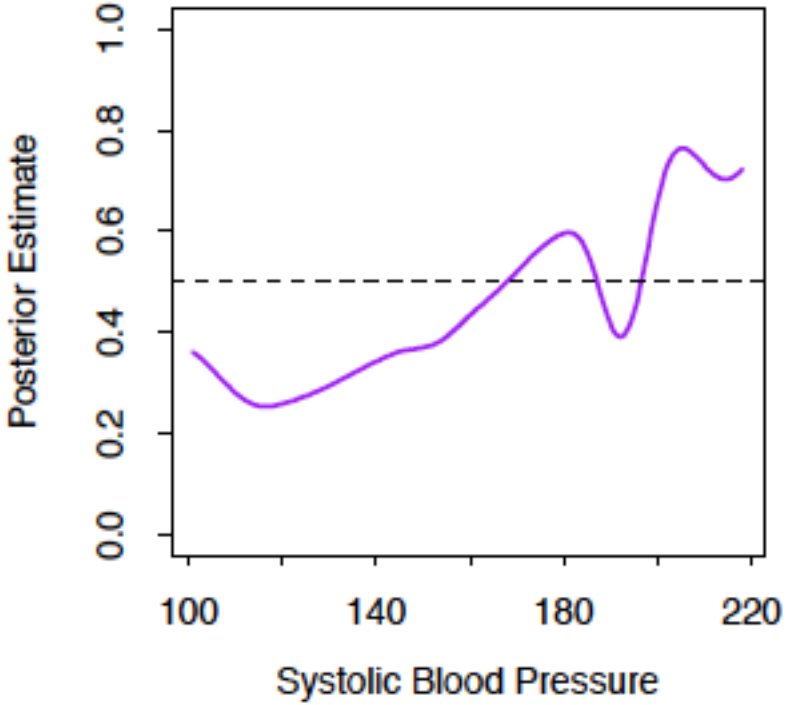
La estimación de las prior $f_k(x) = f(x \mid Y = k)$

se hace por Kernels.

Los Kernels convergen a la densidad que se desea estimar en MISE (Mean Integrated Square Error), lo que permite mostrar consistencia en los clasificadores por Kernel.

La pregunta acerca del ancho de banda óptimo está abierta. El ancho de banda óptimo para la estimación de la densidad no es necesariamente el ancho de banda óptimo para la estimación de las probabilidades.


Density Estimates
0.0
0.010
0.020
CHD
no CHD
100
140
180
220
Systolic Blood Pressure


Posterior Estimate
0.0
0.2
0.4
0.6
0.8
1.0
100
140
180
220
Systolic Blood Pressure