Algunas técnicas de análisis multivariado a través del tiempo

16 de marzo 2005, Johan Van Horebeek, Cimat

Como intro al curso de Prof. Yoon Lee, revisamos:

- Análisis de componentes principales
- Clasificación:

  LDA, Clasificador Bayesiano Optimo, Regresión logística

Algunas técnicas de análisis multivariado a través del tiempo

16 de marzo 2005, Johan Van Horebeek, Cimat

Como intro al curso de Prof. Yoon Lee, revisamos:

- Análisis de componentes principales
- Clasificación:

  LDA, Clasificador Bayesiano Optimo, Regresión logística

Punto de partida: matriz de datos multivariados $\mathbb{X}$:

$\leftarrow d$ **columnas (características)** $\rightarrow$

$\uparrow$

n **reglones**

(observaciones)

$\downarrow$

$$\begin{pmatrix} \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{pmatrix}$$

1. Análisis de Componentes Principales

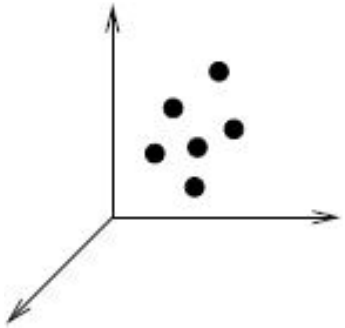
Reducir la dimensión:

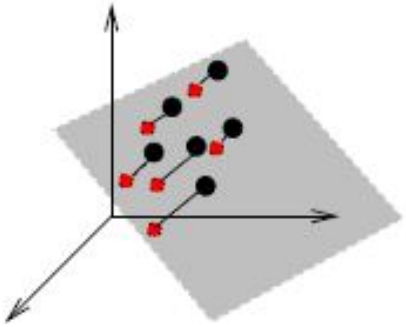
1. Análisis de Componentes Principales

Reducir la dimensión:

No:

sino:



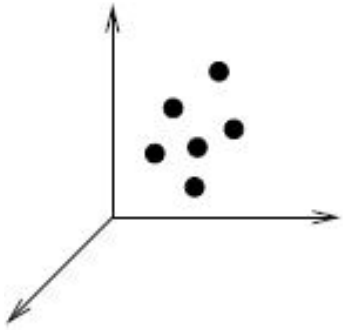


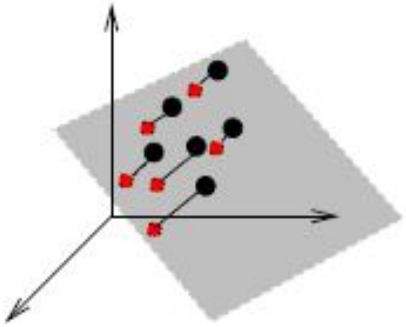
1. Análisis de Componentes Principales

Reducir la dimensión:

No:

sino:



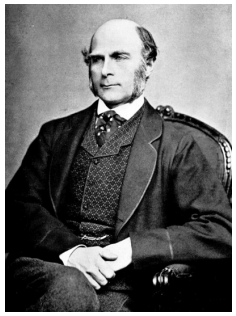


Usado en visualización, compresión, preproceso en general

Ligado con métodos de agrupamiento y selección de variables.

Reducir la dimensión tiene que ver con extracción de información util

1.1 Orígenes



NATURAL INHERITANCE

BY

FRANCIS GALTON, F.R.S.

AUTHOR OF

"HEREDITARY GENIUS," "INQUIRIES INTO HUMAN FACULTY," ETC.

1889

LIII. *On Lines and Planes of Closest Fit to Systems of Points in Space.* By KARL PEARSON, *F.R.S., University College, London* *.

1903, Philosophical Magazine

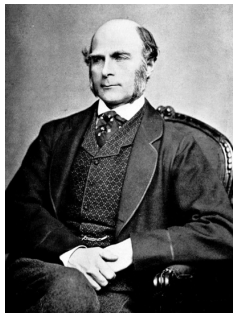
ANALYSIS OF A COMPLEX OF STATISTICAL VARIABLES INTO PRINCIPAL COMPONENTS[1]

HAROLD HOTELLING

Columbia University

1933, J. of Educactional Psychology

1.1 Orígenes



NATURAL INHERITANCE

BY

FRANCIS GALTON, F.R.S.

AUTHOR OF

"HEREDITARY GENIUS," "INQUIRIES INTO HUMAN FACULTY," ETC.

1889

ANTHROPOLOGICAL MISCELLANEA.

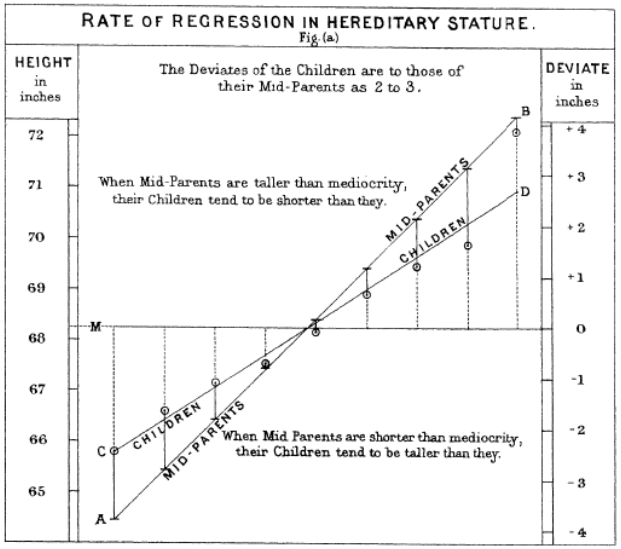
REGRESSION *towards* MEDIOCRITY *in* HEREDITARY STATURE.

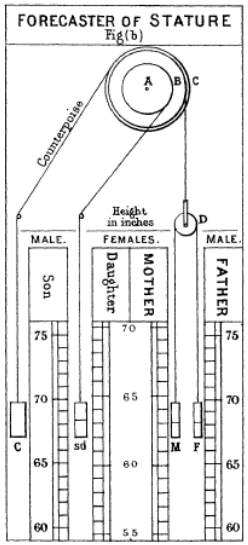
By FRANCIS GALTON, F.R.S., &c.

1887

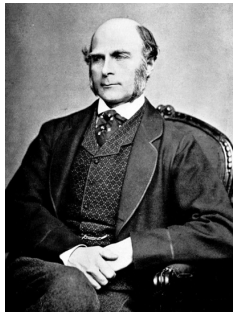
Interés en estudiar la variabilidad entre características físicas individuales sobre generaciones.

$\Rightarrow$ **Modelo de regresión:** $Y = \alpha + \beta X + \epsilon,\ \epsilon \sim \mathcal{N}(0, \sigma^2)$ $\qquad E(Y|X = x) = \alpha + \beta x$





1.1 Orígenes



NATURAL INHERITANCE

BY

FRANCIS GALTON, F.R.S.

AUTHOR OF

"HEREDITARY GENIUS," "INQUIRIES INTO HUMAN FACULTY," ETC.

1889

ANTHROPOLOGICAL MISCELLANEA.

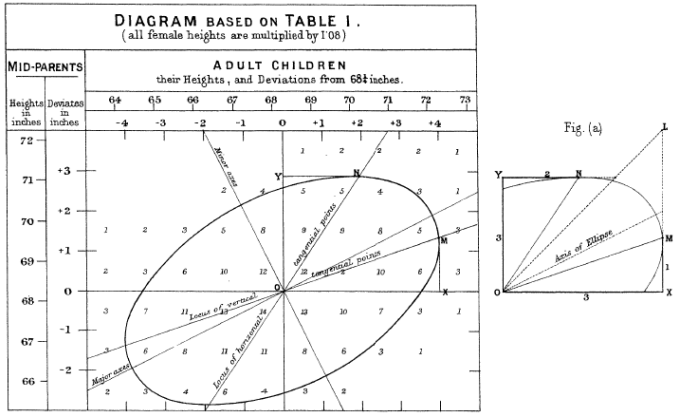
REGRESSION *towards* MEDIOCRITY *in* HEREDITARY STATURE.
By FRANCIS GALTON, F.R.S., &c.

1887

Interés en estudiar la variabilidad entre características físicas individuales sobre generaciones.

$\Rightarrow$ **Modelo de regresión:** $Y = \alpha + \beta X + \epsilon,\ \epsilon \sim \mathcal{N}(0, \sigma^2)$ $\qquad E(Y|X = x) = \alpha + \beta x$

$\Rightarrow$ **Ajustar elipse a los datos:**

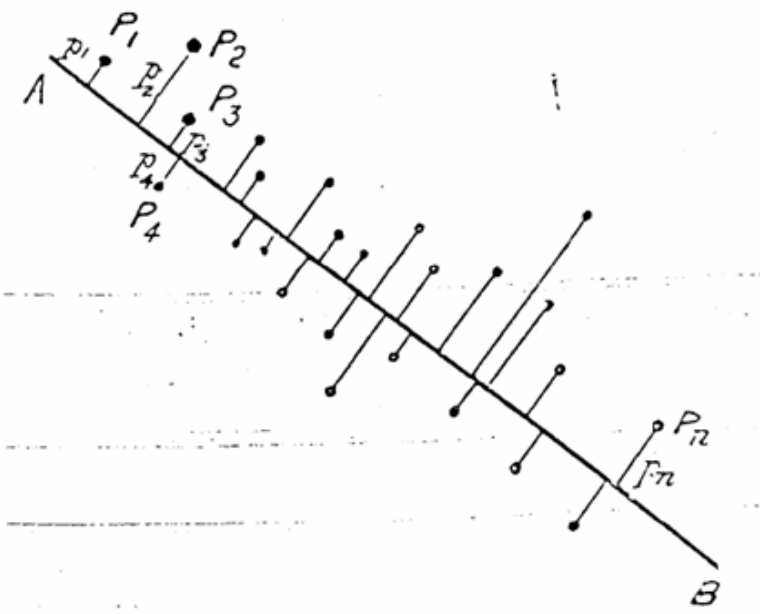


1.1 Orígenes

LIII. *On Lines and Planes of Closest Fit to Systems of Points in Space.* By KARL PEARSON, *F.R.S., University College, London* *.

1903, Philosophical Magazine

⇒ **Representar (visualizar) relación entre variables de manera simétrica:**



Minimizar errores de proyección.

1.1 Orígenes

ON CRIMINAL ANTHROPOMETRY AND THE IDENTIFICATION OF CRIMINALS.

By W. R. MACDONELL, M.A., LL.D.

[Received November 5, 1901.]

1.1 Orígenes

ON CRIMINAL ANTHROPOMETRY AND THE IDENTIFICATION OF CRIMINALS.

By W. R. MACDONELL, M.A., LL.D.

[Received November 5, 1901.]

fuente: Tyne & Wear Archives

(i) To test to what extent the criminal classes diverge in physical characters from other classes of the community.

1.1 Orígenes

ON CRIMINAL ANTHROPOMETRY AND THE IDENTIFICATION OF CRIMINALS.

By W. R. MACDONELL, M.A., LL.D.

[Received November 5, 1901.]

(i) To test to what extent the criminal classes diverge in physical characters from other classes of the community.

⇒ **Construir indices**:

(19) Professor Pearson has pointed out to me that the ideal index characters would be given if we calculated the seven directions of uncorrelated variables, that is, the principal axes of the correlation "ellipsoid." Thus, given x_1, x_2, ... x_7 correlated variables, the seven uncorrelated variables are:

$$X_1 = l_{11}x_1 + l_{12}x_2 + \ldots + l_{17}x_7$$
$$X_2 = l_{21}x_1 + l_{22}x_2 + \ldots + l_{27}x_7$$
$$\&c. \qquad \&c.$$

where the l's give the directions of the principal axes, and X_1, X_2, ... X_7 would be the proper index functions to identify criminals by if we have nothing better than the present correlated characters to work with.

1.2 Formulación: enfoque probabilístico

Consideramos $X = (X_1, \cdots, X_d)^t$ como v.a.

Define $d \times d$ matrix de covarianze $\mathbb{C}$:

$$\mathbb{C}_{i,j} = Cov(X_i, X_j).$$

1. **$\mathbb{C}$ es simétrico, no negativo definido, con eigenvalores no negativos.**
2. **Define $Y = u^t X$, $Var(Y) = u^t \mathbb{C} u$.**

1.2 Formulación: enfoque probabilístico ($EX = 0$)

PCA como transformación (lineal) informativa

Usamos la varianza para medida qué informativa es una transformación; caso 1-D:

$$\max_{||l||=1} Var(l^t X) = \max_{||l||=1} l^t Cov(X) l$$

1.2 Formulación: enfoque probabilístico ($EX = 0$)

PCA como transformación (lineal) informativa

Usamos la varianza para medida qué informativa es una transformación; caso 1-D:

$$\max_{||l||=1} Var(l^t X) = \max_{||l||=1} l^t Cov(X) l$$

Solución: l **es el primer vector propio de** $Cov(X)$

1.2 Formulación: enfoque probabilístico ($EX = 0$)

PCA como transformación (lineal) informativa

Usamos la varianza para medida qué informativa es una transformación; caso 1-D:

$$\max_{||l||=1} Var(l^t X) = \max_{||l||=1} l^t Cov(X) l$$

Solución: l **es el primer vector propio de** $Cov(X)$

En general, buscamos direcciones $\{l_i\}_1^p$ **tal que**

$$\max_{||l_i||=1} Var(l_i^t X), \qquad l_i \perp l_1, \cdots, l_{i-1}$$

Solución: $\{l_i\}$ **son los primeros** p **vectores propios de** $Cov(X)$**.**

Llamamos $Y_i = l_i^t X$**; mapeamos** $(X_1, \cdots, X_d)$ **a** $(Y_1, \cdots, Y_p)$

1.2 Formulación: enfoque probabilístico ($EX = 0$)

PCA como transformación (lineal) informativa

Usamos la varianza para medida qué informativa es una transformación; caso 1-D:

$$\max_{||l||=1} Var(l^t X) = \max_{||l||=1} l^t Cov(X) l$$

Solución: l **es el primer vector propio de** $Cov(X)$

En general, buscamos direcciones $\{l_i\}_1^p$ **tal que**

$$\max_{||l_i||=1} Var(l_i^t X), \qquad l_i \perp l_1, \cdots, l_{i-1}$$

Solución: $\{l_i\}$ **son los primeros** p **vectores propios de** $Cov(X)$**.**

Llamamos $Y_i = l_i^t X$**; mapeamos** $(X_1, \cdots, X_d)$ **a** $(Y_1, \cdots, Y_p)$

Propiedades

$Var(Y_i) = \lambda_i$

$Cov(Y_i, Y_j) = 0$ **si** $i \neq j$

$\sum_i Var(X_i) = \sum_i Var(Y_i) = \sum_i \lambda_i$

PCA como predicción óptima en subespacio

Busca transformaciones lineales $X \Longrightarrow Y \Longrightarrow \hat{X}$

con $dim(Y) = p < d$ **tal que** $E||X - \hat{X}||^2$ **es mínima.**

PCA como predicción óptima en subespacio

Busca transformaciones lineales $X \Longrightarrow Y \Longrightarrow \hat{X}$

con $dim(Y) = p < d$ **tal que** $E||X - \hat{X}||^2$ **es mínima.**

- **Si** $p = 1$**:** $X\ (\in \mathcal{R}^d) \quad \Longrightarrow \quad Y = \langle l, X \rangle\ (\in \mathcal{R}^1) \quad \Longrightarrow \quad \hat{X} = Yl\ (\in \mathcal{R}^d)$

  tal que $E||X - \hat{X}||^2$ **sea mínima**.

- **En general: busca una matriz** $\mathbb{B}$ $(d \times p)$ **que minimiza:** $E||X - \hat{X}||^2$ **donde** $\hat{X} = \mathbb{B}(\mathbb{B}^t X)$

  Solución: $\mathbb{B}$ **formada por primeros** p **vectores propios de** $Cov(X)$**.**

PCA como predicción óptima en subespacio

Busca transformaciones lineales $X \Longrightarrow Y \Longrightarrow \hat{X}$

con $dim(Y) = p < d$ **tal que** $E||X - \hat{X}||^2$ **es mínima.**

- **Si** $p = 1$**:** $X\ (\in \mathcal{R}^d) \quad \Longrightarrow \quad Y = \langle l, X \rangle\ (\in \mathcal{R}^1) \quad \Longrightarrow \quad \hat{X} = Yl\ (\in \mathcal{R}^d)$

  tal que $E||X - \hat{X}||^2$ **sea mínima**.

- **En general: busca una matriz** $\mathbb{B}$ $(d \times p)$ **que minimiza:** $E||X - \hat{X}||^2$ **donde** $\hat{X} = \mathbb{B}(\mathbb{B}^t X)$

  Solución: $\mathbb{B}$ **formada por primeros** p **vectores propios de** $Cov(X)$**.**

Propiedad

Si $\mathbb{B}$ **está formada por los primeros** p **vectores propios de** $Cov(X)$ **y** $EX = 0$**:**

$E||X - \mathbb{B}(\mathbb{B}^t X)||^2 = \sum_{i=p+1}^{d} \lambda_i$**.**

¿Qué bueno es limitarse a aproximaciones lineales?

¿Qué bueno es limitarse a aproximaciones lineales?

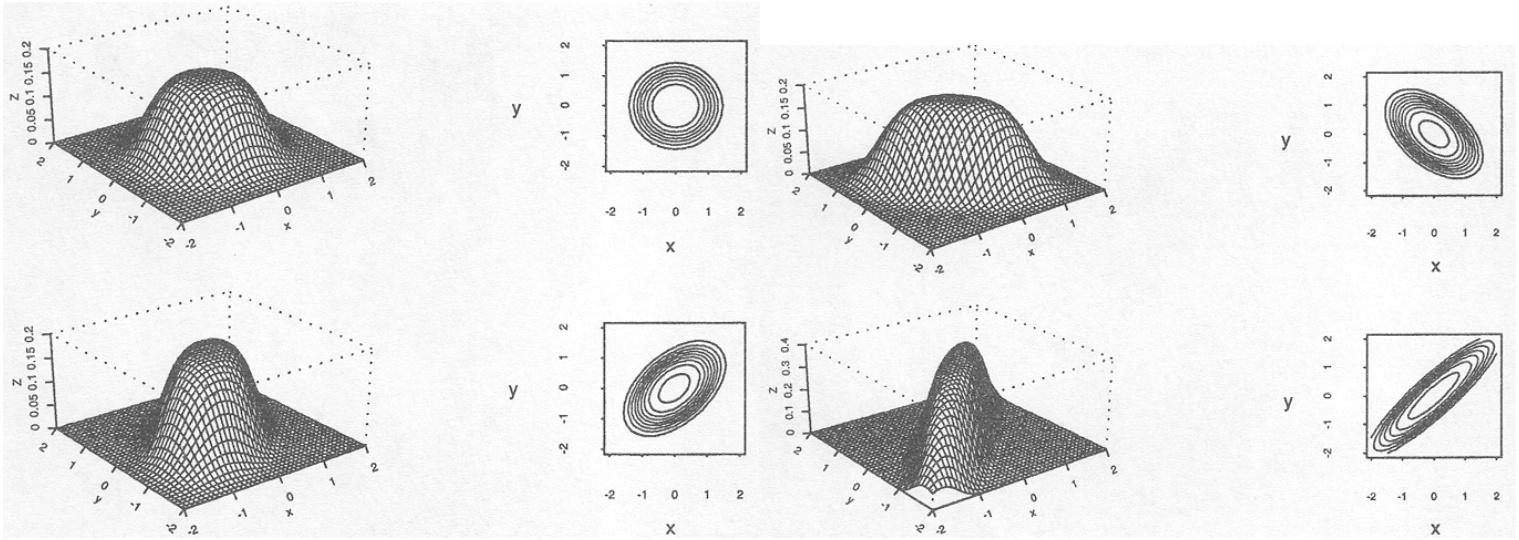
$\Rightarrow$ Depende de la distribución de los datos

¿Qué bueno es limitarse a aproximaciones lineales?

$\Rightarrow$ Depende de la distribución de los datos

Caso óptimo: X proviene de una multivariada gausiana $\mathcal{N}(\mu, \Sigma)$

$$f_{X_1,\cdots,X_d}(x_1, \cdots, x_d) = \frac{1}{(2\pi)^{d/2}\sqrt{|\Sigma|}} \exp \frac{-(x-\mu)^t \Sigma^{-1}(x-\mu)}{2},$$

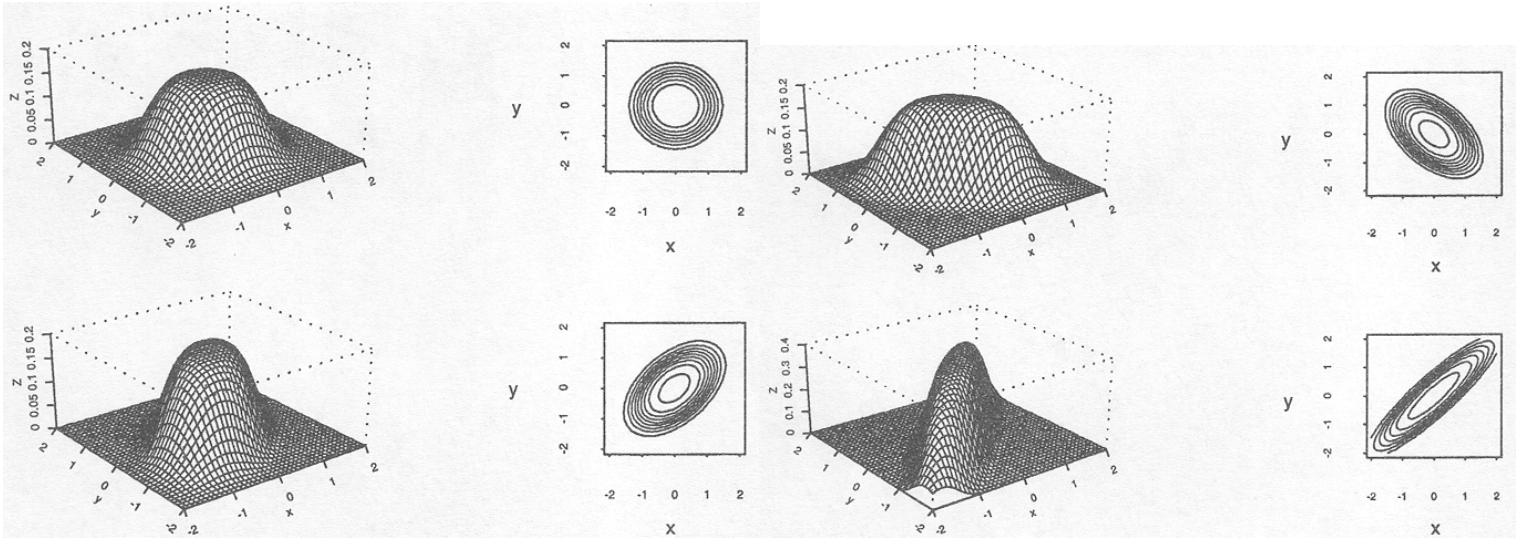


¿Qué bueno es limitarse a aproximaciones lineales?

⇒ Depende de la distribución de los datos

Caso óptimo: X proviene de una multivariada gausiana $\mathcal{N}(\mu, \Sigma)$

$$f_{X_1,\cdots,X_d}(x_1,\cdots,x_d) = \frac{1}{(2\pi)^{d/2}\sqrt{|\Sigma|}} \exp \frac{-(x-\mu)^t\Sigma^{-1}(x-\mu)}{2},$$



Propiedades

Distribuciones marginales y condicionales son también normales y $E(X_A|X_B = x_B)$ es una función lineal de x_B.

1.3 Formulación: enfoque matricial

Se busca una aproximación de bajo rango a una matríz (centrada):

$$min||\mathbb{X} - \hat{\mathbb{X}}|| \textbf{ sujeto a } rank(\hat{\mathbb{X}}) = p < d$$

1.3 Formulación: enfoque matricial

Se busca una aproximación de bajo rango a una matríz (centrada):

$$min||\mathbb{X} - \hat{\mathbb{X}}|| \text{ sujeto a } rank(\hat{\mathbb{X}}) = p < d$$

Recordatorio: SVD

Sea $\mathbb{A}$ matriz $n \times d$. Existen matrices $\mathbb{U}$ $(d \times r)$, $\mathbb{V}$ $(n \times r)$ y $\mathbb{D}(r \times r)$ tal que

$$\mathbb{A} = \mathbb{V}\mathbb{D}\mathbb{U}^t = \sum_{i=1}^{r} \sigma_i v_i u_i^t$$

columnas de $\mathbb{U}$ son vectores propios $\{u_i\}$ de $\mathbb{A}^t\mathbb{A}$,

columnas de $\mathbb{V}$ son vectores propios $\{v_i\}$ de $\mathbb{A}\mathbb{A}^t$,

$D = Diag(\{\sigma_i\})$ **,** $\sigma_i^2 = \lambda_i$**, con** λ_i **valor propio de** u_i **(y** v_i**).**

Además $v_i = \mathbb{A}u_i/\sqrt{\lambda_i}$**,** $u_i = \mathbb{A}^t v_i/\sqrt{\lambda_i}$

1.3 Formulación: enfoque matricial

Se busca una aproximación de bajo rango a una matríz (centrada):

$$min||\mathbb{X} - \hat{\mathbb{X}}|| \text{ sujeto a } rank(\hat{\mathbb{X}}) = p < d$$

Recordatorio: SVD

Sea $\mathbb{A}$ matriz $n \times d$. Existen matrices $\mathbb{U}$ $(d \times r)$, $\mathbb{V}$ $(n \times r)$ y $\mathbb{D}(r \times r)$ tal que

$$\mathbb{A} = \mathbb{V}\mathbb{D}\mathbb{U}^t = \sum_{i=1}^{r} \sigma_i v_i u_i^t$$

columnas de $\mathbb{U}$ son vectores propios $\{u_i\}$ de $\mathbb{A}^t\mathbb{A}$,

columnas de $\mathbb{V}$ son vectores propios $\{v_i\}$ de $\mathbb{A}\mathbb{A}^t$,

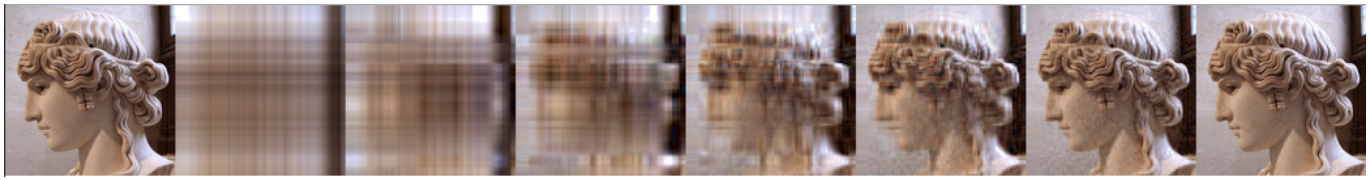
$D = Diag(\{\sigma_i\})$ **,** $\sigma_i^2 = \lambda_i$**, con** λ_i **valor propio de** u_i **(y** v_i**).**

Además $v_i = \mathbb{A}u_i/\sqrt{\lambda_i}$**,** $u_i = \mathbb{A}^t v_i/\sqrt{\lambda_i}$

Solución:

$\hat{\mathbb{X}} = \mathbb{V}(p)\mathbb{D}(p)\mathbb{U}(p)^t = \sum_{i=1}^{p} \sigma_i v_i u_i^t$ **con**

$$\mathbb{V}(p) = \mathbb{V}[, 1:p],\ \mathbb{U}(p) = \mathbb{U}[, 1:p]\ , \mathbb{D}(p) = \mathbb{D}[1:p, 1:p] \text{ de la SVD de } \mathbb{X}.$$



original (256×256)**, approximación con rango=1,2,4,8,16,32,64. fuente: wikipedia**

1.3 Formulación: enfoque matricial

Se busca una aproximación de bajo rango a una matríz (centrada):

$$min||\mathbb{X} - \hat{\mathbb{X}}|| \text{ sujeto a } rank(\hat{\mathbb{X}}) = p < d$$

Recordatorio: SVD

Sea $\mathbb{A}$ matriz $n \times d$. Existen matrices $\mathbb{U}$ $(d \times r)$, $\mathbb{V}$ $(n \times r)$ y $\mathbb{D}(r \times r)$ tal que

$$\mathbb{A} = \mathbb{V}\mathbb{D}\mathbb{U}^t = \sum_{i=1}^{r} \sigma_i v_i u_i^t$$

columnas de $\mathbb{U}$ son vectores propios $\{u_i\}$ de $\mathbb{A}^t\mathbb{A}$,

columnas de $\mathbb{V}$ son vectores propios $\{v_i\}$ de $\mathbb{A}\mathbb{A}^t$,

$D = Diag(\{\sigma_i\})$ **, $\sigma_i^2 = \lambda_i$, con λ_i valor propio de u_i (y v_i).**

Además $v_i = \mathbb{A}u_i/\sqrt{\lambda_i}$, $u_i = \mathbb{A}^t v_i/\sqrt{\lambda_i}$

Solución:

$\hat{\mathbb{X}} = \mathbb{V}(p)\mathbb{D}(p)\mathbb{U}(p)^t = \sum_{i=1}^{p} \sigma_i v_i u_i^t$ **con**

$$\mathbb{V}(p) = \mathbb{V}[, 1:p],\ \mathbb{U}(p) = \mathbb{U}[, 1:p]\ , \mathbb{D}(p) = \mathbb{D}[1:p, 1:p] \text{ de la SVD de } \mathbb{X}.$$

Observa: como $\mathbb{V}(p) \propto \mathbb{X}\mathbb{U}(p)$, obtenemos que $\hat{\mathbb{X}}$ es una transformación lineal de $\mathbb{X}$

Técnicamente: las soluciones de ambos enfoques coinciden al usar $\widehat{Cov}(X) \propto \mathbb{X}^t\mathbb{X}$

Dualidad entre trabajar con $\mathbb{X}\mathbb{X}^t$ y $\mathbb{X}^t\mathbb{X}$ y entre direcciones y proyecciones.

Si $\widehat{Cov}(X) \propto \mathbb{X}^t\mathbb{X}$:

Dualidad entre trabajar con $\mathbb{X}\mathbb{X}^t$ y $\mathbb{X}^t\mathbb{X}$:

$u_i = \mathbb{X}^t v_i/\sqrt{\lambda_i}$, **con** $\{v_i\}$ **vectores propios de** $\mathbb{X}\mathbb{X}^t$

$x^t u_i = x\mathbb{X}^t v_i/\sqrt{\lambda_i}$

Basta conocer productos puntos entre observaciones (transformadas).

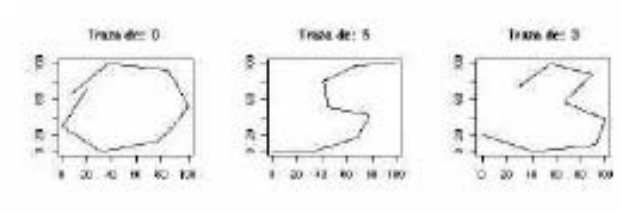
Forma la base de muchas extensiones como kernel PCA.

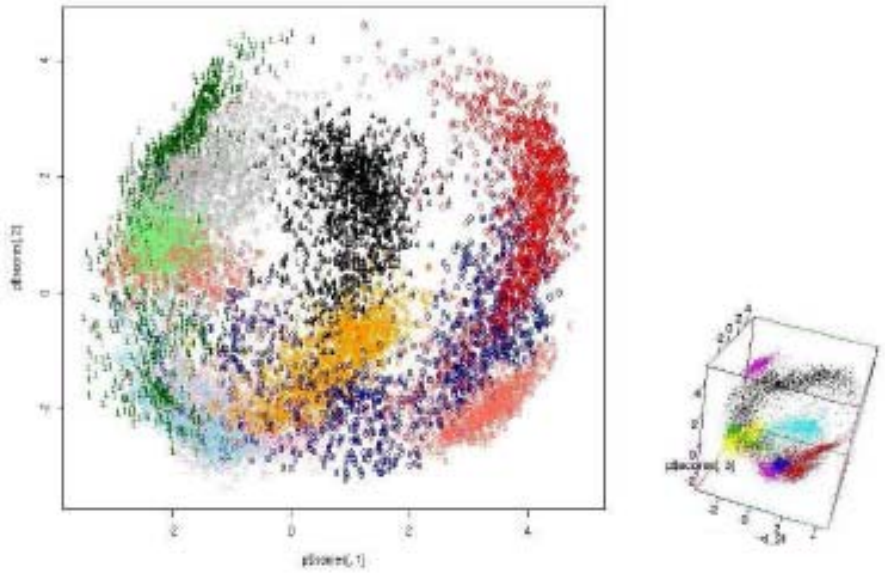
1.3 Ejemplos

Ejemplo 1

Cada dato es la traza de un dígito escrito a mano.

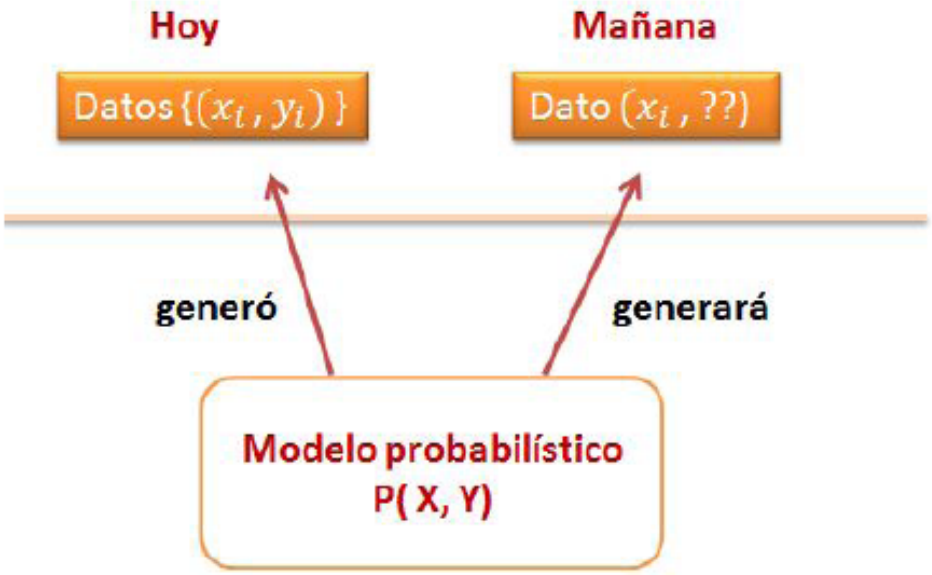
La traza está discretizada en 8 puntos $\{(x(t_i), y(t_i)))\}_{t=1}^{8}$





2. Clasificación

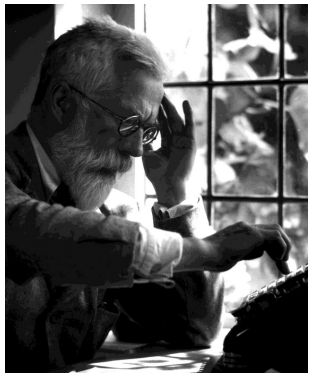
Sea X un vector de características y Y una variable de respuesta (categoría)



Diferentes enfoques con diferentes preguntas de interés.

2.1 Origin

Fisher (1936): discriminar



THE USE OF MULTIPLE MEASUREMENTS IN TAXONOMIC PROBLEMS

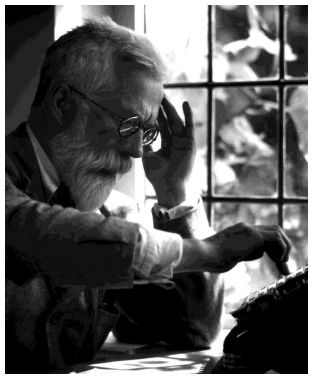
By R. A. FISHER, Sc.D., F.R.S.

I. Discriminant functions

When two or more populations have been measured in several characters, $x_1, \ldots, x_s$, special interest attaches to certain linear functions of the measurements by which the populations are best discriminated.

2.1 Origin

Fisher (1936): discriminar



THE USE OF MULTIPLE MEASUREMENTS IN TAXONOMIC PROBLEMS

By R. A. FISHER, Sc.D., F.R.S.

I. Discriminant functions

When two or more populations have been measured in several characters, $x_1, \ldots, x_s$, special interest attaches to certain linear functions of the measurements by which the populations are best discriminated.

We shall first consider the question: What linear function of the four measurements

$$X = \lambda_1 x_1 + \lambda_2 x_2 + \lambda_3 x_3 + \lambda_4 x_4$$

will maximize the ratio of the difference between the specific means to the standard deviations within species? The observed means and their differences are shown in Table II.

	Petal ancho	Petal largo	Sepal ancho	Sepal largo
1				
2				
3				
4				
5				
6				
7				
8				
9				
10				
11				
12				
13				
14				
15				

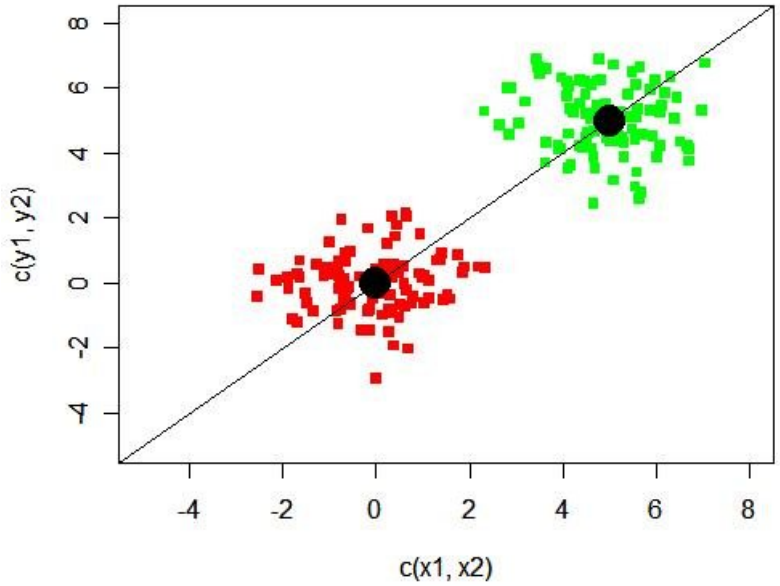
Familia Setosa

Familia Versicolor

??

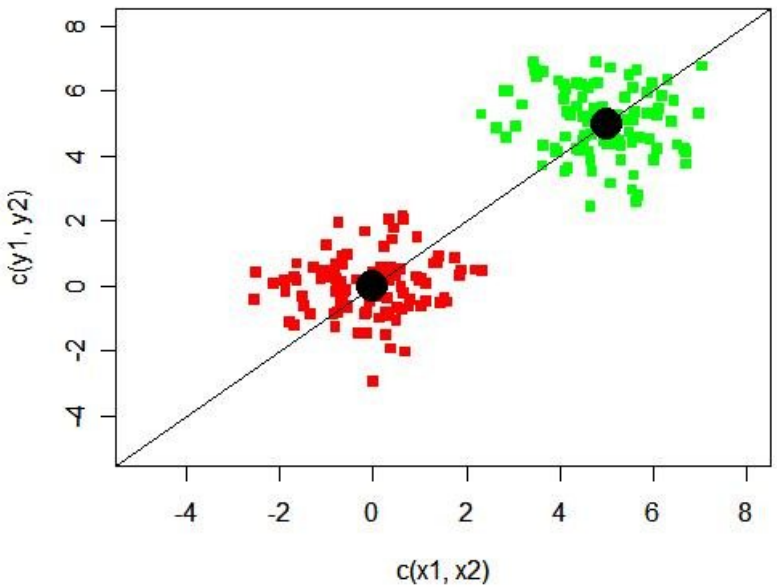
Análisis Discrimimante Lineal

Punto de partida: **¿En qué dirección proyectar los datos para separar los puntos de diferentes clases lo más posible?**



Análisis Discrimimante Lineal

Punto de partida: **¿En qué dirección proyectar los datos para separar los puntos de diferentes clases lo más posible?**



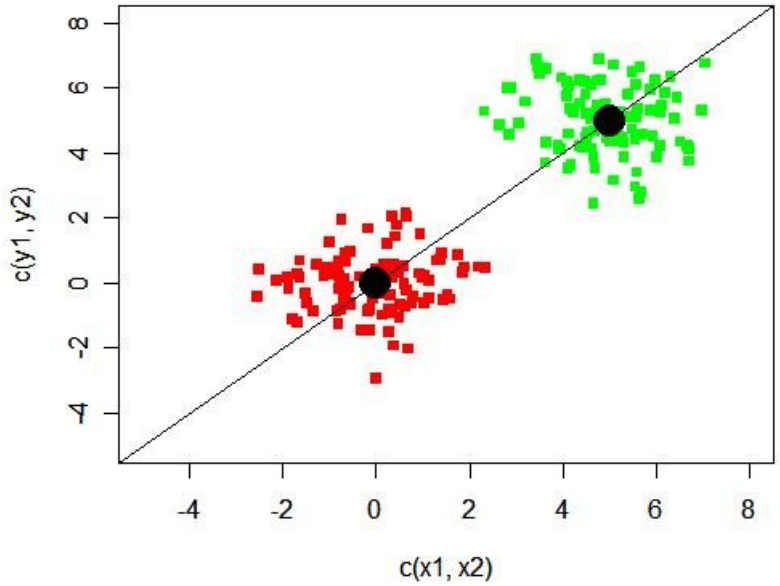
Define c_+, c_- como los centroides de cada clase.

Primera idea: Separamos las proyecciones de los centroides lo más posible:

$$\arg\max_{l}(l^t(c_+ - c_-))^2$$

Análisis Discrimimante Lineal

Punto de partida: **¿En qué dirección proyectar los datos para separar los puntos de diferentes clases lo más posible?**



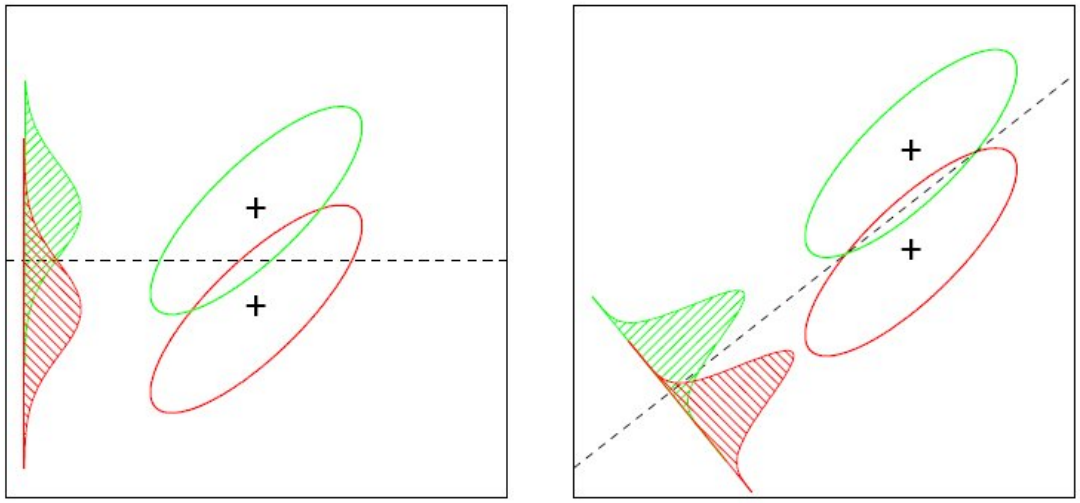
Define c_+, c_- como los centroides de cada clase.

Primera idea: Separamos las proyecciones de los centroides lo más posible:

$$\arg\max_{l}(l^t(c_+ - c_-))^2$$

Problema:

hay que tomar en cuenta la estructura de la covarianza



Vamos a suponer que la estructura de covarianza de ambos clases es igual.

Antes:

$$\arg\max_l (l^t(c_+ - c_-))^2$$

Ahora:

$$\arg\max_l \left(\frac{(l^t(c_+ - c_-))}{\sqrt{Var(l^t X|Y)}}\right)^2 = \frac{l^t(c_+ - c_-)(c_+ - c_-)^t l}{l^t S_W l} := \frac{l^t S_B l}{l^t S_W l}$$

Solución: $l \propto S_W^{-1}(c_+ - c_-)$

2.2 Teoría de decisión óptima

Años 50 (Wald, Chow, Anderson, ...)

Dado un clasificador $\hat{y}(x)$, define una función de perdida $L(y, \hat{y}(x))$.

Buscamos

$$\arg\min_{\hat{y}(\cdot)} E(L(Y, \hat{y}(X)))$$

2.2 Teoría de decisión óptima

Años 50 (Wald, Chow, Anderson, ...)

Dado un clasificador $\hat{y}(x)$, define una función de perdida $L(y, \hat{y}(x))$.

Buscamos

$$\arg\min_{\hat{y}(\cdot)} E(L(Y, \hat{y}(X)))$$

Ejemplos: $L(y, \hat{y}(x)) = I(y \neq \hat{y}(x)), \quad L(y, f(x)) = (y - \hat{y}(x))^2, \quad L(y, f(x)) = |y - \hat{y}(x)|$

2.2 Teoría de decisión óptima

Años 50 (Wald, Chow, Anderson, ...)

Dado un clasificador $\hat{y}(x)$, define una función de perdida $L(y, \hat{y}(x))$.

Buscamos

$$\arg\min_{\hat{y}(\cdot)} E(L(Y, \hat{y}(X)))$$

Ejemplos: $L(y, \hat{y}(x)) = I(y \neq \hat{y}(x)), \quad L(y, f(x)) = (y - \hat{y}(x))^2, \quad L(y, f(x)) = |y - \hat{y}(x)|$

Observa:

$$E_{X,Y} L(Y, \hat{y}(X)) = E_X(E_{Y|X} L(Y, \hat{y}(X))) = \int E_{Y|X=x} L(Y, \hat{y}(X)) dF_X(x)$$

Si minimizamos lo anterior sobre $\hat{y}(x)$, es suficiente para cada x calcular:

$\arg\min_{\hat{y}(x)} E_{Y|X=x} L(Y, \hat{y}(x))$, **solución es clasificador Bayesiano óptimo**

2.2 Teoría de decisión óptima

Años 50 (Wald, Chow, Anderson, ...)

Dado un clasificador $\hat{y}(x)$, define una función de perdida $L(y, \hat{y}(x))$.

Buscamos

$$\arg\min_{\hat{y}(\cdot)} E(L(Y, \hat{y}(X)))$$

Ejemplos: $L(y, \hat{y}(x)) = I(y \neq \hat{y}(x)), \quad L(y, f(x)) = (y - \hat{y}(x))^2, \quad L(y, f(x)) = |y - \hat{y}(x)|$

Observa:

$$E_{X,Y} L(Y, \hat{y}(X)) = E_X(E_{Y|X} L(Y, \hat{y}(X))) = \int E_{Y|X=x} L(Y, \hat{y}(X)) dF_X(x)$$

Si minimizamos lo anterior sobre $\hat{y}(x)$, es suficiente para cada x calcular:

$\arg\min_{\hat{y}(x)} E_{Y|X=x} L(Y, \hat{y}(x))$, **solución es clasificador Bayesiano óptimo**

Ejemplo: Si Y toma solamente dos valores:

$$E_{Y|X=x} L(Y, \hat{y}(x)) = L(0, \hat{y}(x)) P(Y = 0|X = x) + L(1, \hat{y}(x)) P(Y = 1|X = x)$$

$$E_{Y|X=x}L(Y,\hat{y}(x)) = L(0,\hat{y}(x))P(Y=0|X=x) + L(1,\hat{y}(x))P(Y=1|X=x)$$

Toma caso binario y $L(y,\hat{y}(x)) = I(y \neq \hat{y}(x))$**:**

si $\hat{y}(x) = 0 : L(0,\hat{y}(x)) = 0, \;\; L(1,\hat{y}(x)) = 1$ **y el error es** $P(Y=1|X=x)$

si $\hat{y}(x) = 1 : L(0,\hat{y}(x)) = 1, \;\; L(1,\hat{y}(x)) = 0$ **y el error es** $P(Y=0|X=x)$

$$E_{Y|X=x}L(Y,\hat{y}(x)) = L(0,\hat{y}(x))P(Y=0|X=x) + L(1,\hat{y}(x))P(Y=1|X=x)$$

Toma caso binario y $L(y,\hat{y}(x)) = I(y \neq \hat{y}(x))$**:**

si $\hat{y}(x) = 0 : L(0,\hat{y}(x)) = 0, \;\; L(1,\hat{y}(x)) = 1$ **y el error es** $P(Y=1|X=x)$

si $\hat{y}(x) = 1 : L(0,\hat{y}(x)) = 1, \;\; L(1,\hat{y}(x)) = 0$ **y el error es** $P(Y=0|X=x)$

Así el clasificador óptimo es

$$\hat{y}^*(x) = \begin{cases} 0 & \textbf{si } P(Y=0|X=x) > P(Y=1|X=x). \\ 1 & \textbf{si } P(Y=1|X=x) \geq P(Y=0|X=x) \end{cases}$$

En general: asigna x **a la categoría más probable según** $P(Y|X=x)$**.**

Define $L^* = E(L(Y,\hat{y}^*(X)))$**.**

$$E_{Y|X=x}L(Y,\hat{y}(x)) = L(0,\hat{y}(x))P(Y=0|X=x) + L(1,\hat{y}(x))P(Y=1|X=x)$$

Toma caso binario y $L(y,\hat{y}(x)) = I(y \neq \hat{y}(x))$**:**

si $\hat{y}(x) = 0 : L(0,\hat{y}(x)) = 0, \quad L(1,\hat{y}(x)) = 1$ **y el error es** $P(Y=1|X=x)$

si $\hat{y}(x) = 1 : L(0,\hat{y}(x)) = 1, \quad L(1,\hat{y}(x)) = 0$ **y el error es** $P(Y=0|X=x)$

Así el clasificador óptimo es

$$\hat{y}^*(x) = \begin{cases} 0 & \textbf{si } P(Y=0|X=x) > P(Y=1|X=x). \\ 1 & \textbf{si } P(Y=1|X=x) \geq P(Y=0|X=x) \end{cases}$$

En general: asigna x **a la categoría más probable según** $P(Y|X=x)$**.**

Define $L^* = E(L(Y,\hat{y}^*(X)))$**.**

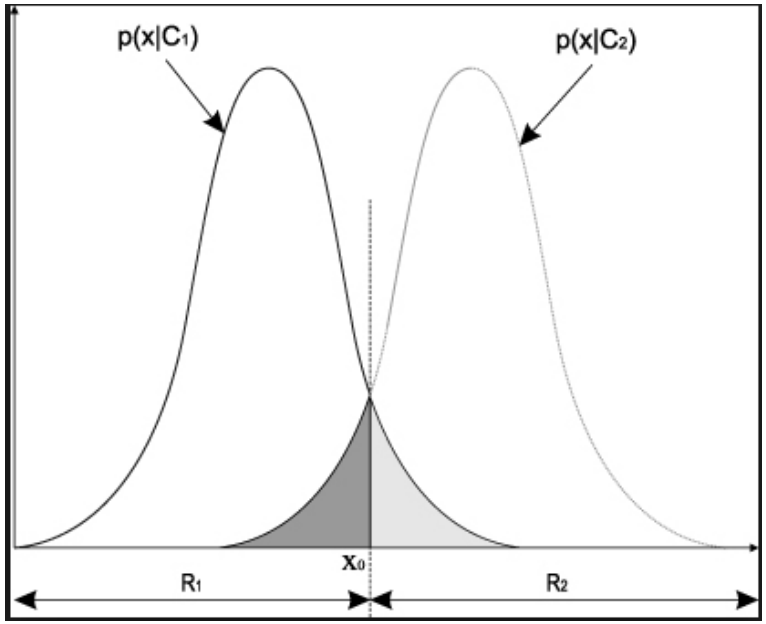
En general, la clasificación será basada en la pregunta

$$\frac{P(Y=1|X=x)}{P(Y=0|X=x)} \overset{?}{>} c \qquad \frac{P(X=x|Y=1)}{P(X=x|Y=0)} \overset{?}{>} constante$$

$$\frac{P(Y=1|X=x)}{P(Y=0|X=x)} \stackrel{?}{>} c \qquad \frac{P(X=x|Y=1)}{P(X=x|Y=0)} \stackrel{?}{>} constante$$

Ejemplo

$X|Y=y \sim \mathcal{N}(\mu_y, \sigma^2)$ $\quad$ $P(Y=1)=P(Y=0)$

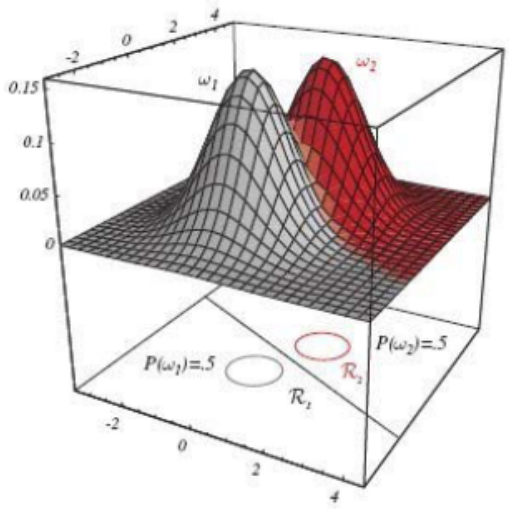


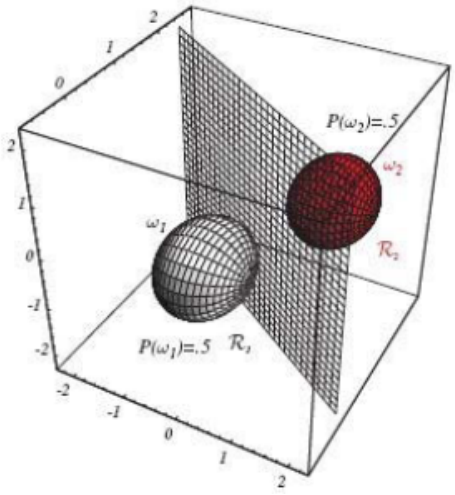
Observa: $y^*(x)$ **es de la forma** $I(x > c)$**.**

$$\frac{P(Y=1|X=x)}{P(Y=0|X=x)} \overset{?}{>} c \qquad \frac{P(X=x|Y=1)}{P(X=x|Y=0)} \overset{?}{>} constante$$

Ejemplo

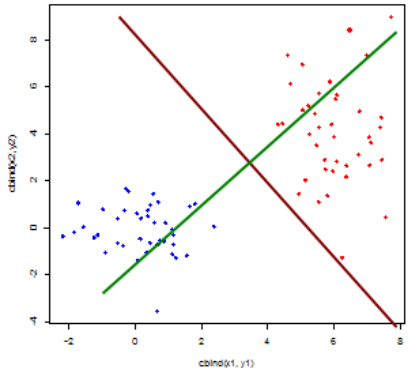
$X|Y=y \sim \mathcal{N}(\mu_y, \Sigma)$





fuente: Duda et al.

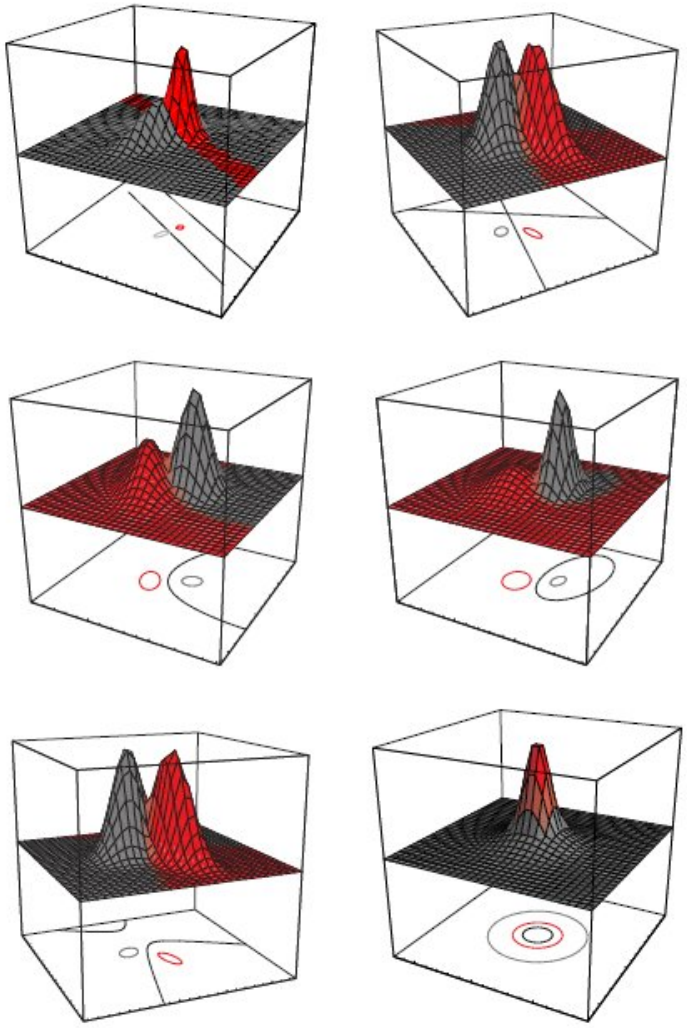
Observa: $y^*(x)$ es de la forma $I(l^t x > c)$. Corresponde a solución LDA.



$$\frac{P(Y=1|X=x)}{P(Y=0|X=x)} \overset{?}{>} c \qquad \frac{P(X=x|Y=1)}{P(X=x|Y=0)} \overset{?}{>} constante$$

Ejemplo

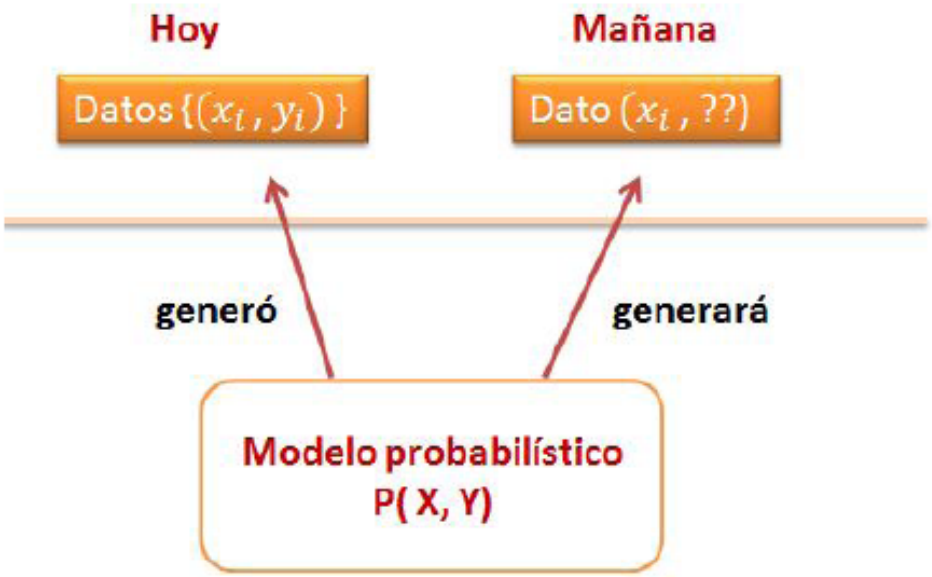
$X|Y=y \sim \mathcal{N}(\mu_y, \Sigma_y)$



fuente: Duda et al.

2.4 Enfoque noparamétrico

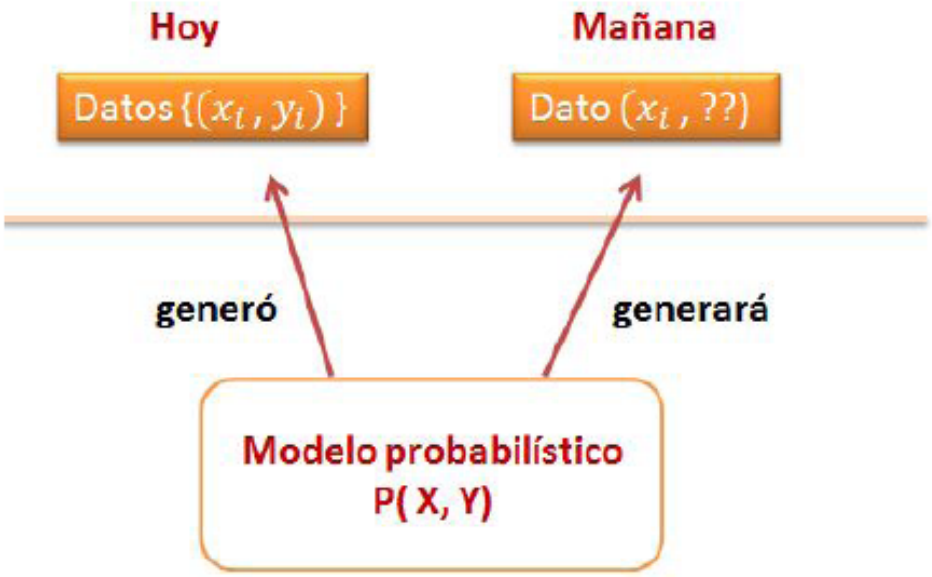
Años 70 (Stone, Vapnik, ...)



¿ Podemos estimar/aprender un clasificador Bayesiano óptimo sin limitar $P()$ a familia paramétrica?

2.4 Enfoque noparamétrico

Años 70 (Stone, Vapnik, ...)



¿ Podemos estimar/aprender un clasificador Bayesiano óptimo sin limitar $P()$ a familia paramétrica?

Teorema no free lunch:

Para n finito, sin ningun supuesto adicional sobre P,

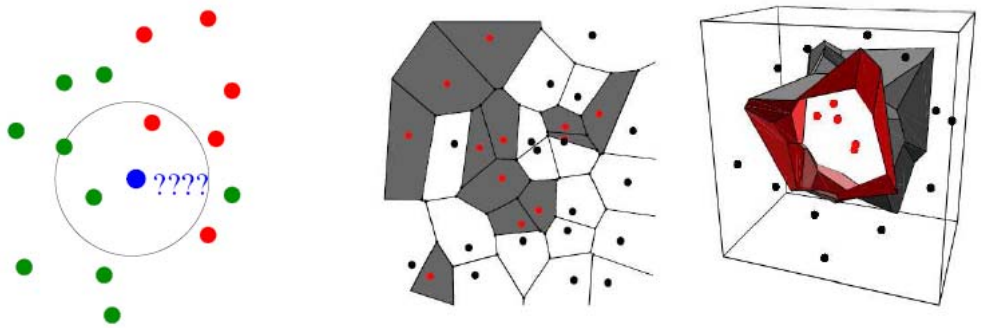
ningun clasificador es mejor que otro.

... pero si $n \to \infty$?

Dado el conjunto $\{(x_i, y_i)\}$

Para una x **fija, busca el conjunto** C **con los** k **vecinos** $\{x_i\}$ **más cercanos a** x

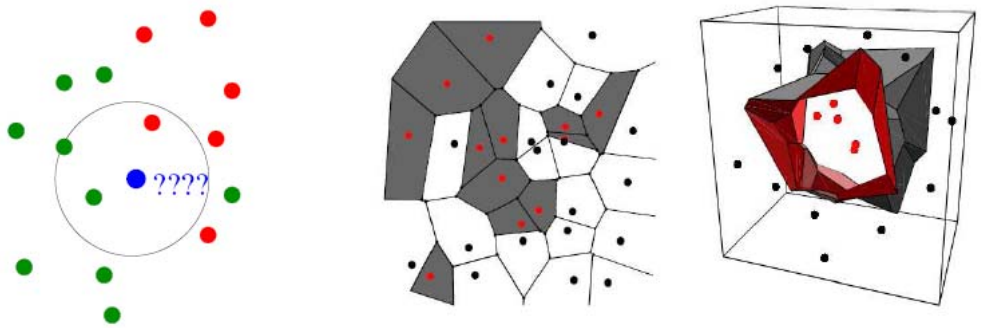
Define $\hat{y}(x)$ **como la clase que más veces ocurre en** C



Dado el conjunto $\{(x_i, y_i)\}$

Para una x **fija, busca el conjunto** C **con los** k **vecinos** $\{x_i\}$ **más cercanos a** x

Define $\hat{y}(x)$ **como la clase que más veces ocurre en** C



Si denotamos con L^* **el error del clasificador Bayesiano óptimo, y** $\hat{y}_n$ **un clasificador basado en** $\{(X_i, Y_i)\}_1^n$ **y** $L(\hat{y}_n) = E(L(Y, \hat{y}_n(X))$**, se puede mostrar:**

Propiedad 1 Si $n \to \infty$ **y** $\hat{y}_n$ **es el 1-NN:**

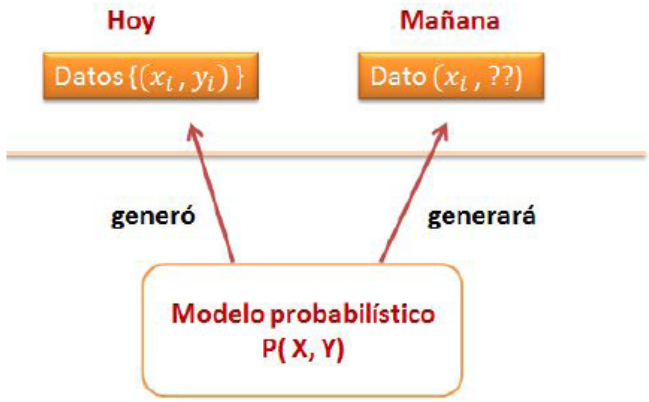
$$L^* \leq EL(\hat{y}_n) \leq 2 * L^*$$

Propiedad 2 (1977) Si $n \to \infty$ **y** $k \to \infty$ **tal que** $k/n \to 0$**, si** $\hat{y}_n$ **es el k-NN:**
para cualquier P**:**

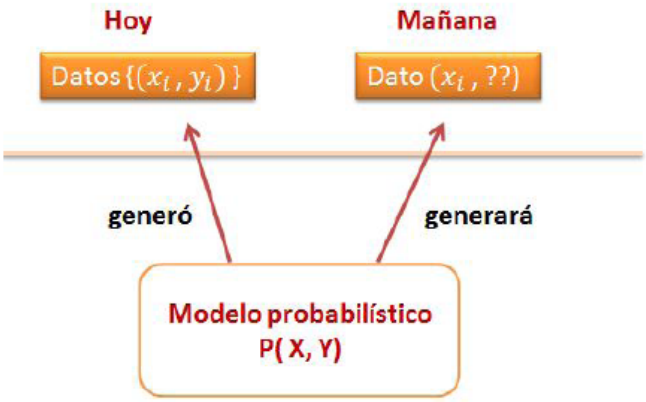
$$EL(\hat{y}_n) \to L^*$$

Donde nos quedamos

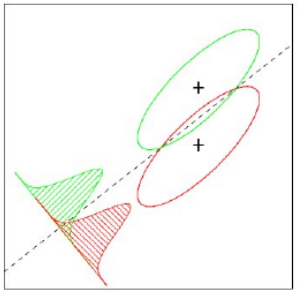
Donde nos quedamos



Donde nos quedamos



Discriminar - LDA



$$\mathbf{l} \propto \mathbf{S_W^{-1}}(\mathbf{c_+} - \mathbf{c_-})$$

Decisión óptima

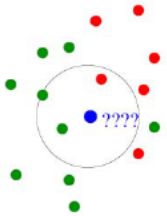
$$\arg\min_{\hat{\mathbf{y}}(.)} \mathbf{E}(\mathbf{L}(\mathbf{Y}, \hat{\mathbf{y}}(\mathbf{X})))$$

$$\frac{\mathbf{P(Y=1|X=x)}}{\mathbf{P(Y=0|X=x)}} \overset{?}{>} \mathbf{c}$$

$$\frac{\mathbf{P(X=x|Y=1)}}{\mathbf{P(X=x|Y=0)}} \overset{?}{>} \textbf{constante}$$

Clasificador Bayesiano óptimo

Enfoque noparamétrica



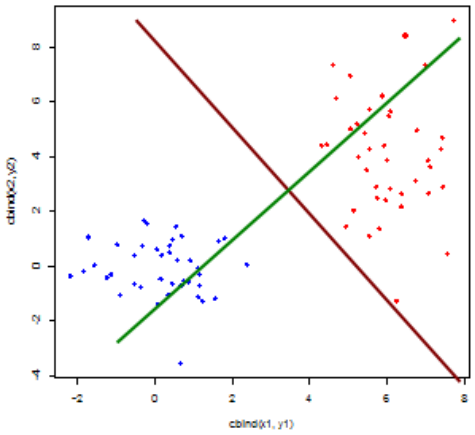
Optimalidad asimptótica de k-NN

2.5 Clasificadores lineales

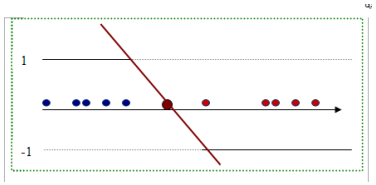
Buscamos clasificadores de la forma

$$\hat{y}(x) = sign(g(x)) = sign(\beta^t x + \alpha).$$

Observa: la frontera entre las dos clases es una linea de contorno de g.



¿ **Cómo determinar α y β?**:
Definir una función de costo:
$C(\alpha, \beta)$



Seguido, se trata de formular $C()$ basada en $\{g(x_i)\}$ y no $\{\hat{y}(x_i)\}$.

Interpretación geométrico: $|g(x)| \propto distancia(x, frontera)$

Base de clasificadores basados en márgenes

2.6 Regresión logística

Clasificador Bayesiano óptimo, caso binario:

$$\frac{P(Y=1|X=x)}{P(Y=0|X=x)} \overset{?}{>} c \qquad \frac{P(X=x|Y=1)}{P(X=x|Y=0)} \overset{?}{>} constante$$

Antes especificamos $X|Y=y$

Es un enfoque generativo.

Problema: tenemos que especificar más de lo que necesitamos.

2.6 Regresión logística

Clasificador Bayesiano óptimo, caso binario:

$$\frac{P(Y=1|X=x)}{P(Y=0|X=x)} \overset{?}{>} c \qquad \frac{P(X=x|Y=1)}{P(X=x|Y=0)} \overset{?}{>} constante$$

Antes especificamos $X|Y=y$

Es un enfoque generativo.

Problema: tenemos que especificar más de lo que necesitamos.

Ahora: parametrizamos y estimamos directamente:

$$\frac{P(Y=1|X=x)}{P(Y=0|X=x)}$$

Es un enfoque discriminatorio.

Vamos a suponer que $\exists \alpha, \beta$:

$$\log \frac{P(Y=1|X=x)}{P(Y=0|X=x)} = \alpha + \beta x.$$

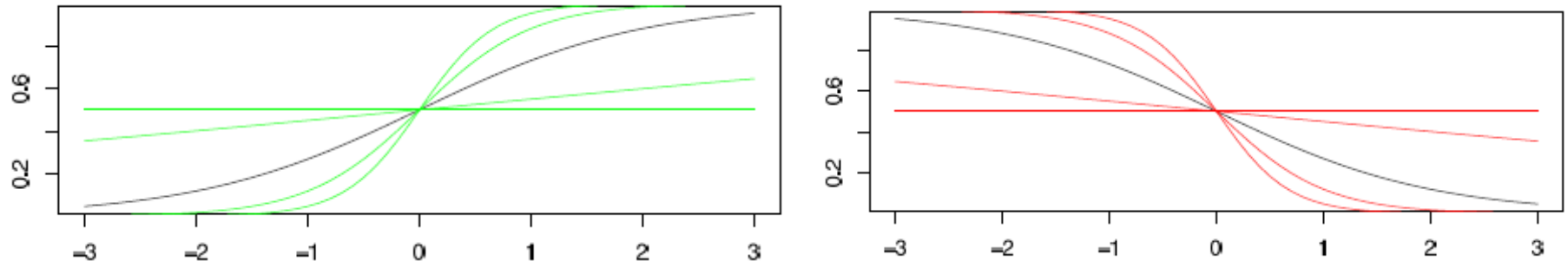
Vamos a suponer que $\exists \alpha, \beta$:

$$\log \frac{P(Y=1|X=x)}{P(Y=0|X=x)} = \alpha + \beta x.$$

Usando $P(Y=0|X=x) = 1 - P(Y=1|X=x)$:

$$P(Y=1|X=x) = \frac{1}{1+\exp(-\alpha-\beta x)}$$

Interpretación β:

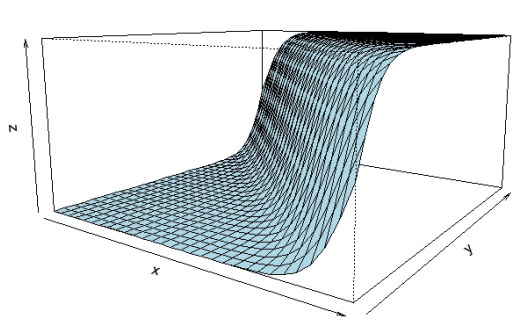


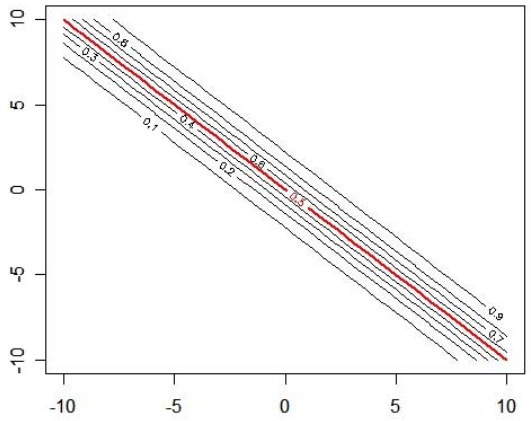
$$\log \frac{P(Y=1|X=x)}{P(Y=0|X=x)} = \alpha + \beta^t x.$$

Para simplificar la notación: incluye en vector x la constante 1:

$$\log \frac{P(Y=1|X=x)}{P(Y=0|X=x)} = \beta^t x.$$

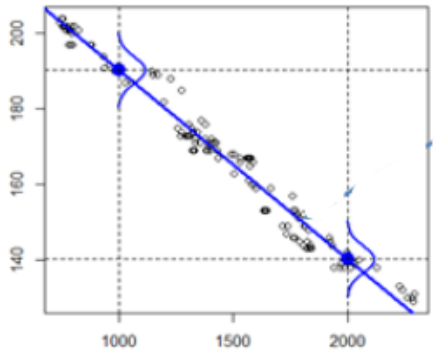
$$P(Y=1|X=x) = \frac{1}{1+\exp(-\beta^t x)}$$

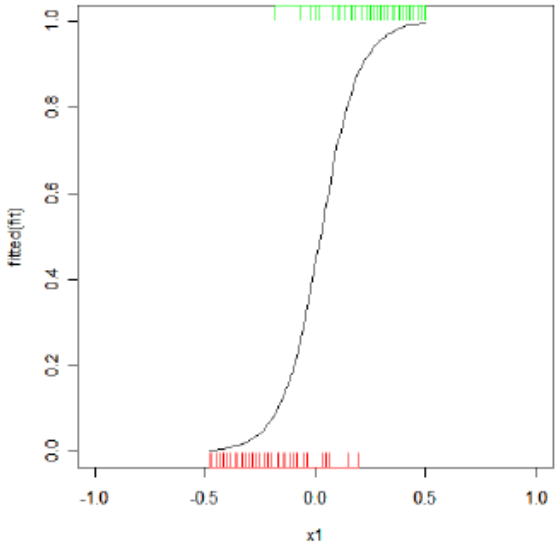




Medimos distancias hacia la frontera de decisión pero ya no Euclidianas (lineas de contornos: paralelas pero no equidistantes)

Compara regresión con regresión logística:

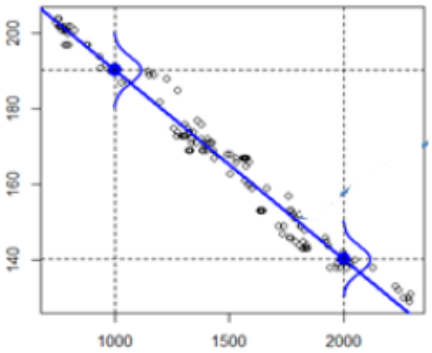


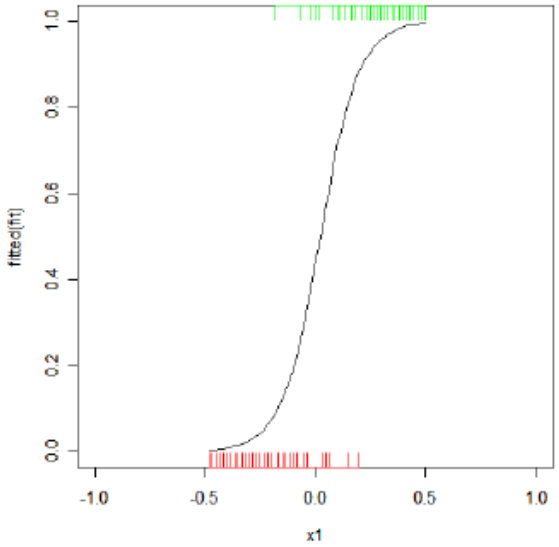


$Y = \beta^t X + \epsilon$ **con** $\epsilon \sim \mathcal{N}(0, \sigma^2)$

$\pi(x) = P(Y = 1|X = x) = (1 + \exp(-\beta^t x))^{-1}$

Compara regresión con regresión logística:





$Y = \beta^t X + \epsilon$ **con** $\epsilon \sim \mathcal{N}(0, \sigma^2)$

$\mathbf{Y} = \mathbb{X}\beta + \mathbf{E} \Rightarrow \mathbf{E} = \mathbf{Y} - \mathbb{X}\beta$

$l(\beta) = \sum_i logP(\epsilon_i = y_i - x_i^t\beta)$

$\hat{\beta} = (\mathbb{X}^t\mathbb{X})^{-1}\mathbb{X}^t\mathbf{Y}$

$\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2(\mathbb{X}^t\mathbb{X})^{-1})$

$\pi(x) = P(Y = 1|X = x) = (1 + \exp(-\beta^t x))^{-1}$

$l(\beta) = \sum_i \log(\pi(x_i)^{y_i}(1 - \pi(x_i))^{1-y_i})$

$= \sum_i (y_i \beta^t x_i - \log(1 + \exp(\beta^t x_i)))$

Recurrir a métodos iterativos

Para datos separables:

$||\beta|| \to \infty$

Estimación de los parámetros

$$l(\beta) = \sum_i \log(\pi(x_i)^{y_i}(1-\pi(x_i))^{1-y_i}),$$

$$= \sum_i (y_i \log(\pi(x_i)) + (1-y_i)\log(1-\pi(x_i))),$$

$$= \sum_i (y_i \log(\pi(x_i)/(1-\pi(x_i))) - \log(1/(1-\pi(x_i)))),$$

$$= \sum_i (y_i \beta^t x_i - \log(1+\exp(\beta^t x_i))),$$

Problema: no-lineal en β.

Por ejemplo se puede usar Newton-Raphson:

$$\beta^n = \beta^{n-1} - [\frac{\delta^2 l(\beta^{n-1})}{\delta\beta\delta\beta^t}]^{-1}(\frac{\delta l(\beta^{n-1})}{\delta\beta})$$

$$\frac{\delta l(\beta)}{\delta\beta} = \sum_i (y_i x_i - \frac{\exp(\beta^t x_i)}{1+\exp(\beta^t x_i)} x_i) = \sum_i x_i(y_i - \pi(x_i)),$$

$$\frac{\delta^2 l(\beta)}{\delta\beta\delta\beta^t} = \sum_i (x_i \pi(x_i)(1-\pi(x_i))x_i^t)$$

Si se escribe lo anterior en términos de matrices:

$$\frac{\delta l(\beta)}{\delta \beta} = \mathbb{X}^t(Y - \Pi), \frac{\delta^2 l(\beta)}{\delta \beta \delta \beta^t} = -\mathbb{X}^t \mathbb{W} \mathbb{X},$$

con $\mathbb{W} = Diag(\pi(x_i)(1 - \pi(x_i))$

De este manera:

$$\beta^n = \beta^{n-1} + (\mathbb{X}^t \mathbb{W} \mathbb{X})^{-1} \mathbb{X}^t (Y - \Pi)$$

$$\beta^n = (\mathbb{X}^t \mathbb{W} \mathbb{X})^{-1} \mathbb{X}^t \mathbb{W} (\mathbb{X} \beta^{n-1} + \mathbb{W}^{-1}(Y - \Pi)).$$

Es de la forma de regresión lineal ponderada:

$$\beta^n = (\mathbb{X}^t \mathbb{W} \mathbb{X})^{-1} \mathbb{X}^t \mathbb{W} Z.$$

Ejemplo

$x \in \mathcal{R}^{3+1}$, $\beta = (0, 10, 0, 0)$

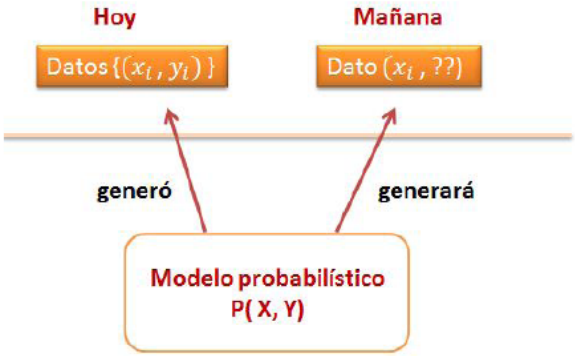
$$P(Y = 1|X = x) = (1 + \exp(-\beta^t x))^{-1}$$

```
> fit <- glm(y ~ x1+x2+x3, family=binomial(link=logit))
> summary(fit)

Call:
glm(formula = y ~ x1 + x2 + x3, family = binomial(link = logit))

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.12293  -0.09891   0.01052   0.08903   1.54857

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.5028     0.5966   0.843  0.39931
x1           20.2036     6.9865   2.892  0.00383 **
x2           -0.6924     0.5715  -1.212  0.22569
x3           -0.7142     0.6514  -1.096  0.27289
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 82.911  on 59  degrees of freedom
Residual deviance: 21.442  on 56  degrees of freedom
AIC: 29.442

Number of Fisher Scoring iterations: 8
```

Seleción de modelo basado en *likelihood ratio approach* **tipo**

$$-2(l_2 - l_1) \sim \chi^2_q$$

2.7 Comentarios finales



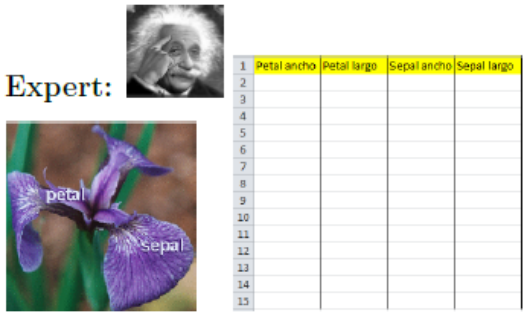
1. **Cubre amplio espectro de preguntas:**

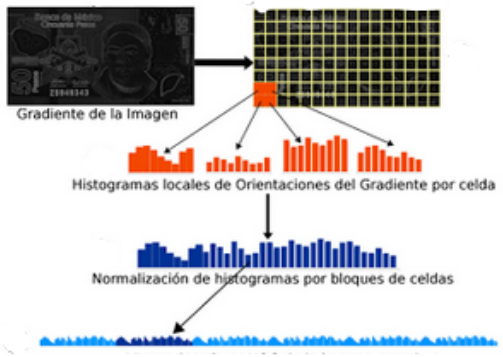
   como predecir (bien) ..... entender $P()$

ejemplo

2. **Obtención x:**

   experto ..... greedy





Retos nuevos en construcción, evaluación y elección de modelo