APRENDIZAJE DE MÁQUINA

CLASIFICACIÓN LINEAL

AGENDA

01 **Introducción**

Problema de clasificación y regresión lineal

02 **Clasificación binaria**

Regresión logística y método de Newton

03 **Modelos Lineales Generalizados**

Familia exponencial, Construyendo GLMs, Softmax

04 **Modelos discriminativos y generativos**

Diferencias, Análisis de Gaussianas, Clasificador de Bayes

05 **Métricas de evaluación binarias**

Sensibilidad, Especificidad, Puntaje F1, Curva ROC

INTRODUCCIÓN

INTRODUCCIÓN

PROBLEMA DE CLASIFICACIÓN

El **problema** es el **mismo** que el de **regresión**: **predecir** un conjunto de variables de **salidas** $\boldsymbol{y}$ dado un **conjunto** de **datos** de **entrada** $\boldsymbol{X}$.

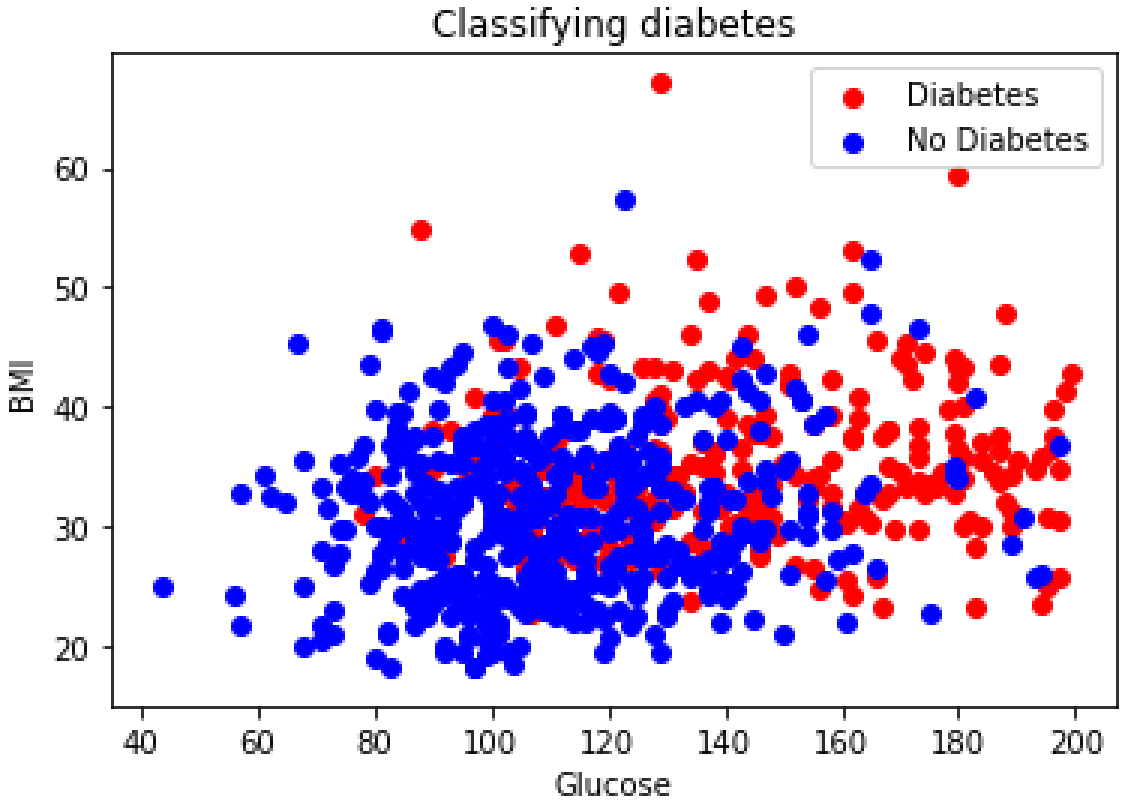
La única **deferencia** recae en que los **valores** de $\boldsymbol{y}$ toman un **conjunto** de **valores discretos.**

x_1=Glucosa [mg/dl]	x_2=IMC	y = Presencia de Diabetes
148	**33.6**	**1**
85	**26.6**	**0**
183	**23.3**	**1**
89	**28.1**	**0**
137	**43.1**	**1**
⋮	⋮	⋮

INTRODUCCIÓN

PROBLEMA DE CLASIFICACIÓN

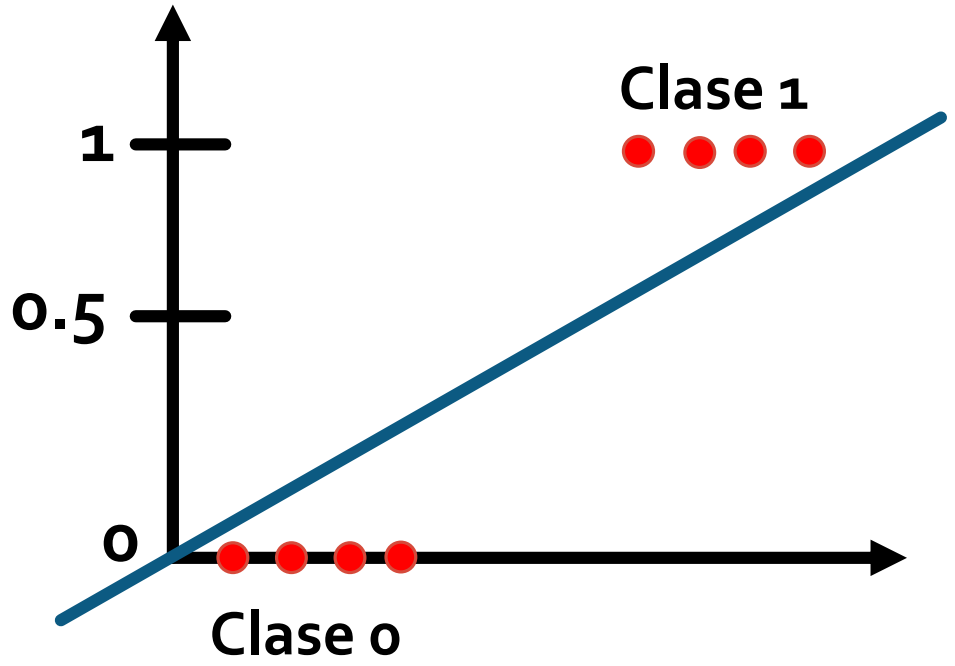
Graficamos los 752 datos con Python:



INTRODUCCIÓN

REGRESIÓN LINEAL

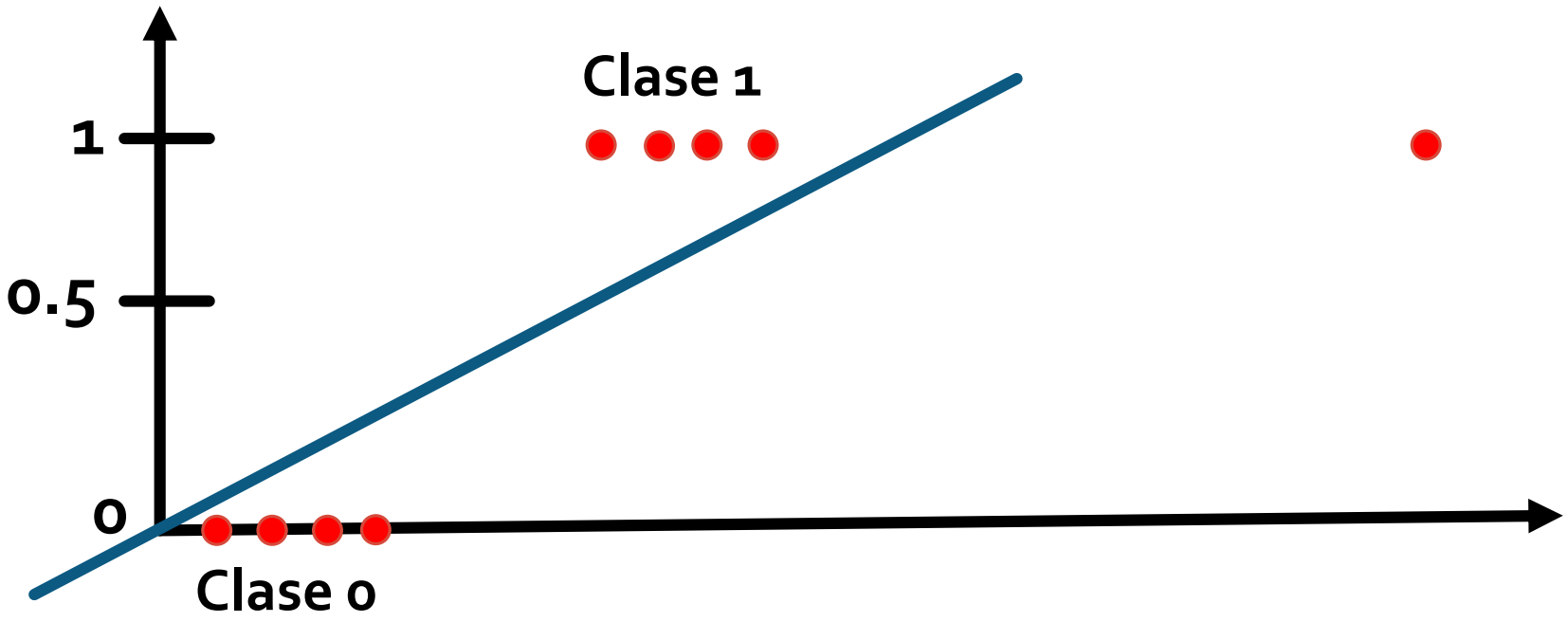
¿PORQUÉ NO APLICAR LA REGRESIÓN LINEAL PARA CLASIFICAR?



INTRODUCCIÓN

REGRESIÓN LINEAL

¿PORQUÉ NO APLICAR LA REGRESIÓN LINEAL PARA CLASIFICAR?



La función $w^T X$ no describe el valor esperado de $Y \in \{0, 1\}$

CLASIFICACIÓN BINARIA

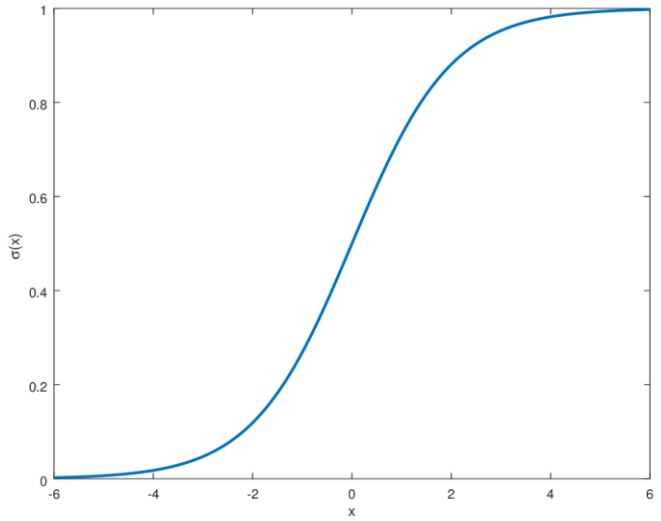
REGRESIÓN LOGÍSTICA

HIPÓTESIS

La **hipótesis** que utilizamos **para** la **regresión lineal no** es **adecuada** para resolver nuestro **problema** de **clasificación**.

Se propone una **hipótesis diferente** que se ajuste **mejor** al **problema**:

$$h_w(x) = g\left(w^T X\right) = \frac{1}{1 + e^{-w^T X}}$$



A esta función se le denomina **función sigmoide** o **logística**.

Se **garantiza** que $g(z) \in \{0, 1\}$.

REGRESIÓN LOGÍSTICA

FUNCIÓN SIGMOIDE

TAREA

Demostrar que:

$$g'(z) = g(z)(1 - g(z))$$

REGRESIÓN LOGÍSTICA

SUPOSICIONES

Ahora se define el **mismo problema** de **encontrar** la **mejor combinación** de **pesos** que se **ajuste** al **problema** de **clasificación**.

Se **aplican** las **siguientes suposiciones**:

1. Suponemos que la **hipótesis** define una **medida** de **probabilidad (Bernoulli)**:

$$\boldsymbol{P(y = 1/x; w) = h_w(x)}$$

$$\boldsymbol{P(y = 0/x; w) = 1 - h_w(x)}$$

$$\boldsymbol{P(y/x; w) = h_w(x)^y (1 - h_w(x))^{1-y}}$$

$$\boldsymbol{L(w) = P(y/x; w)}$$

REGRESIÓN LOGÍSTICA

SUPOSICIONES

2. **Suponiendo** que se **toman m muestras iid**, se tiene que la **verosimilitud** se puede **escribir** como:

$$L(w) = \prod_{i=1}^{m} P\left(y^{(i)} / x^{(i)}; w\right)$$

$$L(w) = \prod_{i=1}^{m} h_w\left(x^{(i)}\right)^{y^{(i)}} \left(1 - h_w\left(x^{(i)}\right)\right)^{1-y^{(i)}}$$

REGRESIÓN LOGÍSTICA

PÉRDIDA LOGARÍTMICA

Se puede escribir **la pérdida logarítmica:**

$$l(w) = log \prod_{i=1}^{m} P\left(y^{(i)} / x^{(i)}; w\right)$$

$$l(w) = \sum_{i=1}^{m} y^{(i)} log\, h_w\left(x^{(i)}\right) + \left(1 - y^{(i)}\right) log \left(1 - h_w\left(x^{(i)}\right)\right)$$

$$l(w) = \sum_{i=1}^{m} y^{(i)} log \left(\frac{1}{1 + e^{-w^T X}}\right) + \left(1 - y^{(i)}\right) log \left(1 - \frac{1}{1 + e^{-w^T X}}\right)$$

REGRESIÓN LOGÍSTICA

FUNCIÓN DE COSTO

Se puede escribir **la función de costo:**

$$J(w) = -l(w)$$

$$J(w) = -\sum_{i=1}^{m} y^{(i)} log\left(\frac{1}{1+e^{-w^T X}}\right) + \left(1 - y^{(i)}\right) log\left(1 - \frac{1}{1+e^{-w^T X}}\right)$$

REGRESIÓN LOGÍSTICA

FUNCIÓN DE COSTO

Interpretando **la función de costo:**

$$J(w) = -\sum_{i=1}^{m} y^{(i)} log\, h_w\left(x^{(i)}\right) + \left(1 - y^{(i)}\right) log\left(1 - h_w\left(x^{(i)}\right)\right)$$

Si $y = 0$	Si $y = 1$
$J(y = 0, \hat{y}) = -log(1 - \hat{y})$	$J(y = 1, \hat{y}) = -log(\hat{y})$
$\lim_{\hat{y} \to 0} log(1 - \hat{y}) \to 0$	$\lim_{\hat{y} \to 0} log(\hat{y}) \to \infty$
$\lim_{\hat{y} \to 1} log(1 - \hat{y}) \to \infty$	$\lim_{\hat{y} \to 1} log(\hat{y}) \to 0$

REGRESIÓN LOGÍSTICA

OPTIMIZACIÓN

¿CÓMO ENCONTRAR LOS MEJORES PESOS W?

REGRESIÓN LOGÍSTICA

DESCENSO POR GRADIENTE

Encontrar la **mejor combinación** de **pesos** w por descenso de gradiente se **ajuste** al **problema** de **clasificación**.

$$w := w - \alpha \nabla_w J(w)$$

REGRESIÓN LOGÍSTICA

DESCENSO POR GRADIENTE

TAREA

Demostrar que el **gradiente** respecto a un **solo dato** de **entrenamiento** $(\boldsymbol{x}, \boldsymbol{y})$ y **peso** $\boldsymbol{w_j}$ está dado por:

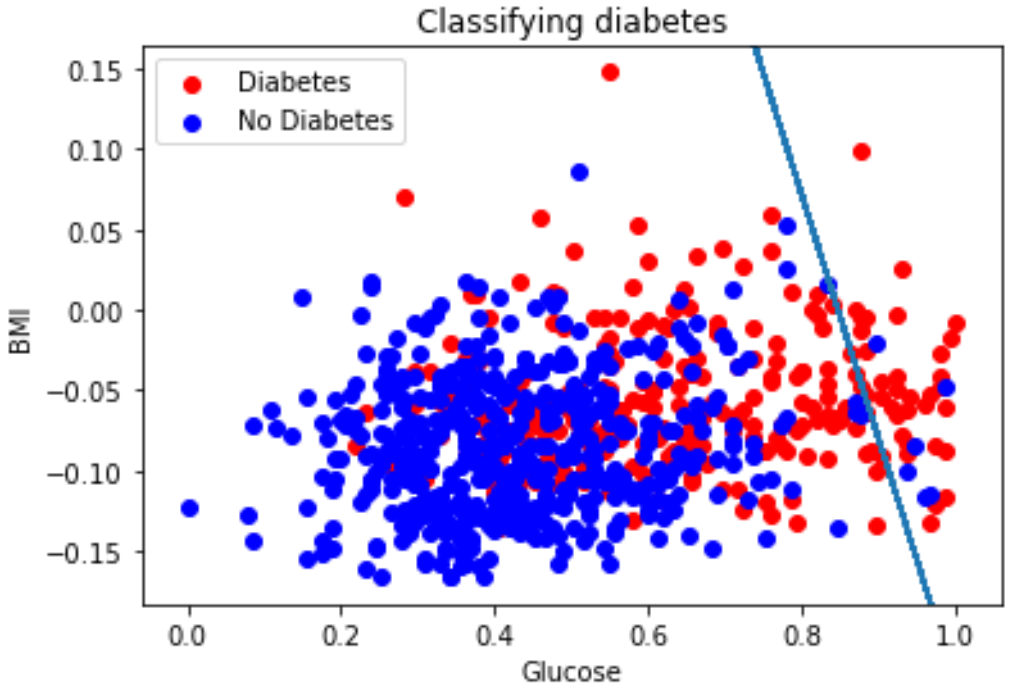
$$\frac{\partial}{\partial w_j} J(w) = (y - h_w(x))x_j$$

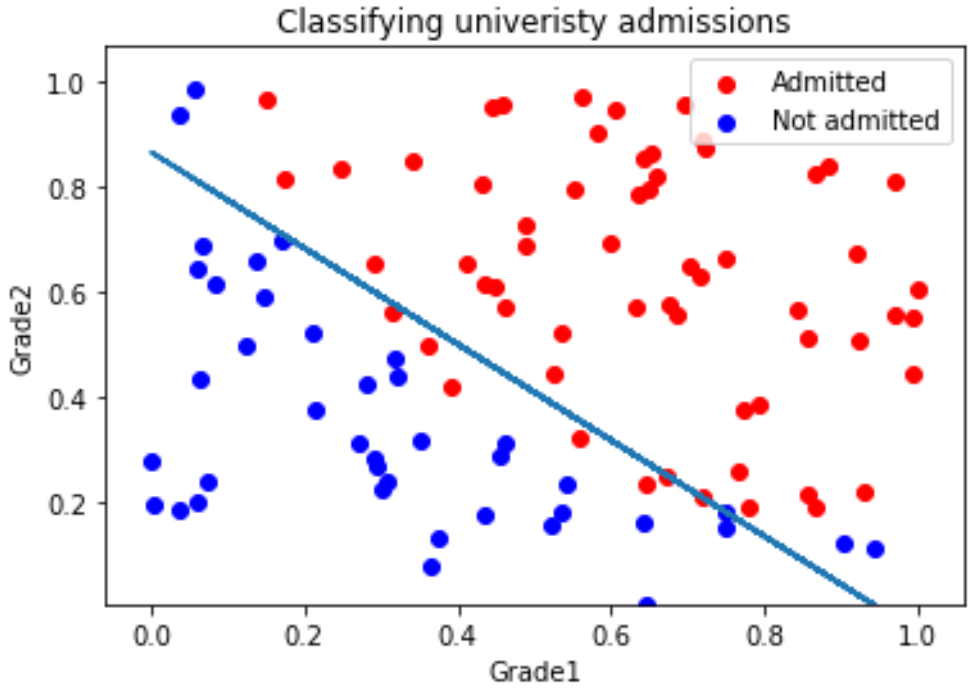
Donde $\boldsymbol{x}, \boldsymbol{w_j} \in \mathbb{R}^n$ y $\boldsymbol{y} \in \{0, 1\}$

RECORDAR EL DESCENSO POR GRADIENTE EN REGRESIÓN LINEAL → MISMO RESULTADO

REGRESIÓN LOGÍSTICA

EJEMPLO REAL



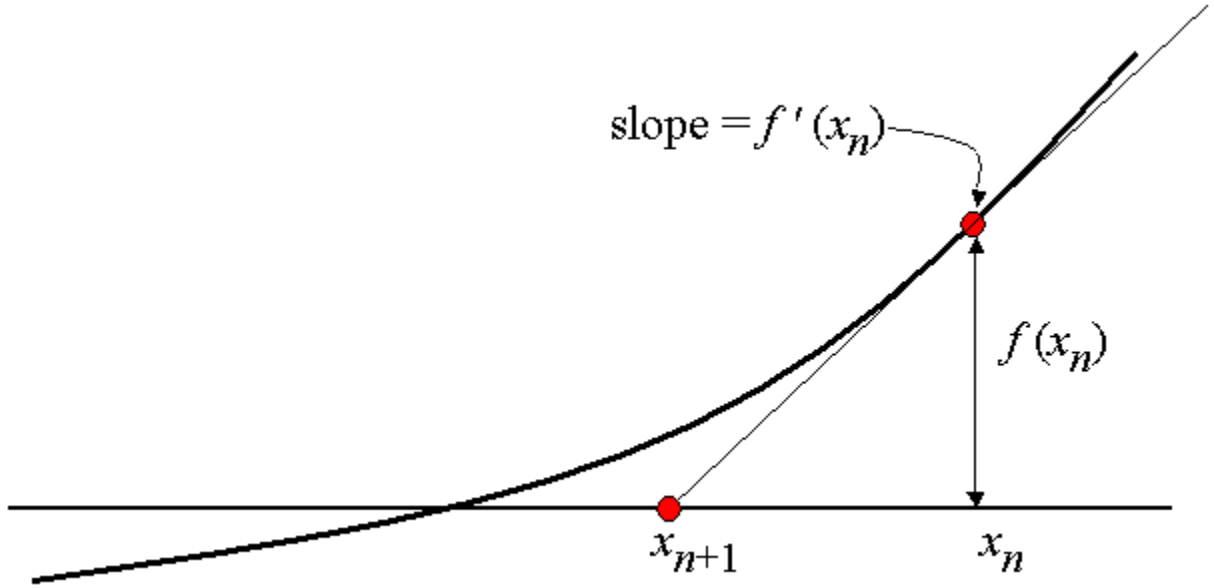


MÉTODO DE NEWTON

OPTIMIZACIÓN

Recordamos de métodos numéricos, el **método** de **Newton,** en donde se encuentra el valor de $\boldsymbol{w}$ para el cual $\boldsymbol{f}(\boldsymbol{w}) = \boldsymbol{0}$:

$$\boldsymbol{w} := \boldsymbol{w} - \frac{\boldsymbol{f}(\boldsymbol{w})}{\boldsymbol{f}'(\boldsymbol{w})}$$



MÉTODO DE NEWTON

OPTIMIZACIÓN

Si queremos **encontrar** el **mínimo** de la **función** de **costo** $\boldsymbol{J(w)}$**,** esto correspondería a encontrar los **puntos** en donde $\boldsymbol{\nabla_w J(w) = 0}$, por lo que se puede usar el **Método** de **Newton** (mucho **más rápido**):

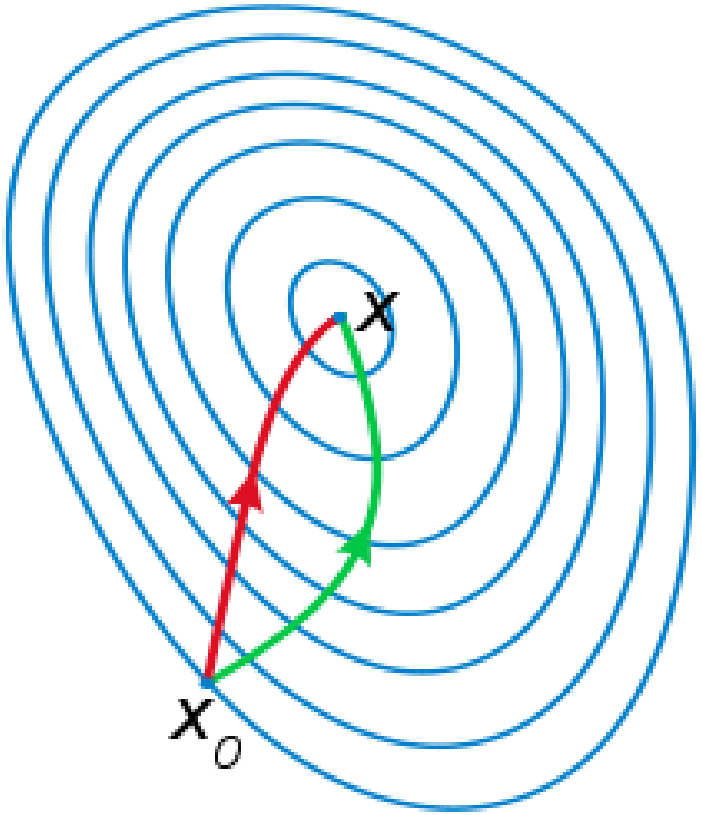
$$\boldsymbol{w := w - H^{-1} \nabla_w J(w)}$$

donde $\boldsymbol{H^{-1}}$ es la matriz Hessiana y $\boldsymbol{\nabla_w}$ es el gradiente.

$$\boldsymbol{H = \begin{bmatrix} \frac{\partial^2 J(w)}{\partial {w_1}^2} & \cdots & \frac{\partial^2 J(w)}{\partial w_1 \partial w_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 J(w)}{\partial w_n \partial w_1} & \cdots & \frac{\partial^2 J(w)}{\partial {w_n}^2} \end{bmatrix}}$$

MÉTODO DE NEWTON

OPTIMIZACIÓN



Descenso por gradiente
(Convergencia lineal)

Método de Newton
(Convergencia cuadrática)

MÉTODO DE NEWTON

OPTIMIZACIÓN

¿PORQUÉ NO USAR EL MÉTODO DE NEWTON EN LUGAR DE DESCENSO POR GRADIENTE?

MODELOS LINEALES GENERALIZADOS

MODELOS LINEALES GENERALIZADOS

F A M I L I A E X P O N E N C I A L

Recordamos que **realizamos** las siguientes **suposiciones:**

MÍNIMOS CUADRADOS

$\boldsymbol{y} \in \mathbb{R} \sim \boldsymbol{Gaussiana}$ (Regresión lineal)

REGRESIÓN LOGÍSTICA

$\boldsymbol{y} \in \{\mathbf{0}, \mathbf{1}\} \sim \boldsymbol{Bernoulli}$ (Clasificación binaria)

MODELOS LINEALES GENERALIZADOS

F A M I L I A E X P O N E N C I A L

Decimos que una **distribución** de **probabilidad** pertenece a la **familia exponencial** si se puede **escribir** de la **siguiente forma**:

$$p(y;\eta) = b(y)e^{\left(\eta^T T(y) - a(\eta)\right)}$$

Donde:

- η es denominado como el **parámetro natural** o **canónico** de la distribución.
- $T(y)$ **estadístico suficiente** (estadístico que **resume** la **información** completa de una **muestra**).
- $a(\eta)$ la **función logística** de **partición**.

NOTA: $e^{-a(\eta)}$ funciona como **constante** de **normalización** para asegurar que $p(y;\eta)$ **integre a 1.**

MODELOS LINEALES GENERALIZADOS

F A M I L I A E X P O N E N C I A L

$$p(y;\eta) = b(y)e^{\left(\eta^T T(y) - a(\eta)\right)}$$

Sí **fijamos** a, b **y** T, entonces podemos decir que tenemos una **distribución parametrizada** solo por η, donde al **variar** η se tienen **diferentes distribuciones.**

MODELOS LINEALES GENERALIZADOS

EJEMPLOS DE FAMILIA EXPONENCIAL

Probamos que la **distribución** de **Bernoulli** es **parte** de la **familia exponencial:**

$$\boldsymbol{p(y;\phi) = \phi^{y}(1-\phi)^{1-y}}$$

Al **variar** $\boldsymbol{\phi}$ se tienen **diferentes distribuciones** $\boldsymbol{p(y;\phi)}$.

MODELOS LINEALES GENERALIZADOS

EJEMPLOS DE FAMILIA EXPONENCIAL

Encontramos:

$$T(y) = y$$

$$b(y) = 1$$

$$\eta = log\left(\frac{\phi}{1-\phi}\right)$$

$$a(\eta) = -log(1-\phi)$$

$$p(y;\phi) = e^{\left(log\left(\frac{\phi}{1-\phi}\right)y + log(1-\phi)\right)}$$

MODELOS LINEALES GENERALIZADOS

EJEMPLOS DE FAMILIA EXPONENCIAL

Probamos que la **distribución Gaussiana (donde σ no importa)** es **parte** de la **familia exponencial (**por lo que $\sigma^2 = 1$**):**

$$p(y;\mu) = \frac{1}{\sqrt{2\pi}} e^{\frac{(y-\mu)^2}{2}}$$

TAREA 1

MODELOS LINEALES GENERALIZADOS

CONSTRUYENDO LOS MODELOS

Vamos a **realizar** las siguientes **suposiciones:**

1. La variable de salida $\boldsymbol{y}$ / $\boldsymbol{X}$; $\boldsymbol{w}\sim\boldsymbol{FamExp}$ $(\boldsymbol{\eta})$

2. **Objetivo:** dada una **matriz** de **características $\boldsymbol{X}$, calcular** $\boldsymbol{E}(\boldsymbol{T}(\boldsymbol{y});\boldsymbol{\eta})$ denominada como **función canónica de respuesta.**

$$\boldsymbol{h}(\boldsymbol{\eta}) = \boldsymbol{E}(\boldsymbol{T}(\boldsymbol{y});\boldsymbol{\eta})$$

3. Se define que la **relación entre $\boldsymbol{\eta}$, $\boldsymbol{w}$ y $\boldsymbol{X}$ es lineal (solo sí $\boldsymbol{\eta} \in \mathbb{R}$)**:

$$\boldsymbol{\eta} = \boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{X}$$

MODELOS LINEALES GENERALIZADOS

CONSTRUYENDO LOS MODELOS

EJEMPLO: Bernoulli

Para valores fijos de $\boldsymbol{X}$ y $\boldsymbol{w}$ se tiene que el **objetivo** es **calcular**:

$$\boldsymbol{h(w) = E(T(y); \eta)}$$

Pero sabemos que $\boldsymbol{T(y) = y}$ por lo que el **valor esperado** de una **variable** que se **distribuye** como **Bernoulli** es:

$$\boldsymbol{E(y;\ \eta) = P(y = 1;\ \eta)}$$

De igual forma, sabemos de la familia exponencial que:

$$\boldsymbol{\phi = P(y = 1;\ \eta)}$$

MODELOS LINEALES GENERALIZADOS

EJEMPLOS DE FAMILIA EXPONENCIAL

Se tiene de la definición de la **familia exponencial** que el parámetro natural $\boldsymbol{\eta}$ está **expresado como**:

$$\boldsymbol{\eta = log\left(\frac{\phi}{1-\phi}\right)}$$

Al **despejar** $\boldsymbol{\phi}$ de la igualdad anterior:

$$\boldsymbol{\eta = log\left(\frac{\phi}{1-\phi}\right)}$$

Se **tiene** lo **siguiente**:

$$\boldsymbol{\phi = \frac{1}{1+e^{-\eta}}}$$

MODELOS LINEALES GENERALIZADOS

CONSTRUYENDO LOS MODELOS

Por lo que el **valor esperado** es:

$$E(y;\ \eta) = \phi = \frac{1}{1 + e^{-\eta}}$$

Pero gracias a la tercera suposición $\eta = w^T X$

$$E(y; w, X) = \phi = \boxed{\frac{1}{1 + e^{-W^T X}}}$$

FUNCIÓN SIGMOIDE

MODELOS LINEALES GENERALIZADOS

CONSTRUYENDO LOS MODELOS

EJEMPLO: Gaussiana

Probar que la **función canónica** de **respuesta** es equivalente a $\boldsymbol{W^T X}$ suponiendo que $\boldsymbol{y \sim Gaussiana}$ y que $\boldsymbol{\eta = W^T X}$

$$\boldsymbol{h(w) = E(T(y);\ \eta) = w^T X}$$

TAREA 2

MODELOS LINEALES GENERALIZADOS

R E G R E S I Ó N S O F T M A X

Se define el **siguiente problema**, en donde la **variable** de **salida** $\boldsymbol{y} \in \{\boldsymbol{1}, \ldots, \boldsymbol{k}\}$ se **distribuye** como una **función multinomial** $\boldsymbol{y \sim Multinomial}$.

- Parámetros: $\phi_1, \phi_2, \ldots, \phi_{k-1}$
- $P(y = i) = \phi_i$
- $\phi_k = 1 - (\phi_1 + \phi_2 + \cdots + \phi_{k-1})$

Se define el **estadístico suficiente** $\boldsymbol{T(y)}$ como **vector**:

$$T(1) = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} T(2) = \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix} \ldots T(k-1) = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix} T(k) = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \in \mathbb{R}^{k-1}$$

Función indicador:

$$1\{True\} = 1 \qquad 1\{False\} = 0$$

$$T(y)_i = 1\{y = i\}$$

MODELOS LINEALES GENERALIZADOS

R E G R E S I Ó N S O F T M A X

Se expresa la **distribución multinomial** en **forma** de la **familia exponencial**:

$$P(y) = \phi_1^{1\{y=1\}} \phi_2^{1\{y=2\}} \dots \phi_k^{1\{y=k\}}$$

$$P(y) = \phi_1^{T(y)_1} \phi_2^{T(y)_2} \dots \phi_k^{T(y)_{k-1}} \phi_k^{1-\sum_{j=1}^{k-1} T(y)_j}$$

$$P(y) = e^{\left[T(y)_1 \log(\phi_1) + T(y)_2 \log(\phi_2) + \dots + (1-\sum_{j=1}^{k-1} T(y)_j) \log(\phi_k)\right]}$$

$$P(y) = e^{\left[T(y)_1 \log\left(\frac{\phi_1}{\phi_k}\right) + T(y)_2 \log\left(\frac{\phi_2}{\phi_k}\right) + \dots + T(y)_{k-1} \log\left(\frac{\phi_{k-1}}{\phi_k}\right) + \log(\phi_k)\right]}$$

MODELOS LINEALES GENERALIZADOS

R E G R E S I Ó N S O F T M A X

Se **compara** con la **forma** de la **familia exponencial**:

$$P(y) = e^{\left[T(y)_1 \log\left(\frac{\phi_1}{\phi_k}\right) + T(y)_2 \log\left(\frac{\phi_2}{\phi_k}\right) + \ldots + T(y)_{k-1} \log\left(\frac{\phi_{k-1}}{\phi_k}\right) + \log(\phi_k)\right]}$$

$$\boldsymbol{p(y;\eta) = b(y)e^{\left(\eta^T T(y) - a(\eta)\right)}}$$

Por lo tanto:

$$\boldsymbol{b(y) = 1}$$

$$\boldsymbol{a(\eta) = -log(\phi_k)}$$

MODELOS LINEALES GENERALIZADOS

R E G R E S I Ó N S O F T M A X

$$P(y) = e^{\left[T(y)_1 \log\left(\frac{\phi_1}{\phi_k}\right) + T(y)_2 \log\left(\frac{\phi_2}{\phi_k}\right) + \ldots + T(y)_{k-1} \log\left(\frac{\phi_{k-1}}{\phi_k}\right) + \log(\phi_k)\right]}$$

$$\boldsymbol{p(y;\eta) = b(y)e^{\left(\eta^T T(y) - a(\eta)\right)}}$$

Finalmente:

$$\boldsymbol{\eta = \begin{bmatrix} log\left(\frac{\phi_1}{\phi_k}\right) \\ \vdots \\ log\left(\frac{\phi_{k-1}}{\phi_k}\right) \end{bmatrix}}$$

MODELOS LINEALES GENERALIZADOS

R E G R E S I Ó N S O F T M A X

La **función canónica** de **respuesta** sería:

$$\eta_i = log\left(\frac{\phi_i}{\phi_k}\right)$$

$$\eta_k = log\left(\frac{\phi_k}{\phi_k}\right) = 0$$

$$\phi_k e^{\eta_i} = \phi_i$$

MODELOS LINEALES GENERALIZADOS

R E G R E S I Ó N S O F T M A X

La **suma** de **probabilidades** debe ser **igual** a **1**:

$$\sum_{i=1}^{k} \boldsymbol{\phi_k} e^{\eta_i} = \sum_{i=1}^{k} \boldsymbol{\phi_i} = \mathbf{1}$$

$$\boldsymbol{\phi_k} \sum_{i=1}^{k} e^{\eta_i} = \sum_{i=1}^{k} \boldsymbol{\phi_i} = \mathbf{1}$$

$$\boldsymbol{\phi_k} = \frac{\mathbf{1}}{\sum_{i=1}^{k} e^{\eta_i}}$$

MODELOS LINEALES GENERALIZADOS

R E G R E S I Ó N S O F T M A X

Sustituyendo $\boldsymbol{\phi_k} = \frac{\mathbf{1}}{\sum_{i=1}^{k} e^{\eta_i}}$ en la función canónica de respuesta:

$$\boldsymbol{\phi_k} e^{\eta_i} = \boldsymbol{\phi_i}$$

$$\boldsymbol{\phi_i} = \frac{e^{\eta_i}}{\sum_{j=1}^{k} e^{\eta_j}}$$

FUNCIÓN SOFTMAX

MODELOS LINEALES GENERALIZADOS

R E G R E S I Ó N S O F T M A X

Utilizando la **tercera suposición** de que $\boldsymbol{\eta_i = w_i^T X}$:

$$\boldsymbol{\phi_i = \frac{e^{w_i^T X}}{\sum_{j=1}^{k} e^{w_j^T X}}}$$

Sabiendo que el objetivo es calcular $\boldsymbol{h(w) = E(T(y);\ \eta)}$:

$$\boldsymbol{E\left(\begin{bmatrix} T(y)_1 \\ T(y)_2 \\ \vdots \\ T(y)_{k-1} \end{bmatrix}\right) = E\left(\begin{bmatrix} 1\{y=1\} \\ 1\{y=2\} \\ \vdots \\ 1\{y=k-1\} \end{bmatrix}\right) = \begin{bmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_{k-1} \end{bmatrix}}$$

MODELOS LINEALES GENERALIZADOS

R E G R E S I Ó N S O F T M A X

La hipótesis $\boldsymbol{h(w)}$ queda **expresada como** un **vector**:

$$\boldsymbol{h(w)} = \begin{bmatrix} \dfrac{e^{w_1^T X}}{\sum_{j=1}^{k} e^{w_j^T X}} \\ \dfrac{e^{w_2^T X}}{\sum_{j=1}^{k} e^{w_j^T X}} \\ \vdots \\ \dfrac{e^{w_{k-1}^T X}}{\sum_{j=1}^{k} e^{w_j^T X}} \end{bmatrix} = \begin{bmatrix} P(y=1) \\ P(y=2) \\ \vdots \\ P(y=k-1) \end{bmatrix}$$

que tiene como **componentes** la **probabilidad** cada **clase** $\boldsymbol{i}$: $\boldsymbol{P(y=i)}$

MODELOS LINEALES GENERALIZADOS

R E G R E S I Ó N S O F T M A X

Calculamos la **pérdida logarítmica** y **función de costo** para **un solo** dato de **entrenamiento**:

$$\boldsymbol{l(w) = \sum_{l=1}^{k} 1(y = l) log(p(y/x)) = \sum_{i=1}^{k} 1(y = l) log\left(\frac{e^{w_l^T X}}{\sum_{j=1}^{k} e^{w_j^T X}}\right)}$$

$$\boldsymbol{J(w) = -l(w)}$$

MODELOS LINEALES GENERALIZADOS

R E G R E S I Ó N S O F T M A X

Calculamos la **pérdida logarítmica** y **función de costo** para $\boldsymbol{m}$ datos de **entrenamiento**:

$$\boldsymbol{l(w) = \sum_{i=1}^{m} \sum_{l=1}^{k} 1\left(y^{(i)} = l\right) log\left(\frac{e^{w_l^T x^{(i)}}}{\sum_{j=1}^{k} e^{w_j^T x^{(i)}}}\right)}$$

$$\boldsymbol{J(w) = -l(w)}$$

SE PUEDE USAR EL MÉTODO DE NEWTON
O DESCENSO POR GRADIENTE

MODELOS DISCRIMINATIVOS Y GENERATIVOS

MODELOS DISCRIMINATIVOS Y GENERATIVOS

DIFERENCIAS

Modelos discriminativos:

- Los **modelos** que se han visto hasta ahora, se denominan como **modelos discriminativos**, en donde se realiza el **aprendizaje** sobre la **distribución** de **probabilidad** $\boldsymbol{p(y/x)}$ **directamente**.

- **Ejemplos**: regresión logística y mínimos cuadrados.

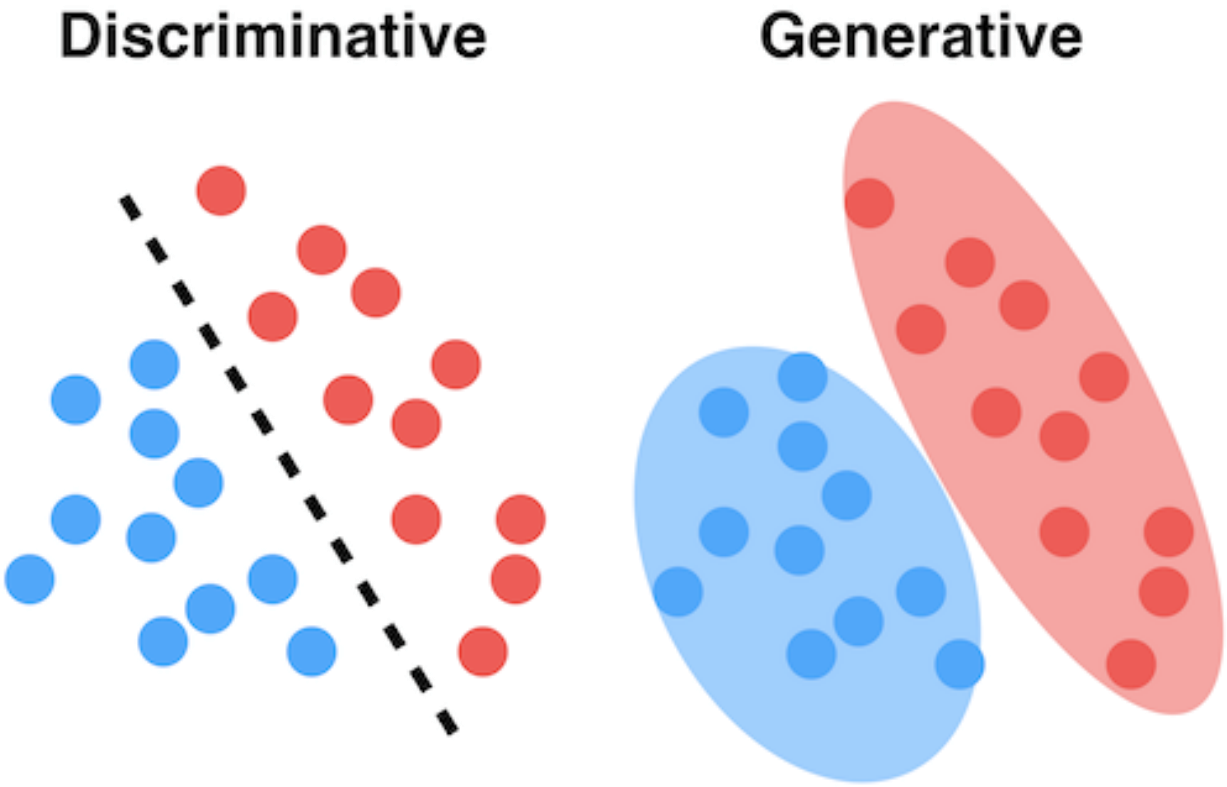
Modelos generativos:

- Modelos que tratan de **aprender** $\boldsymbol{p(x/y)}$ y $\boldsymbol{p(y)}$. Para el caso de **clasificación binaria** el algoritmo **aprendería tres distribuciones** distintas $\boldsymbol{p(x/y=0)}$, $\boldsymbol{p(x/y=1)}$ y $\boldsymbol{p(y)}$.

- Entonces se **clasificaría** un **nuevo dato comparándolo** con ambas distribuciones $\boldsymbol{p(x/y=0)}$ y $\boldsymbol{p(x/y=1)}$.

MODELOS DISCRIMINATIVOS Y GENERATIVOS

DIFERENCIAS



MODELOS GENERATIVOS

DESCRIPCIÓN

Un **modelo generativo modela** las **características** que están **condicionadas** por la **variable** de **respuesta** $p(x/y)$**.**

Utilizando el **Teorema** de **Bayes** se puede calcular $p(y/x)$**.**

$$p(y/x) = \frac{p(x/y)p(y)}{p(x)}$$

Se tiene que el **denominador** se puede calcular de la **siguiente forma (clasificación binaria):**

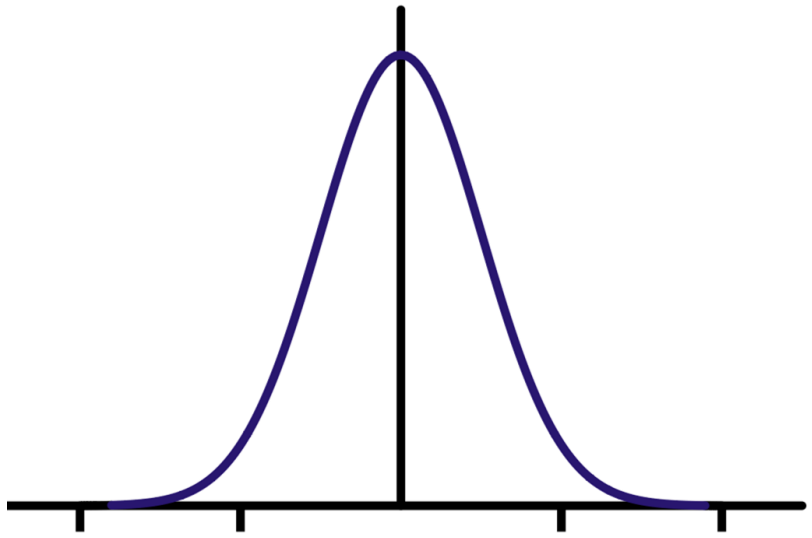
$$p(x) = p(x/y = 1)p(y = 1) + p(x/y = 0)p(y = 0)$$

MODELOS GENERATIVOS

ANÁLISIS DISCRIMINATIVO GAUSSIANO

En un **análisis discriminativo Gaussiano suponemos** que:

$$p(x/y) \sim Gaussiana$$



MODELOS GENERATIVOS

GAUSSIANAS MULTIVARIADAS

Una **distribución Gaussiana** de $\boldsymbol{d}$ **dimensiones** es parametrizada por:

- **Vector** de **media**: $\mu \in \mathbb{R}^d$
- **Matriz** de **covarianza**: $\Sigma \in \mathbb{R}^{dxd}$

donde Σ es simétrica y positiva semi-definida. Su **densidad** se puede **calcular** como:

$$\boldsymbol{p(x;\mu,\Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} e^{\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)}}$$

Propiedades:

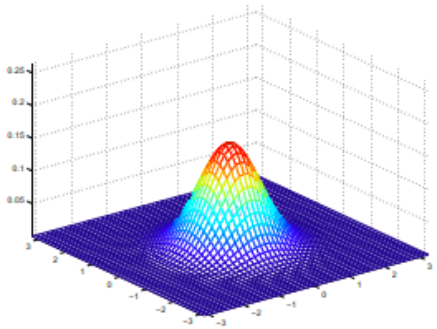
$$\boldsymbol{E[X] = \int_x x\, p(x;\mu,\Sigma)\, dx = \mu}$$

$$\boldsymbol{Cov(X) = E[(X - E[X])(X - E[X])^T] = \Sigma}$$

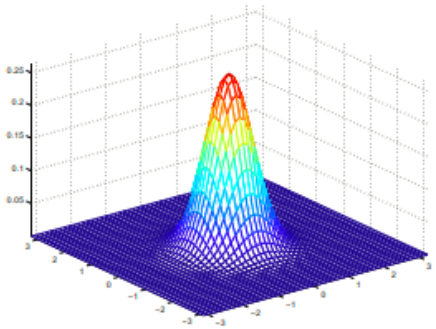
MODELOS GENERATIVOS

GAUSSIANAS MULTIVARIADAS

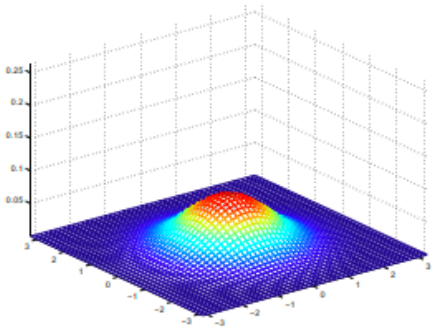
Ejemplo: Variando Σ.



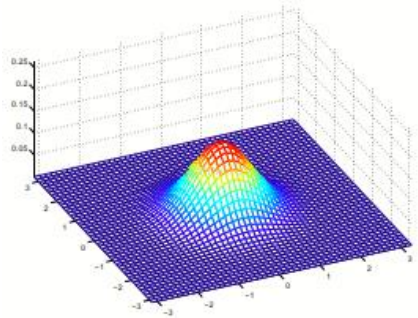
$$\boldsymbol{\Sigma} = \begin{bmatrix} \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{1} \end{bmatrix}$$



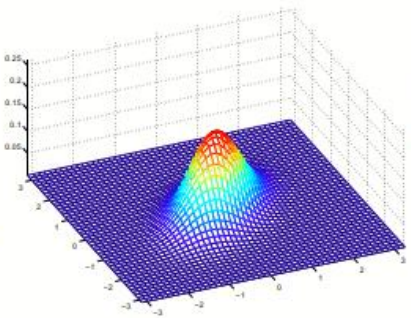
$$\boldsymbol{\Sigma} = \begin{bmatrix} \mathbf{0.6} & \mathbf{0} \\ \mathbf{0} & \mathbf{0.6} \end{bmatrix}$$



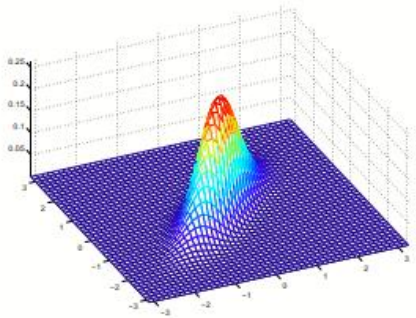
$$\boldsymbol{\Sigma} = \begin{bmatrix} \mathbf{2} & \mathbf{0} \\ \mathbf{0} & \mathbf{2} \end{bmatrix}$$



$$\boldsymbol{\Sigma} = \begin{bmatrix} \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{1} \end{bmatrix}$$



$$\boldsymbol{\Sigma} = \begin{bmatrix} \mathbf{1} & \mathbf{0.5} \\ \mathbf{0.5} & \mathbf{1} \end{bmatrix}$$

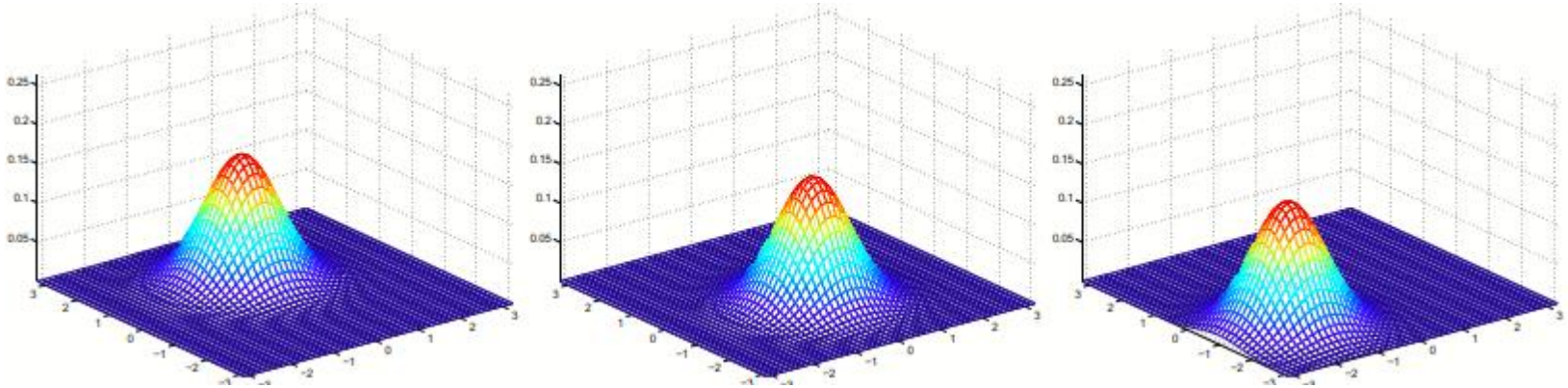


$$\boldsymbol{\Sigma} = \begin{bmatrix} \mathbf{1} & \mathbf{0.8} \\ \mathbf{0.8} & \mathbf{1} \end{bmatrix}$$

MODELOS GENERATIVOS

GAUSSIANAS MULTIVARIADAS

Ejemplo: Variando μ.



$$\boldsymbol{\mu} = \begin{bmatrix} \mathbf{1} \\ \mathbf{0} \end{bmatrix} \qquad \boldsymbol{\mu} = \begin{bmatrix} \mathbf{-0.5} \\ \mathbf{0} \end{bmatrix} \qquad \boldsymbol{\mu} = \begin{bmatrix} \mathbf{-1} \\ \mathbf{-1.5} \end{bmatrix}$$

MODELOS GENERATIVOS

ANÁLISIS DISCRIMINATIVO GAUSSIANO

El **análisis discriminativo Gaussiano resuelve** el problema de **clasificación binaria** donde:

$$\boldsymbol{x} \in \mathbb{R}^{\boldsymbol{n}}$$
$$\boldsymbol{y} \in \{\mathbf{0}, \mathbf{1}\}$$

Las **suposiciones** del **modelo son**:

$$y \sim Bernoulli(\emptyset)$$

$$x \,/\, y = 0 \sim N(\mu_0, \Sigma)$$

$$x \,/\, y = 1 \sim N(\mu_1, \Sigma)$$

MODELOS GENERATIVOS

ANÁLISIS DISCRIMINATIVO GAUSSIANO

Las **distribuciones** están **dadas** por:

$$
\begin{aligned}
p(y) &= \phi^y(1-\phi)^{1-y} \\
p(x|y=0) &= \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)\right) \\
p(x|y=1) &= \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)\right)
\end{aligned}
$$

Los **parámetros** de las distribuciones **serían**: $\boldsymbol{\phi}, \boldsymbol{\mu_0}, \boldsymbol{\mu_1}, \boldsymbol{\Sigma}$

MODELOS GENERATIVOS

ANÁLISIS DISCRIMINATIVO GAUSSIANO

La **verosimilitud conjunta** de los datos estaría dada por:

$$l(\phi, \mu_k, \Sigma) = \prod_{i=1}^{m} p(x^{(i)}, y^{(i)}; \phi, \mu_k, \Sigma)$$

$$l(\phi, \mu_k, \Sigma) = \prod_{i=1}^{m} p(x^{(i)}/y^{(i)};\ \mu_k, \Sigma) p(y^{(i)}; \phi)$$

$$\log l(\phi, \mu_k, \Sigma) = \log \prod_{i=1}^{m} p(x^{(i)}/y^{(i)}; \mu_k, \Sigma) p(y^{(i)}; \phi)$$

$$\log l(\phi, \mu_k, \Sigma) = \sum_{i=1}^{m} \log(p(x^{(i)}/y^{(i)}; \mu_k, \Sigma) + \log p(y^{(i)}; \phi)$$

MODELOS GENERATIVOS

ANÁLISIS DISCRIMINATIVO GAUSSIANO

La **verosimilitud conjunta** de los datos estaría dada por:

$$log\, l(\phi, \mu_k, \Sigma) = \sum_{i=1}^{m} log(p(x^{(i)}/y^{(i)}; \mu_k, \Sigma) + log\, p(y^{(i)}; \phi)$$

$$log\, l(\phi, \mu_k, \Sigma) = \sum_{i=1}^{m} \log\left(\frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}}\right) + log\left[e^{\left(-\frac{1}{2}(x^{(i)}-\mu_k)^T\Sigma^{-1}(x^{(i)}-\mu_k)\right)}\right] + \log(\Phi^{y^{(i)}}(1-\Phi)^{1-y^{(i)}})$$

$$log\, l(\phi, \mu_k, \Sigma) = \sum_{i=1}^{m} log(\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}) - \frac{1}{2}\left(x^{(i)} - \mu_k\right)^T\Sigma^{-1}\left(x^{(i)} - \mu_k\right) + y^{(i)}\, log(\phi) + \left(1 - y^{(i)}\right)\log((1-\Phi))$$

MODELOS GENERATIVOS

ANÁLISIS DISCRIMINATIVO GAUSSIANO

Maximizando respecto a ϕ:

$$\nabla_{\emptyset} \log l(\phi, \mu_k, \Sigma) = \nabla_{\phi} \sum_{i=1}^{m} \boldsymbol{y^{(i)}} \log\left(\frac{\phi}{1-\phi}\right) + \log((1-\Phi)) = 0$$

$$\nabla_{\emptyset} \log l(\phi, \mu_k, \Sigma) = \sum_{i=1}^{m} \left(\frac{\boldsymbol{y^{(i)}}}{\phi(1-\phi)} - \frac{1}{1-\phi}\right) = 0$$

$$\sum_{i=1}^{m} \frac{\boldsymbol{y^{(i)}}}{\phi(1-\phi)} = \frac{m}{1-\phi}$$

$$\nabla_{\emptyset} \log l(\phi, \mu_k, \Sigma) = \sum_{i=1}^{m} \frac{\boldsymbol{y^{(i)}}}{m} = \sum_{i=1}^{m} \frac{1(\boldsymbol{y^{(i)}} = \boldsymbol{k})}{m} = \phi = \frac{numero\ de\ datos\ de\ la\ clase\ k}{total\ de\ datos}$$

MODELOS GENERATIVOS

ANÁLISIS DISCRIMINATIVO GAUSSIANO

Maximizando respecto a $\boldsymbol{\mu_k}$:

$$\nabla_{\mu_k} \log l(\phi, \mu_k, \Sigma) = \nabla_{\mu_k} \sum_{i=1}^{m} -\frac{\mathbf{1}}{\mathbf{2}} \left(\boldsymbol{x^{(i)}} - \boldsymbol{\mu_k}\right)^{\boldsymbol{T}} \boldsymbol{\Sigma^{-1}} \left(\boldsymbol{x^{(i)}} - \boldsymbol{\mu_k}\right) = 0$$

Aplicando la propiedad $\nabla_x x^T A x = 2Ax$:

$$\nabla_{\mu_k} \log l(\phi, \mu_k, \Sigma) = \sum_{i=1}^{m} \boldsymbol{\Sigma^{-1}} \left(\boldsymbol{x^{(i)}} - \boldsymbol{\mu_k}\right) = 0$$

$$\nabla_{\mu_k} \log l(\phi, \mu_k, \Sigma) = \boldsymbol{\Sigma^{-1}} \sum_{i=1}^{m} \left(\boldsymbol{x^{(i)}} - \boldsymbol{\mu_k}\right) = 0$$

$$\nabla_{\mu_k} \log l(\phi, \mu_k, \Sigma) = \sum_{i=1}^{m} \boldsymbol{x^{(i)}} - \sum_{i=1}^{m} \boldsymbol{\mu_k} = 0$$

MODELOS GENERATIVOS

ANÁLISIS DISCRIMINATIVO GAUSSIANO

Solo nos **interesan las** $\boldsymbol{x^{(i)}}$ que pertenezcan a la **clase k**:

$$\nabla_{\mu_k} \log l(\phi, \mu_k, \Sigma) = \sum_{i=1}^{m} \boldsymbol{x^{(i)}} - \sum_{i=1}^{m} \boldsymbol{\mu_k} = 0$$

$$\nabla_{\mu_k} \log l(\phi, \mu_k, \Sigma) = \sum_{i=1}^{m} 1\left(\boldsymbol{y^{(i)} = k}\right)\boldsymbol{x^{(i)}} - \sum_{i=1}^{m} 1(\boldsymbol{y^{(i)} = k})\boldsymbol{\mu_k} = 0$$

$$\boldsymbol{\mu_k} = \frac{\sum_{i=1}^{m} 1(\boldsymbol{y^{(i)} = k})\boldsymbol{x^{(i)}}}{\sum_{i=1}^{m} 1(\boldsymbol{y^{(i)} = k})} = \frac{\boldsymbol{suma\ de\ x\ que\ pertenecen\ a\ clase\ k}}{\boldsymbol{número\ de\ datos\ de\ clase\ k}}$$

MODELOS GENERATIVOS

ANÁLISIS DISCRIMINATIVO GAUSSIANO

Maximizando respecto a $\boldsymbol{\Sigma^{-1}}$:

$$\nabla_{\boldsymbol{\Sigma^{-1}}} \log l(\phi, \mu_k, \Sigma) = \nabla_{\boldsymbol{\Sigma^{-1}}} \sum_{i=1}^{m} \boldsymbol{log}\left(\frac{\mathbf{1}}{(\boldsymbol{2\pi})^{d/2}|\boldsymbol{\Sigma}|^{1/2}}\right) - \frac{\mathbf{1}}{\mathbf{2}}\left(\boldsymbol{x^{(i)}} - \boldsymbol{\mu_k}\right)^{\boldsymbol{T}} \boldsymbol{\Sigma^{-1}}\left(\boldsymbol{x^{(i)}} - \boldsymbol{\mu_k}\right) = 0$$

$$\nabla_{\boldsymbol{\Sigma^{-1}}} \log l(\phi, \mu_k, \Sigma) = \nabla_{\boldsymbol{\Sigma^{-1}}} \sum_{i=1}^{m} \boldsymbol{log}\left(\frac{\mathbf{1}}{(\boldsymbol{2\pi})^{d/2}|\boldsymbol{\Sigma}|^{1/2}}\right) - \frac{\mathbf{1}}{\mathbf{2}}\left(\boldsymbol{x^{(i)}} - \boldsymbol{\mu_k}\right)^{\boldsymbol{T}}\left(\boldsymbol{x^{(i)}} - \boldsymbol{\mu_k}\right)\boldsymbol{\Sigma^{-T}} = 0$$

Como $\boldsymbol{\Sigma}$ es simétrica a $\boldsymbol{\Sigma^{-1}} = \boldsymbol{\Sigma^{-T}}$:

$$\nabla_{\boldsymbol{\Sigma^{-1}}} \log l(\phi, \mu_k, \Sigma) = \nabla_{\boldsymbol{\Sigma^{-1}}} \sum_{i=1}^{m} -\frac{\boldsymbol{d}}{\mathbf{2}}\boldsymbol{log}(\boldsymbol{2\pi}) + \frac{\mathbf{1}}{\mathbf{2}}\boldsymbol{log}\left(\left|\boldsymbol{\Sigma^{-1}}\right|\right) - \frac{\mathbf{1}}{\mathbf{2}}\left(\boldsymbol{x^{(i)}} - \boldsymbol{\mu_k}\right)^{\boldsymbol{T}}\left(\boldsymbol{x^{(i)}} - \boldsymbol{\mu_k}\right)\boldsymbol{\Sigma^{-1}} = 0$$

$$\nabla_{\boldsymbol{\Sigma^{-1}}} \log l(\phi, \mu_k, \Sigma) = \sum_{i=1}^{m} \nabla_{\boldsymbol{\Sigma^{-1}}} \frac{\mathbf{1}}{\mathbf{2}}\boldsymbol{log}\left(\left|\boldsymbol{\Sigma^{-1}}\right|\right) - \nabla_{\boldsymbol{\Sigma^{-1}}} \frac{\mathbf{1}}{\mathbf{2}}\left(\boldsymbol{x^{(i)}} - \boldsymbol{\mu_k}\right)^{\boldsymbol{T}}\left(\boldsymbol{x^{(i)}} - \boldsymbol{\mu_k}\right)\boldsymbol{\Sigma^{-1}} = 0$$

MODELOS GENERATIVOS

ANÁLISIS DISCRIMINATIVO GAUSSIANO

Aplicando la propiedad $\nabla_x b^T x = b$:

$$\nabla_{\boldsymbol{\Sigma^{-1}}} \log l(\phi, \mu_k, \Sigma) = \frac{\mathbf{1}}{\mathbf{2}} \sum_{i=1}^{m} \nabla_{\boldsymbol{\Sigma^{-1}}} \left[\boldsymbol{log}(\left|\boldsymbol{\Sigma^{-1}}\right|) \right] - \left(\boldsymbol{x^{(i)}} - \boldsymbol{\mu_k}\right)\left(\boldsymbol{x^{(i)}} - \boldsymbol{\mu_k}\right)^{\boldsymbol{T}} = 0$$

Aplicando la propiedad $\nabla_x \log(|X|) = X^{-T}$:

$$\nabla_{\boldsymbol{\Sigma^{-1}}} \log l(\phi, \mu_k, \Sigma) = \sum_{i=1}^{m} \left(\boldsymbol{\Sigma^{-1}}\right)^{-T} - \left(\boldsymbol{x^{(i)}} - \boldsymbol{\mu_k}\right)\left(\boldsymbol{x^{(i)}} - \boldsymbol{\mu_k}\right)^{\boldsymbol{T}} = 0$$

$$m\boldsymbol{\Sigma^{T}} - \sum_{i=1}^{m} \left(\boldsymbol{x^{(i)}} - \boldsymbol{\mu_k}\right)\left(\boldsymbol{x^{(i)}} - \boldsymbol{\mu_k}\right)^{\boldsymbol{T}} = 0$$

$$\boldsymbol{\Sigma^{T}} = \frac{1}{m} \sum_{i=1}^{m} \left(\boldsymbol{x^{(i)}} - \boldsymbol{\mu_k}\right)\left(\boldsymbol{x^{(i)}} - \boldsymbol{\mu_k}\right)^{\boldsymbol{T}}$$

MODELOS GENERATIVOS

ANÁLISIS DISCRIMINATIVO GAUSSIANO

Como Σ es simétrica a $\boldsymbol{\Sigma^T = \Sigma}$:

$$\Sigma = \frac{1}{m}\sum_{i=1}^{m}\left(\boldsymbol{x^{(i)}} - \boldsymbol{\mu_k}\right)\left(\boldsymbol{x^{(i)}} - \boldsymbol{\mu_k}\right)^{\boldsymbol{T}}$$

VARIANZA DE LOS DATOS DE LA CLASE K

Recordando que **es un "outer"** product**:** $\Sigma \in \mathbb{R}^{nxn}$

$$\Sigma = \begin{bmatrix} \frac{1}{m}\sum_{i=1}^{m}\left(\boldsymbol{x^{(i)}_1} - \boldsymbol{\mu_k}\right)^2 & \cdots & \frac{1}{m}\sum_{i=1}^{m}\left(\boldsymbol{x^{(i)}_1} - \boldsymbol{\mu_k}\right)\left(\boldsymbol{x^{(i)}_n} - \boldsymbol{\mu_k}\right) \\ \vdots & \ddots & \vdots \\ \frac{1}{m}\sum_{i=1}^{m}\left(\boldsymbol{x^{(i)}_n} - \boldsymbol{\mu_k}\right)\left(\boldsymbol{x^{(i)}_1} - \boldsymbol{\mu_k}\right) & \cdots & \frac{1}{m}\sum_{i=1}^{m}\left(\boldsymbol{x^{(i)}_n} - \boldsymbol{\mu_k}\right)^2 \end{bmatrix}$$

MODELOS GENERATIVOS

ANÁLISIS DISCRIMINATIVO GAUSSIANO

Una vez que hemos **encontrado** los **parámetros**:

$$\phi = \sum_{i=1}^{m} \frac{1(\boldsymbol{y^{(i)}} = \boldsymbol{k})}{m}$$

$$\boldsymbol{\mu_k} = \frac{\sum_{i=1}^{m} 1(\boldsymbol{y^{(i)}} = \boldsymbol{k})\boldsymbol{x^{(i)}}}{\sum_{i=1}^{m} 1(\boldsymbol{y^{(i)}} = \boldsymbol{k})}$$

$$\Sigma = \frac{1}{m}\sum_{i=1}^{m}\left(\boldsymbol{x^{(i)}} - \boldsymbol{\mu_k}\right)\left(\boldsymbol{x^{(i)}} - \boldsymbol{\mu_k}\right)^{\boldsymbol{T}}$$

Ya podemos realizar **predicciones**.

MODELOS GENERATIVOS

ANÁLISIS DISCRIMINATIVO GAUSSIANO

Para realizar una **predicción** calculamos la **probabilidad** a **posteriori** $\boldsymbol{P(y=1/x)}$ con el nuevo dato $\boldsymbol{x}$:

$$\boldsymbol{p(y=1/x) = \frac{p(x/y=1)p(y=1)}{p(x)}}$$

Asignamos un umbral: si $\boldsymbol{P(y=1/x) \geq 0.5}$ pertenece a la clase 1, sino pertenece a la clase 0. Es decir estamos **maximizando** la **probabilidad**:

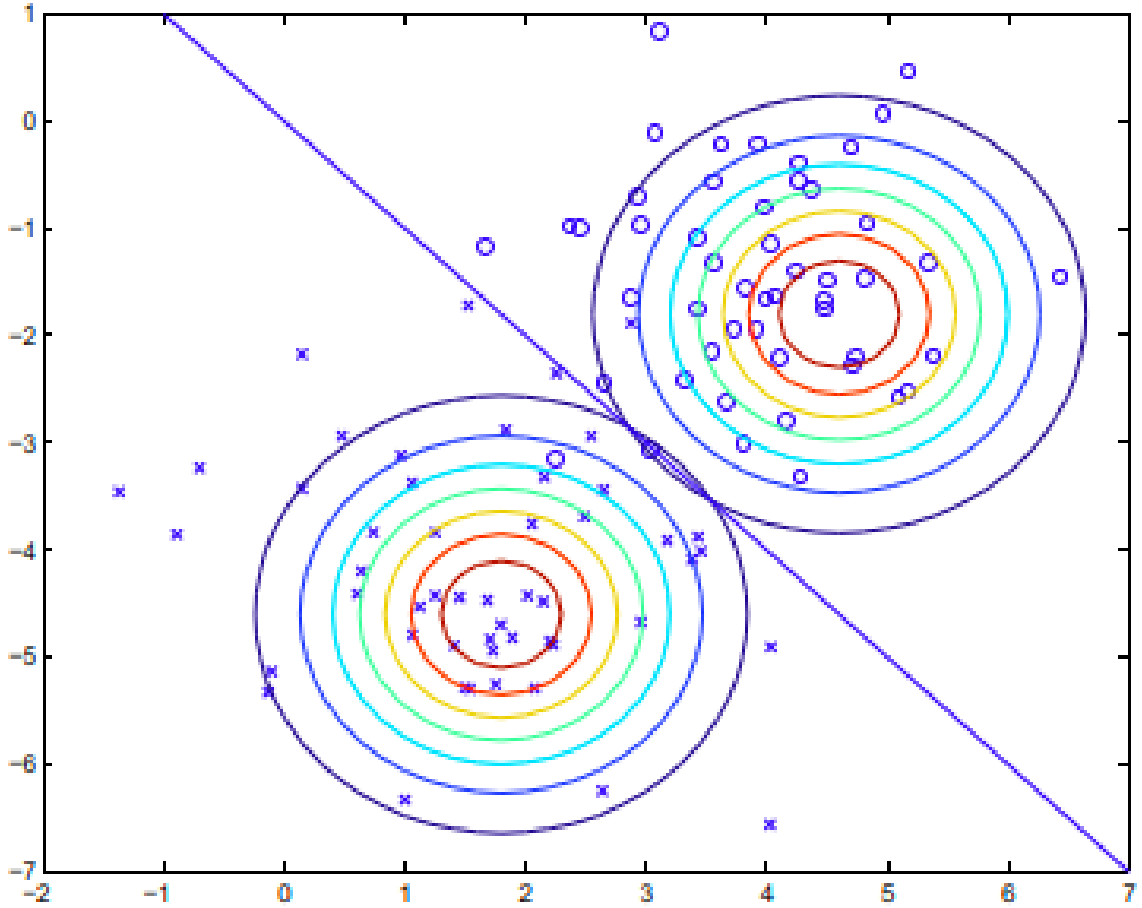
$$\boldsymbol{argmax_y\, p(y/x) = argmax_y \frac{p(x/y)p(y)}{p(x)}}$$

Sí $\boldsymbol{p(y=1) = p(y=0) = 0.5}$ se tiene:

$$\boldsymbol{argmax_y p(y/x) = argmax_y\, p(x/y)}$$

MODELOS GENERATIVOS

ANÁLISIS DISCRIMINATIVO GAUSSIANO



MODELOS GENERATIVOS

ANÁLISIS DISCRIMINATIVO GAUSSIANO

RESUMEN:

1. Obtengo los parámetros para cada Gaussiana (en clasificación binaria serían dos curvas) a partir de los datos.

$$\phi, \mu, \Sigma$$

2. Calculo las probabilidades

$$P(x/y = k) \;\; P(y) \;\; P(x)$$

3. Utilizo Teorema de Bayes para hacer una nueva predicción.

$$\boldsymbol{p(y = k/x) = \frac{p(x/y = k)p(y = k)}{p(x)}}$$

MODELOS GENERATIVOS

COMPARACIÓN MODELOS DISCRIMINATIVOS

Si **analizamos** la **probabilidad** $p(y = 1/x)$, en el **Teorema de Bayes**, podremos darnos cuenta que **obtenemos** la **misma curva** de **regresión logística.**

$$p(y = 1/x) = \frac{p(x/y = 1)p(y = 1)}{p(x)}$$

$$p(y = 1/x) = \frac{p(x/y = 1)p(y = 1)}{p(x/y = 1)p(y = 1) + p(x/y = 0)p(y = 0)} \left(\frac{\frac{1}{p(x/y = 1)p((y = 1)}}{\frac{1}{p(x/y = 1)p((y = 1)} \right)$$

$$p(y = 1/x) = \frac{1}{1 + \frac{p(x/y = 0)p(y = 0)}{p(x/y = 1)p((y = 1)}}$$

MODELOS GENERATIVOS

COMPARACIÓN MODELOS DISCRIMINATIVOS

$$\frac{p(x/y=0)p(y=0)}{p(x/y=1)p((y=1)} = e^{\left(-\frac{1}{2}(x^{(i)}-\mu_0)^T\Sigma^{-1}(x^{(i)}-\mu_0)+\frac{1}{2}(x^{(i)}-\mu_1)^T\Sigma^{-1}(x^{(i)}-\mu_1)\right)}\left(\frac{1-\phi}{\phi}\right)$$

$$\frac{p(x/y=0)p(y=0)}{p(x/y=1)p((y=1)} = e^{\left(-\frac{1}{2}(x^{(i)}-\mu_0)^T\Sigma^{-1}(x^{(i)}-\mu_0)+\frac{1}{2}(x^{(i)}-\mu_1)^T\Sigma^{-1}(x^{(i)}-\mu_1)\right)}\left(e^{\log\left(\frac{1-\phi}{\phi}\right)}\right)$$

$$\frac{p(x/y=0)p(y=0)}{p(x/y=1)p((y=1)} = e^{\left(-\frac{1}{2}\Sigma^{-1}\left[(x^{(i)}-\mu_0)^T(x^{(i)}-\mu_0)-(x^{(i)}-\mu_1)^T(x^{(i)}-\mu_1)\right]+\log\left(\frac{1-\phi}{\phi}\right)\right)}$$

$$\frac{p(x/y=0)p(y=0)}{p(x/y=1)p((y=1)} = e^{\left(-\frac{1}{2}\Sigma^{-1}\left[\sum_{j=1}^{n}(x^{(i)}{}_j-\mu_0)^2-(x^{(i)}{}_j-\mu_1)^2\right]+\log\left(\frac{1-\phi}{\phi}\right)\right)}$$

MODELOS GENERATIVOS

COMPARACIÓN MODELOS DISCRIMINATIVOS

$$\frac{p(x/y=0)p(y=0)}{p(x/y=1)p((y=1)} = e^{\left(-\frac{1}{2}\Sigma^{-1}\left[\sum_{j=1}^{n} -2\mu_0 x^{(i)}{}_j+\mu_0{}^2+2\mu_1 x^{(i)}{}_j-\mu_1{}^2\right]+\log\left(\frac{1-\phi}{\phi}\right)\right)}$$

Como $x^{(i)}{}_0 = 1$

$$\frac{p(x/y=0)p(y=0)}{p(x/y=1)p((y=1)} = e^{\left(-\left[\sum_{j=1}^{n}\frac{1}{2}\Sigma^{-1}(\mu_0-\mu_1)(x^{(i)}{}_j)\right]+\left[\frac{1}{2}\Sigma^{-1}(\mu_0{}^2-\mu_1{}^2)+\log\left(\frac{1-\phi}{\phi}\right)\right](x^{(i)}{}_0)\right)}$$

Comparando $w^T X$:

$$w^T = \begin{bmatrix} -\frac{1}{2}\Sigma^{-1}(\mu_0{}^2-\mu_1{}^2) - \log\left(\frac{1-\phi}{\phi}\right) \\ \frac{1}{2}\Sigma^{-1}(\mu_0-\mu_1) \\ \vdots \\ \frac{1}{2}\Sigma^{-1}(\mu_0-\mu_1) \end{bmatrix}$$

MODELOS GENERATIVOS

COMPARACIÓN MODELOS DISCRIMINATIVOS

Por lo tanto:

$$p(y=1/x)=\frac{1}{1+e^{-\left(\left[\sum_{j=1}^{n}\frac{1}{2}\Sigma^{-1}(\mu_0-\mu_1)(x^{(i)}{}_j)\right]+\left[-\frac{1}{2}\Sigma^{-1}(\mu_0{}^2-\mu_1{}^2)+\log\left(\frac{1-\phi}{\phi}\right)\right](x^{(i)}{}_0)\right)}}=\boxed{\frac{1}{1+e^{-W^TX}}}$$

Donde w^T :

$$w^T=\begin{bmatrix}-\frac{1}{2}\Sigma^{-1}(\mu_0{}^2-\mu_1{}^2)-\log\left(\frac{1-\phi}{\phi}\right)\\ \frac{1}{2}\Sigma^{-1}(\mu_0-\mu_1)\\ \vdots\\ \frac{1}{2}\Sigma^{-1}(\mu_0-\mu_1)\end{bmatrix}$$

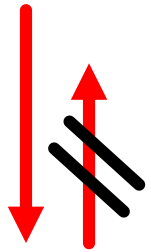
MODELOS GENERATIVOS

COMPARACIÓN MODELOS DISCRIMINATIVOS

De manera más general:

Función de Verosimilitud

$$x/y \sim FamiliaExponencial(\eta)$$



Distribución Posterior

$$P(y/x) = \textit{función sigmoide}$$

ESA ES LA RAZÓN PORQUE SE USA REGRESIÓN LOGÍSTICA → FUNCIONA PARA MUCHOS TIPOS DE SUPOSICIONES

MODELOS GENERATIVOS

COMPARACIÓN MODELOS DISCRIMINATIVOS

Un **modelo generativo** tendrá **mejor rendimiento** que uno discriminativo, sí la **suposición** que hicimos de la forma de la **distribución** $\boldsymbol{P(x/y)}$ **se cumple para los datos reales** (se provee de más información al algoritmo)**.**

De otra forma, si los **datos no** se **comportan** como **supusimos**, el **modelo discriminativo** tendrá **mejores resultados**, porque **aún** cuando nuestras **suposiciones no** fueron tan **certeras**, la **regresión logística** funciona para **muchas suposiciones distintas**.

GENERATIVO	DISCRIMINATIVO
Mejor rendimiento cuando se conoce la distribución P(x/y)	Mejor rendimiento cuando se desconoce P(x/y) (robusto a suposiciones incorrectas)
Eficiencia asintótica	Puede existir un mejor modelo
Necesita pocos datos	Necesita muchos datos

MODELOS GENERATIVOS

CLASIFICADOR INGENUO DE BAYES

MODELO DE EVENTOS MULTIVARIADO DE BERNOULLI:

Para el caso del **análisis discriminativo Gaussiano**, se tenía el caso de que $x^{(i)}_j \in \mathbb{R}$**,** es decir, eran **valores continuos**.

Para casos**,** como **clasificación** de **textos (“correo deseado”** y **“correo no deseado”)** los **vectores x** pueden tener una **dimensionalidad muy alta (**el **número de palabras** es **inmenso).**

$$\boldsymbol{x \in \{0,1\}^{5000}}$$

$$x = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} \begin{matrix} \text{a} \\ \text{aardvark} \\ \text{aardwolf} \\ \vdots \\ \text{buy} \\ \vdots \\ \text{zygmurgy} \end{matrix}$$

Se **necesitarían** $\boldsymbol{2^{5000}}$ **parámetros**.

MODELOS GENERATIVOS

CLASIFICADOR INGENUO DE BAYES

SOLUCIÓN: suposición ingenua de Bayes (muy fuerte)

Suponemos que $x^{(i)}_j$ son **condicionalmente independientes dado** y.

Es decir, si se sabe que un texto es **"correo no deseado"** $y = 1$, el hecho de que aparezca la palabra $x_{2087} = "comprar"$ en el texto, **no afecta las creencias** sobre la aparición de cualquier otra palabra, como $x_{39831} = "precio"$.

$$p(x_1, \dots x_j, \dots, x_n/y) = p(x_1)p(x_2/x_1, y)p(x_3/x_1, x_2, y) \dots = p(x_1/y)p(x_2/y) \dots = \prod_{j=1}^{n} p(x_j/y)$$

AÚN CUANDO LA SUPOSICIÓN NO ES CIERTA, EL ALGORITMO TIENE BUEN RENDIMIENTO EN MUCHAS APLICACIONES

MODELOS GENERATIVOS

CLASIFICADOR INGENUO DE BAYES

Como el **clasificador** de **Bayes** es un **modelo generativo**, es de interés **modelar** las **distribuciones** $p(x_j/y)$ y $p(y)$. De manera particular se quiere **maximizar** la **verosimilitud conjunta**:

$$l(\phi_y, \phi_{j/y=0}\phi_{j/y=1}) = \prod_{i=1}^{m} p(x^{(i)}, y^{(i)})$$

Donde las **suposiciones son**:

$$p(x/y) = p(x_1, \dots x_j, \dots, x_n/y) = \prod_{j=1}^{n} p(x_j/y)$$

$$x_j \,/\, y = 0 \sim Bernoulli(\phi_{j/y=0})$$

$$x_j \,/\, y = 1 \sim Bernoulli(\phi_{j/y=1})$$

$$y \sim Bernoulli(\phi_j)$$

MODELOS GENERATIVOS

CLASIFICADOR INGENUO DE BAYES

El **resultado** de la **estimación** por **máxima verosimilitud** da:

$$\phi_{j/y=1} = \frac{\sum_{i=1}^{m} 1\{x_j^{(i)}, y^{(i)} = 1\}}{\sum_{i=1}^{m} 1\{y^{(i)} = 1\}} = \frac{\#\ veces\ que\ se\ repite\ la\ palabra\ j\ en\ "no\ deseado"}{total\ de\ correos\ "no\ deseados"}$$

$$\phi_{j/y=0} = \frac{\sum_{i=1}^{m} 1\{x_j^{(i)}, y^{(i)} = 0\}}{\sum_{i=1}^{m} 1\{y^{(i)} = 0\}} = \frac{\#\ veces\ que\ se\ repite\ la\ palabra\ j\ en\ "deseado"}{total\ de\ correos\ "deseados"}$$

$$\phi_y = \frac{\sum_{i=1}^{m} 1\{y^{(i)} = 1\}}{m} = \frac{total\ de\ correos\ "no\ deseados"}{total\ de\ correos}$$

Si las **características** $x_j^{(i)}$ toman **más valores**, se pueden modelar como **multinomiales**.

Si las **características** $x_j^{(i)}$ toman valores **continuos**, se **discretiza** en **intervalos**.

MODELOS GENERATIVOS

CLASIFICADOR INGENUO DE BAYES

Para **realizar** una **nueva predicción** usamos el **teorema** de **Bayes**:

$$p(y = 1/x) = \frac{p(x/y = 1)p(y = 1)}{p(x)}$$

Donde:

$$p(x/y = 1\,) = \prod_{j=1}^{n} p(x_j/y = 1) = \prod_{j=1}^{n} (\phi_j/y = 1)$$

$$p(y) = \phi_y = \frac{\sum_{i=1}^{m} 1\{y^{(i)} = 1\}}{m}$$

$$p(x) = \prod_{j=1}^{n} p\big(x_j/y = 1\big)\, p(y = 1) + \prod_{j=1}^{n} p\big(x_j/y = 0\big)\, p(y = 0)$$

MODELOS GENERATIVOS

CLASIFICADOR INGENUO DE BAYES

¿Qué pasa sí aparece en un correo una palabra que nunca ha visto el algoritmo en otro correo?

MODELOS GENERATIVOS

CLASIFICADOR INGENUO DE BAYES

La **probabilidad** que asignará, de **ver** la **palabra** en **cualquiera** de los **dos correos** será **0**.

$$\phi_{j/y=1} = \frac{\sum_{i=1}^{m} 1\{x_j^{(i)}, y^{(i)} = 1\}}{\sum_{i=1}^{m} 1\{y^{(i)} = 1\}} = \frac{\#\ veces\ que\ se\ repite\ la\ palabra\ j\ en\ "no\ deseado"}{total\ de\ correos\ "no\ deseados"} = \frac{0}{m_{neg}}$$

$$\phi_{j/y=0} = \frac{\sum_{i=1}^{m} 1\{x_j^{(i)}, y^{(i)} = 0\}}{\sum_{i=1}^{m} 1\{y^{(i)} = 0\}} = \frac{\#\ veces\ que\ se\ repite\ la\ palabra\ j\ en\ "deseado"}{total\ de\ correos\ deseados"} = \frac{0}{m_{pos}}$$

Al hacer la **predicción** tendremos una **incongruencia**:

$$\boldsymbol{p(y = 1/x) = \frac{p(x/y = 1)p(y = 1)}{p(x)} = \frac{0}{0}}$$

MODELOS GENERATIVOS

CLASIFICADOR INGENUO DE BAYES

ESTADÍSTICAMENTE NO ES ADECUADO DECIR QUE LA PROBABILIDAD DE UN EVENTO ES CERO SOLO PORQUE NO LO HAYAS VISTO EN TUS DATOS

NOTA: LAPLACE Y EL SOL

MODELOS GENERATIVOS

CLASIFICADOR INGENUO DE BAYES

SUAVIZADO DE LAPLACE:

La **solución** es **cambiar** nuestra **estimación** (caso general multinomial donde $y = \{1, 2, \ldots, k\}$):

$$\boldsymbol{p(y = 1) = \frac{(\#\ \text{"1"s} + 1)}{(\#\text{"0"s} + 1) + (\#\ \text{"1"s} + 1)}}$$

$$\boldsymbol{p(y = j) = \phi_y = \frac{1 + \sum_{i=1}^{m} 1\{y^{(i)} = j\}}{m + k}}$$

Al calcular los **otros estimadores** se tiene:

$$\boldsymbol{\phi_{j/y=1} = \frac{1 + \sum_{i=1}^{m} 1\{x_j^{(i)}, y^{(i)} = 1\}}{2 + \sum_{i=1}^{m} 1\{y^{(i)} = 1\}}}$$

$$\boldsymbol{\phi_{j/y=0} = \frac{1 + \sum_{i=1}^{m} 1\{x_j^{(i)}, y^{(i)} = 0\}}{2 + \sum_{i=1}^{m} 1\{y^{(i)} = 0\}}}$$

MODELOS GENERATIVOS

CLASIFICADOR INGENUO DE BAYES

MODELO DE EVENTOS MULTINOMIAL:

Vamos a ver el caso cuando $\boldsymbol{x_j = \{1, 2, \dots, k\}}$.

Por ejemplo, ahora se tiene que el valor de $\boldsymbol{x_j}$ representa la **posición** de la **palabra** en el **diccionario**, y el índice $\boldsymbol{j}$ indica la **posición** de la **palabra** en el **correo**.

Entonces $\boldsymbol{n}$ ahora representa la **longitud** del correo (y **varía acorde** a cada **correo**).

Las **suposiciones** serían:

$$\boldsymbol{p(x/y) = p\left(x_1, \dots x_j, \dots, x_n/y\right) = \prod_{j=1}^{n} p(x_j/y)}$$

$$\boldsymbol{x_j \,/\, y = 0 \sim Multinomial(\phi_{j/y=0})}$$

$$\boldsymbol{x_j \,/\, y = 1 \sim Multinomial(\phi_{j/y=1})}$$

$$\boldsymbol{y \sim Bernoulli(\phi_j)}$$

MODELOS GENERATIVOS

CLASIFICADOR INGENUO DE BAYES

Los **parámetros**, al **calcular** la **máxima verosimilitud** serían (con suavizado de Laplace):

$$\phi_{k/y=1} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i} 1\{x_j^{(i)} = k, y^{(i)} = 1\} + 1}{\sum_{i=1}^{m} 1\{y^{(i)} = 1\}\, d_i + n_i + k} = \frac{\#\ de\ veces\ que\ aparece\ la\ palabra\ k\ en\ correos\ ND}{total\ de\ palabras\ en\ correos\ ND}$$

$$\phi_{k/y=0} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i} 1\left\{x_j^{(i)} = k, y^{(i)} = 0\right\} + 1}{\sum_{i=1}^{m} 1\{y^{(i)} = 0\}\, d_i + n_i + k} = \frac{\#\ de\ veces\ que\ aparece\ la\ palabra\ k\ en\ correos\ D}{total\ de\ palabras\ en\ correos\ D}$$

$$\phi_y = \frac{\sum_{i=1}^{m} 1\left\{y^{(i)} = 1\right\} + 1}{m}$$

MÉTRICAS DE EVALUACIÓN BINARIAS

MÉTRICAS DE EVALUACION BINARIAS

D E F I N I C I O N E S

Falso positivo (error tipo I):

- **Estadística**: Se **rechaza incorrectamente** una **hipótesis nula verdadera**.
- **Aprendizaje máquina**: el **modelo predice** que la **clase** de un dato de entrenamiento $\boldsymbol{p(y/x) = 1}$ es **positiva**, cuando la **realidad** es que el **dato pertenecía** a la clase **negativa** $\boldsymbol{y = 0}$**.**

MÉTRICAS DE EVALUACION BINARIAS

D E F I N I C I O N E S

Falso negativo (error tipo II):

- **Estadística**: Se **acepta incorrectamente** la **hipótesis nula falsa**.
- **Aprendizaje máquina**: el **modelo predice** que la **clase** de un dato de entrenamiento $\boldsymbol{p(y/x) = 0}$ es **negativa**, cuando la **realidad** es que el **dato pertenecía** a la clase **positiva** $\boldsymbol{y = 1}$**.**

MÉTRICAS DE EVALUACION BINARIAS

D E F I N I C I O N E S

Verdadero positivo:

- **Estadística**: Se **acepta correctamente** la **hipótesis nula falsa**.
- **Aprendizaje máquina**: el **modelo predice** que la **clase** de un dato de entrenamiento $\boldsymbol{p(y/x) = 1}$ es **positiva**, cuando la **realidad** es que el **dato pertenecía** a la clase **positiva** $\boldsymbol{y} = 1$**.**

MÉTRICAS DE EVALUACION BINARIAS

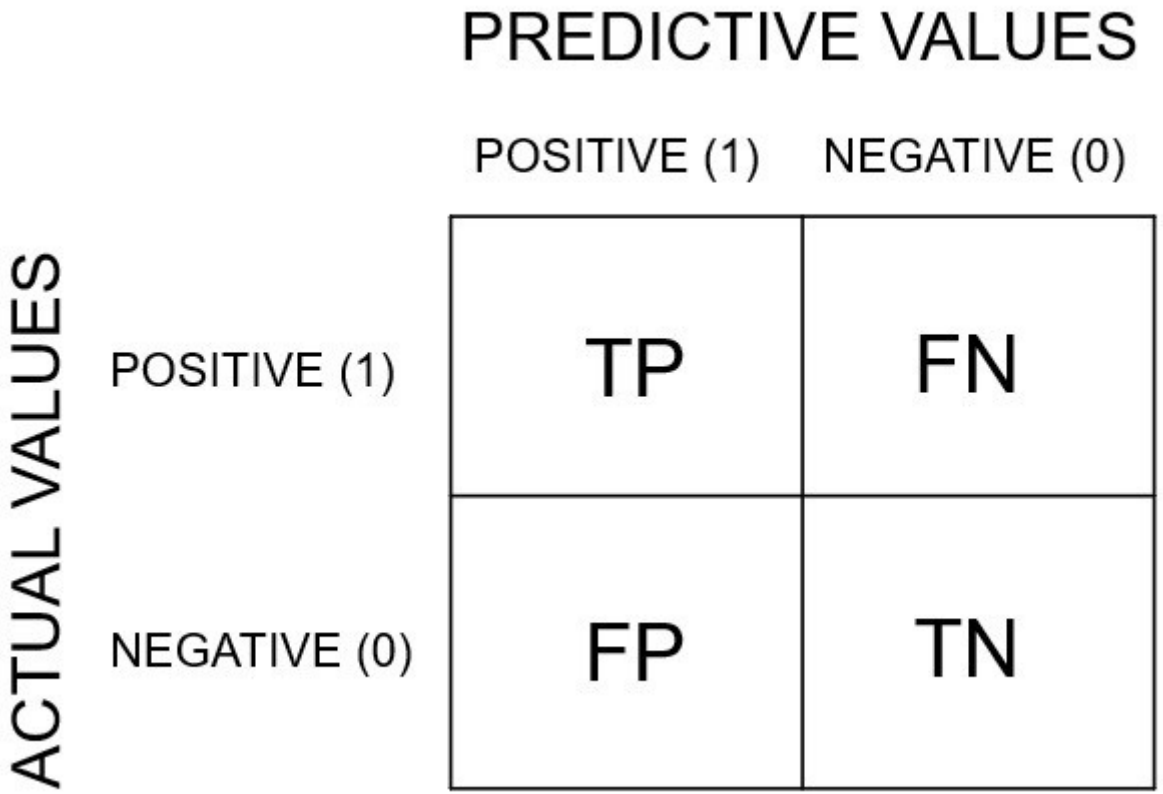
D E F I N I C I O N E S

Verdadero negativo:

- **Estadística**: Se **acepta correctamente** la **hipótesis nula verdadera**.
- **Aprendizaje máquina**: el **modelo predice** que la **clase** de un dato de entrenamiento $\boldsymbol{p(y/x) = 0}$ es **negativa**, cuando la **realidad** es que el **dato pertenecía** a la clase **negativa** $\boldsymbol{y = 0}$**.**

MÉTRICAS DE EVALUACION BINARIAS

MATRIZ DE CONFUSIÓN



MÉTRICAS DE EVALUACION BINARIAS

S E N S I B I L I D A D

La **sensibilidad** (“**true positive rate**”, “**recall**”): mide el **rendimiento** del **modelo** para **predecir** la **clase** positiva $\boldsymbol{y} = 1$.

Es difícil que un **algoritmo sensible** se **equivoque** en **predecir** la clase **positiva**.

Aún así, una **alta sensibilidad puede** venir **acompañada** de **muchos falsos positivos**.

$$\boldsymbol{Sens} = \frac{\boldsymbol{TP}}{\boldsymbol{TP} + \boldsymbol{FN}}$$

MÉTRICAS DE EVALUACION BINARIAS

E S P E C I F I C I D A D

La **especificidad** (“**true negative rate**”): mide el **rendimiento** del **modelo** para **predecir** la **clase** negativa $\boldsymbol{y} = 0$.

Es difícil que un **algoritmo negativo** se **equivoque** en **predecir** la clase **negativa**.

Aún así, una **alta especificidad puede** venir **acompañada** de **muchos falsos negativos**.

$$\boldsymbol{Spec} = \frac{\boldsymbol{TN}}{\boldsymbol{TN} + \boldsymbol{FP}}$$

MÉTRICAS DE EVALUACION BINARIAS

P R E C I S I Ó N

La **precisión** o “**accuracy**” (no confundir con “**positive predictive value” o “precisión”)** mide el **rendimiento** del **modelo** de **manera general. Cuanto** se **desvían** las **predicciones** de los **valores reales.**

$$\boldsymbol{Acc = \frac{TP + TN}{TP + TN + FN + FP}}$$

MÉTRICAS DE EVALUACION BINARIAS

VALOR POSITIVO PREDICHO

El “**positive predictive value” o “precisión”,** mide la **fracción** de **positivos** que fueron **correctamente predichos**.

$$PPV = \frac{TP}{TP + FP}$$

MÉTRICAS DE EVALUACION BINARIAS

P U N T A J E F 1

El **puntaje F1** es la **media armónica** (promedio para **tasas**) entre la **sensibilidad** (“**recall**”) y el **PPV:**

$$\boldsymbol{Sens = \frac{TP}{TP + FN}}$$

$$\boldsymbol{PPV = \frac{TP}{TP + FP}}$$

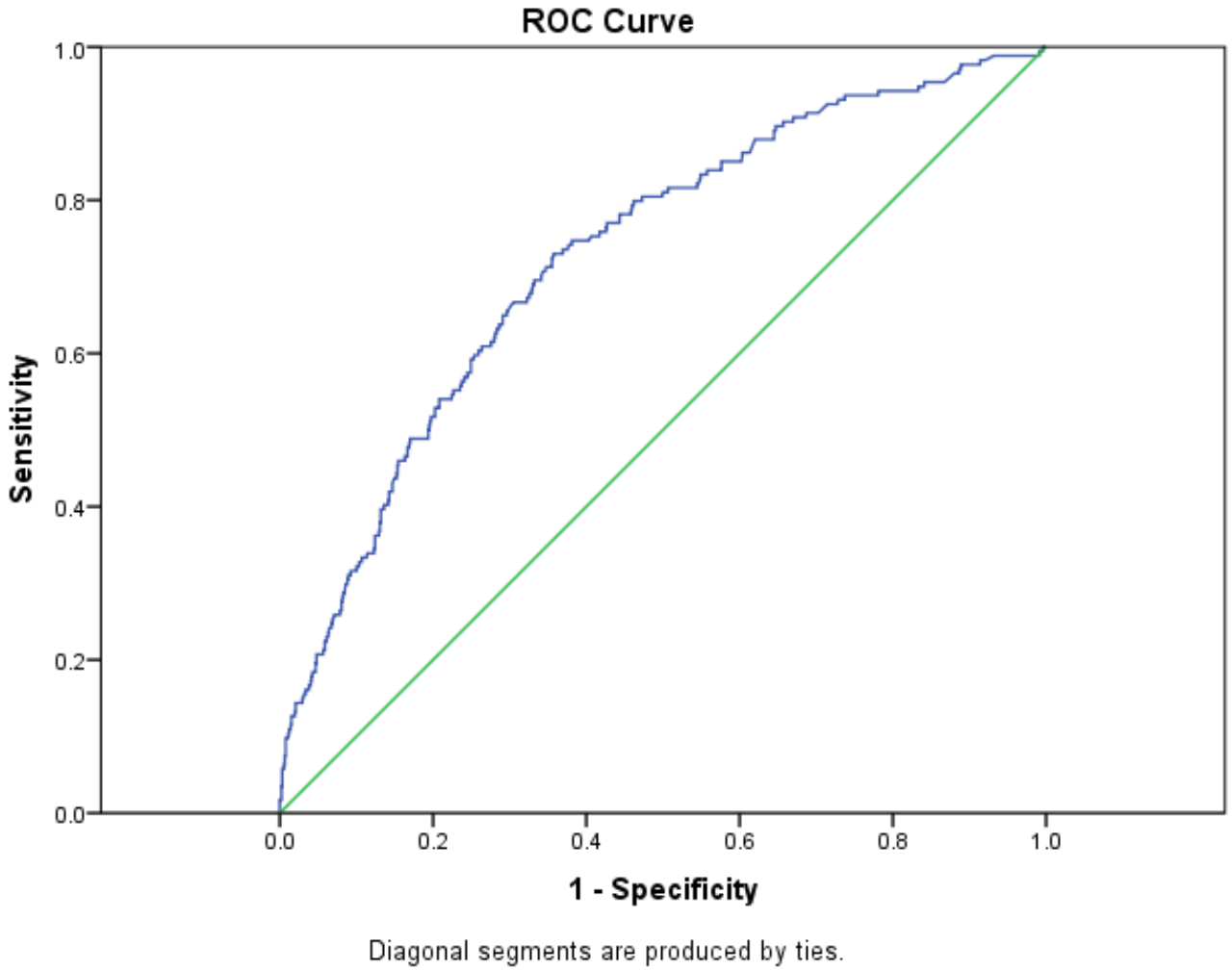
$$\boldsymbol{F1 = \frac{2}{Sens^{-1} + PPV^{-1}}}$$

MÉTRICAS DE EVALUACION BINARIAS

C U R V A R O C

FPR vs TPR

Se varía el umbral



MÉTRICAS DE EVALUACION BINARIAS

C U R V A R O C