Tema 4: CLASIFICADORES PARAMÉTRICOS: MODELOS DE REGRESIÓN

MINERÍA DE DATOS

Alicia Pérez
alicia.perez@ehu.eus

Lengoaia eta Sistema Informatikoak Saila
Bilboko Ingeniaritza Eskola

Universidad del País Vasco / Euskal Herriko Unibertsitatea

Índice

Introducción
Definición del problema

Modelización: [Orallo et al., 2004, Cap. 7]

- Variables de entrada: independientes, predictoras, explicativas, *regressor variables*, . . .
- Variable de salida: dependiente, variable de respuesta, *outcome variable*, . . .
- Regresión: tipo variables entrada/salida cuantitativas

Sean $(x_{i1}, \cdots, x_{ip})$, las variables predictoras de la instancia i, entonces, la variable respuesta, y_i, se determina como sigue:

$$y_i = r(x_{i1}, \cdots, x_{ip}) + \epsilon_i \quad (1)$$

donde:

- r: función de las variables predictoras, representa la parte estructural determinista del modelo
- ϵ_i: parte específica asociado al individuo i, representa la parte aleatoria, y se denomina "término de error"

Introducción
Objetivo

- Objetivo: estimar $r(x_1, \cdots, x_p)$
- Método:
  - Buscar la función $r(\cdot)$ que minimice el error cuadrático medio de la muestra $(\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n)$:

$$\text{mín } E[(e_i)^2] = \text{mín } E[(y_i - r(x_{i1}, \cdots, x_{ip}))^2] \quad (2)$$

  - ¿Qué función minimiza el error cuadrático medio?
    La media condicional de y_i respecto de $x_{i1}, \cdots, x_{ip}$

$$r(x_1, \cdots, x_p) = E\left[y_i | x_{i1}, \cdots, x_{ip}\right] \text{ función de regresión} \quad (3)$$

Modelo de regresión lineal
Un atributo

Función lineal de 1 variable (atributo):

$$\widehat{Y} = f(X) = w_0 + w_1 \cdot X \tag{4}$$

donde los parámetros que caracterizan el modelo predictor son: [James et al., 2013]

- w_0: *intercept*
- w_1: *pendiente*

Utilizamos el conjunto de instancias, $\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^{n}$, para estimar w_0 y w_1 minimizando la suma del cuadrado de los residuos (RSS).

$$RSS = \sum_{i=1}^{n} \left(e^{(i)}\right)^2 = \sum_{i=1}^{n} \left(y^{(i)} - \widehat{y}^{(i)}\right)^2 = \sum_{i=1}^{n} \left(y^{(i)} - w_0 - w_1 \cdot x^{(i)}\right)^2 \tag{5}$$

Modelo de regresión lineal
Un atributo

Ejercicio:

Calcular los coeficientes de forma analítica:

$$\begin{cases} 0 = \frac{\partial RSS}{\partial w_0} = \text{calcular} \\ 0 = \frac{\partial RSS}{\partial w_1} = \text{calcular} \end{cases} \Rightarrow \begin{cases} w_1 = \\ w_0 = \end{cases}$$

Ejercicio:

Alternativamente, calcular los coeficientes de forma estadística (en base a los datos)

$$\Rightarrow \begin{cases} w_1 = \\ w_0 = \end{cases}$$

$$\begin{cases} w_1 = \dfrac{\sum_{i=1}^{n} x^{(i)} y^{(i)} - \frac{1}{n}\left[\left(\sum_{i=1}^{n} x^{(i)}\right) \cdot \left(\sum_{i=1}^{n} y^{(i)}\right)\right]}{\sum_{i=1}^{n} (x^{(i)})^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x^{(i)}\right)^2} \\ w_0 = \bar{y} - w_1 \bar{x} \end{cases} \tag{6}$$

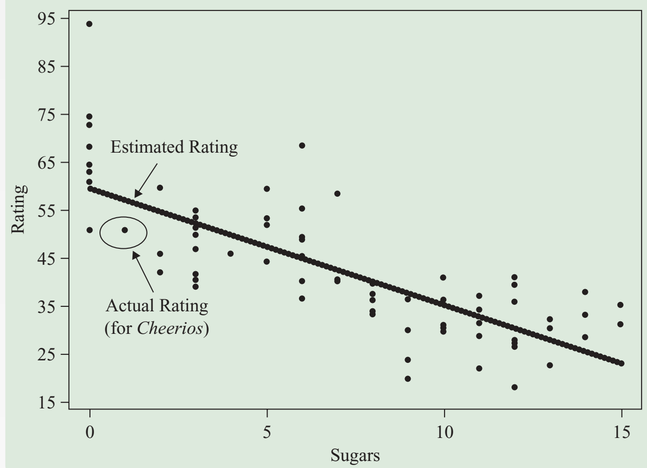
Modelo de regresión lineal
Un atributo

Ejercicio: Calcular el modelo de regresión lineal con los siguientes datos

Cereal Name	Sugars, x	Rating, y	xy	x^2
100% Bran	6	68.4030	410.418	36
100% Natural Bran	8	33.9837	271.870	64
All-Bran	5	59.4255	297.128	25
All-Bran Extra Fiber	0	93.7049	0.000	0
Almond Delight	8	34.3848	275.078	64
Apple Cinnamon Cheerios	10	29.5095	295.095	100
Apple Jacks	14	33.1741	464.437	196
Basic 4	8	37.0386	296.309	64
Bran Chex	6	49.1203	294.722	36
Bran Flakes	5	53.3138	266.569	25
Cap'n Crunch	12	18.0429	216.515	144
Cheerios	1	50.7650	50.765	1
Cinnamon Toast Crunch	9	19.8236	178.412	81
Clusters	7	40.4002	282.802	49
Cocoa Puffs	13	22.7364	295.573	169
⋮	⋮	⋮		
Wheaties Honey Gold	8	36.1876	289.501	64
	$\sum x_i = 534$ $\bar{x} = 534/77$ $= 6.935$	$\sum y_i = 3285.26$ $\bar{y} = 3285.26/77$ $= 42.6657$	$\sum x_i y_i$ $= 19,186.7$	$\sum x_i^2$ $= 5190$

Modelo de regresión lineal
Un atributo

Solución:



$\widehat{y} = 59{,}4 - 2{,}42x$

Fuente: [Larose, 2006]

Modelo de regresión lineal
Múltiples atributos

Función lineal de k variables (atributos):

$$f(X_1, X_2, \cdot, X_k) = w_0 + w_1 \cdot X_1 + w_2 \cdot X_2 + \cdots + w_k \cdot X_k = \sum_{j=0}^{k} w_j X_j \qquad x_0 \equiv 1 \qquad (7)$$

Estimación de la función de regresión lineal: [Witten et al., 2011, Sec.4.6]

- Disponemos de un conjunto de n instancias de entrenamiento. La instancia (i):

  Clase real: $y^{(i)}$

  atributos: $(x_1^{(i)}, x_2^{(i)}, \cdots, x_k^{(i)})$

- Clase estimada:

$$\widehat{y}^{(i)} = f(x_1^{(i)}, x_2^{(i)}, \cdots, x_k^{(i)}) = \sum_{j=0}^{k} w_j x_j^{(i)} \qquad (8)$$

- Objetivo: selecionar $(w_0, w_1, w_2, \cdots, w_k)$ tal que se minimice la diferencia de sumas de cuadrados (*Residual Sum of Squares*)

$$RSS = \sum_{i=0}^{n} (e_i)^2 = \sum_{i=0}^{n} \left(y^{(i)} - \widehat{y}^{(i)}\right)^2 = \sum_{i=0}^{n} \left(y^{(i)} - \sum_{j=0}^{k} w_j x_j^{(i)}\right)^2 \qquad (9)$$

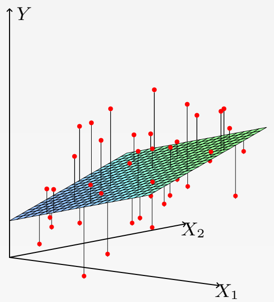
Modelo de regresión lineal
Múltiples atributos

Ejemplo: Weka

- Instancias: `diabetes.arff`
- Convertir clase: nominal → {0,1}
  - Set class: `class (Nom)` → `No class`
  - Filtro `attribute.NominalToBinary`
- Clasificador: `functions.LinearRegression`
- Options: ✓ `Oputput predictions`

Modelo de regresión lineal
Múltiples atributos

- 1 atributo: calcular la recta que minimice el RSS.
- 2 atributos: calcular el plano que minimice el RSS.

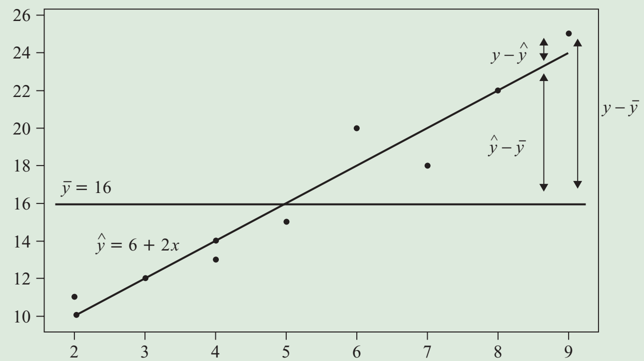


- k atributos: calcular el hiper-plano que minimice el RSS.

Modelo de regresión lineal
Múltiples atributos

Ejercicio: ¿qué modelo elegir: $\bar{y}$ o $\widehat{y}$?

Para el conjunto de datos que se muestra en la figura se proponen dos modelos de regresión: 1) constante que estima la variable respuesta como media de las respuestas observadas: $f_1(x) = \bar{y}$; 2) un modelo de regresión lineal $f_2(x) = 6 + 2x$. ¿Cual de los dos modelos tiene un error predictivo menor? [Larose, 2006]



Modelo de regresión lineal
Evaluación

- SSE≡RSS:

$$SSE \equiv \sum_{i=1}^{n}(y - \hat{y})^2 \quad (10)$$

- SSR:

$$SSR \equiv \sum_{i=1}^{n}(\hat{y} - \bar{y})^2 \quad (11)$$

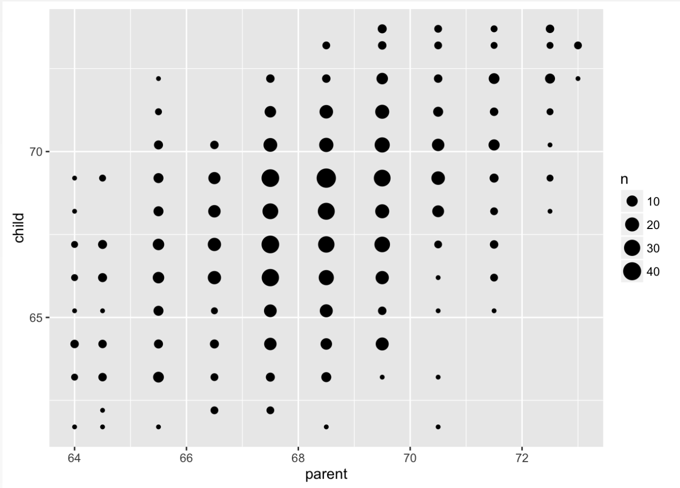
- SST:

$$SST \equiv \sum_{i=1}^{n}(y - \bar{y})^2 \quad (12)$$

- R^2:

$$R^2 \equiv \frac{SSR}{SST} \quad (13)$$

Modelo de regresión lineal
Evaluación

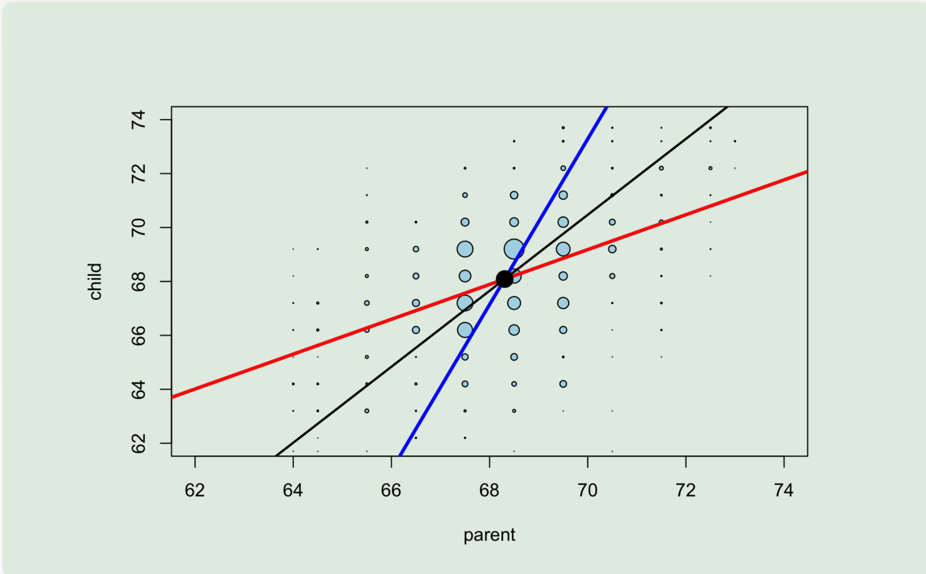


Francis Galton,“Regression towards mediocrity in hereditary stature”, Journal of the Anthropological Institute of Great Britain and Ireland, Vol. 15, (1886).

Modelo de regresión lineal
Evaluación

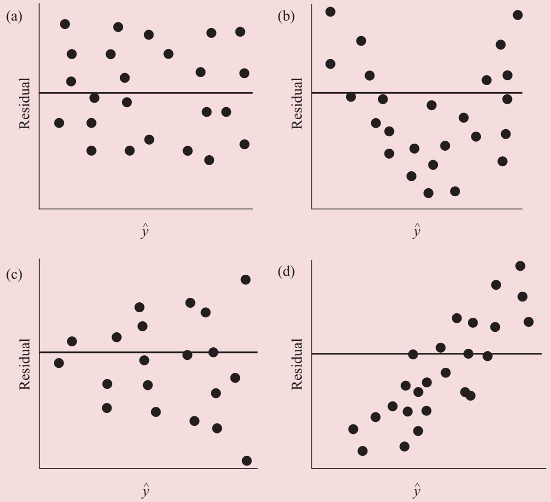
```
freqData <- as.data.frame(table(galton$child, galton$parent))
names(freqData) <- c("child", "parent", "freq")
plot(as.numeric(as.vector(freqData$parent)),
     as.numeric(as.vector(freqData$child)),
     pch = 21, col = "black", bg = "lightblue",
     cex = .05 * freqData$freq,
     xlab = "parent", ylab = "child", xlim = c(62, 74), ylim = c(62, 74))
abline(mean(y) - mean(x) * cor(y, x) * sd(y) / sd(x), sd(y) / sd(x) * cor(y, x), lwd = 3, col = "red")
abline(mean(y) - mean(x) * sd(y) / sd(x) / cor(y, x), sd(y) / sd(x) / cor(y, x), lwd = 3, col = "blue")
abline(mean(y) - mean(x) * sd(y) / sd(x), sd(y) / sd(x), lwd = 2)
points(mean(x), mean(y), cex = 2, pch = 19)
```


Modelo de regresión lineal
Evaluación



Modelo de regresión lineal
Evaluación

¿Los datos se corresponden con el modelo propuesto?



Regresión polinómica
Un atributo

Función polinómica de orden n:
$y = f(x) = a_0 + a_1 \cdot x + a_2 \cdot x^2 + \cdots + a_n \cdot x^n$

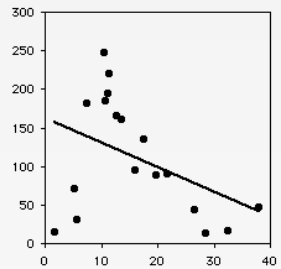


Figura: *Linear regression:* Significance: r^2=0.174, 16 d.f., P=0.08

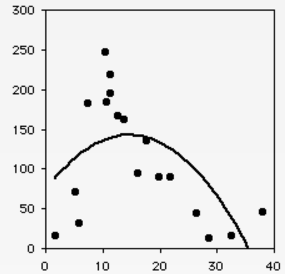


Figura: *Quadratic regression:* Significance: r^2=0.372, 15 d.f., P=0.03

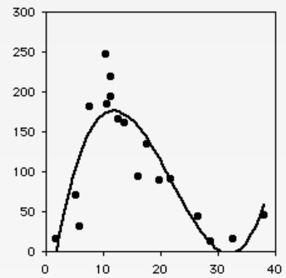


Figura: *Cubic regression:* Significance: r^2=0.765, 14 d.f., P=0.0001

Fuente de las figuras: `http://udel.edu/~mcdonald/statcurvreg.html`

Regresión logística

- Regresión lineal: la respuesta es continua, se aproxima mediante un modelo lineal de las variables de predictoras
- Regresión logística: la respuesta es discreta, construye un modelo lineal basado en una transformada de la variable de salida.

Regresión logística

Ejemplo: 1 atributo

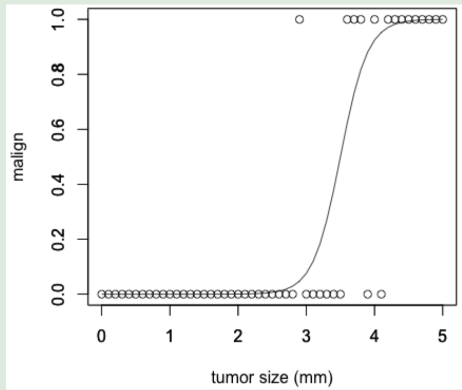


Figura: Datos de entrenamiento y sigmoide

- Problema de clasificación
  - x: tamaño del tumor
  - $y \in \{benigno, maligno\} = \{0, 1\}$
- Sigmoide (función logística):

$$h_\theta(x) = \frac{1}{1 + exp(-(\theta_0 + \theta_1 x))} \tag{14}$$

- Sigmoide para clasificar:

$$\hat{y} = \begin{cases} 0 & si \quad h_\theta(x) < 0{,}5 \\ 1 & si \quad h_\theta(x) \geq 0{,}5 \end{cases}$$

- Interpretación: $h_\theta(x)$ estimación de la probabilidad de que y=1 para x

Regresión logística

Sigmoide: 1 atributo

$$h_\theta(x) = \frac{1}{1+exp(-\theta x)} \equiv g(\theta x)$$

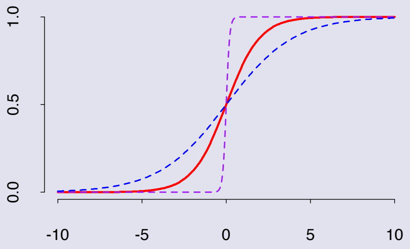


Figura: Sigmoides con distintos valores de θ

Sigmoide: múltiples atributos

$$h_\theta(\mathbf{x}) = \frac{1}{1+exp(-\theta^T\cdot\mathbf{x})} \equiv g(\theta^T\cdot\mathbf{x})$$

Observación:

- $\theta^T\cdot\mathbf{x} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots$
- $\mathbf{x} = (x_0 \equiv 1, x_1, x_2, \cdots)$

Regresión logística

Múltiples atributos:

- Sigmoide (función logística):

$$h_\theta(\mathbf{x}) = \frac{1}{1+exp(-\theta^T\cdot\mathbf{x})} \equiv g(\theta^T\cdot\mathbf{x}) \quad (15)$$

- Predicción:

$$\hat{y} = \begin{cases} 1 & si \quad h_\theta(\mathbf{x}) \geq 0{,}5 \Leftrightarrow \theta^T\cdot\mathbf{x} \geq 0 \\ 0 & si \quad h_\theta(\mathbf{x}) < 0{,}5 \Leftrightarrow \theta^T\cdot\mathbf{x} < 0 \end{cases}$$

- Interpretación: $h_\theta(\mathbf{x})$ estimación de la probabilidad de que y=1 para x

$$P(y=1|\mathbf{x};\theta) = h_\theta(\mathbf{x}) \quad (16)$$

$$P(y=0|\mathbf{x};\theta) = 1 - P(y=1|\mathbf{x};\theta) \quad (17)$$

Regresión logística

Ejemplo: 2 atributos

$$h_\theta(\mathbf{x}) = h_\theta(x_1, x_2) = \frac{1}{1+exp(-(\theta_0+\theta_1 x_1+\theta_2 x_2))} \equiv g(\theta_0+\theta_1 x_1+\theta_2 x_2)$$

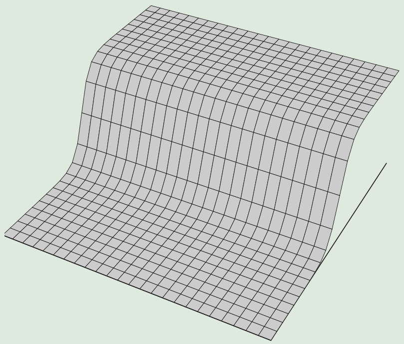


Figura: 2 atributos

Regresión logística

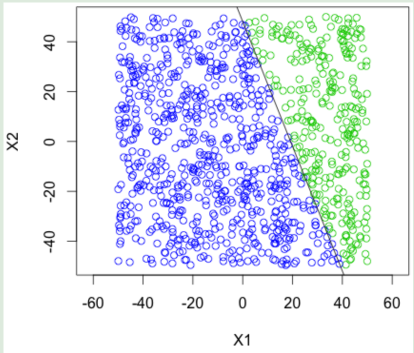
Fronteras de decisión

Predicción:

$$\hat{y} = \begin{cases} 1 & si \quad \Leftrightarrow \theta^T\cdot\mathbf{x} \geq 0 \\ 0 & si \quad \Leftrightarrow \theta^T\cdot\mathbf{x} < 0 \end{cases}$$

Ejemplo: frontera de decisión lineal

Atributos: (x_1, x_2): (lightness, width); Clase $\{0,1\}$: {salmon, sea-bass}



Frontera: la recta $x_2 = 48 - 2{,}5x_1$

Clasificación:

$$\hat{y} = \begin{cases} 1 & si \quad -48 + 2{,}5x_1 + x_2 \geq 0 \\ 0 & si \quad -48 + 2{,}5x_1 + x_2 < 0 \end{cases}$$

Función de regresión logística donde $\theta^T\cdot\mathbf{x} = -48 + 2{,}5x_1 + x_2$

- $\theta_0 = -48$
- $\theta_1 = 2{,}5$
- $\theta_2 = 1$

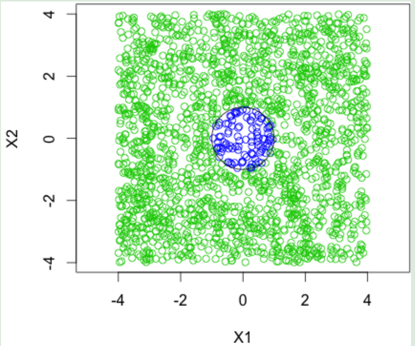
Regresión logística
Fronteras de decisión

Predicción:

$$\hat{y} = \begin{cases} 1 & si \quad \Leftrightarrow z \geq 0 \\ 0 & si \quad \Leftrightarrow z < 0 \end{cases}$$

Ejemplo: frontera de decisión no-lineal

2 atributos (x_1, x_2); clase nominal {verde (1), azul (0)}



Frontera: círculo de radio 1 centrado en el origen. Clasificación:

$$\hat{y} = \begin{cases} 1 & si \quad x_1^2 + x_2^2 \geq 1 \\ 0 & si \quad x_1^2 + x_2^2 < 1 \end{cases}$$

$$\hat{y} = \begin{cases} 1 & si \quad -1 + x_1^2 + x_2^2 \geq 0 \\ 0 & si \quad -1 + x_1^2 + x_2^2 < 0 \end{cases}$$

Función de regresión logística:
$g(z) = g(-1 + x_1^2 + x_2^2)$

Regresión logística
Fronteras de decisión

- Los modelos quedan definidos mediante los parámetros $\theta = (\theta_0, \theta_1, \theta_2, \dots)$.
- Elegir los parámetros que mejor se ajusten a los datos.
  - Criterio: estimación por máxima verosimilitud (MLE).

Estimación paramétrica

Estrategia [Alpaydin, 2010, Chap.4]

- Asume una forma para $p(x|\theta)$
  - Ejemplo: $\mathcal{N}(\mu, \sigma^2)$ donde $\theta = \{\mu, \sigma^2\}$
- Estima θ (aka *sufficient statitstics*) usando $\mathcal{X}$
  - Criterio: *Maximum likelihood*, *Maximum A-Posteriori*, ...

Estimación paramétrica

Repasar!
Teorema de Bayes [Alpaydin, 2010, Sec:3.2] [Duda et al., 2000, Sec:2.1]

$$P(\theta|\mathcal{X}) = \frac{P(\mathcal{X}|\theta)P(\theta)}{P(\mathcal{X})} \tag{18}$$

Terminología:

- $P(\theta)$: probabilidad *a priori* de la hipótesis θ
- $P(\mathcal{X})$: probabilidad *a priori* de los datos $\mathcal{X}$ (*evidencia*)
- $P(\mathcal{X}, \theta)$: probabilidad conjunta de θ y $\mathcal{X}$
- $P(\theta|\mathcal{X})$: probabilidad condicionada de θ dado $\mathcal{X}$ (*a posteriori*)
- $P(\mathcal{X}|\theta)$: probabilidad condicionada de $\mathcal{X}$ dada θ es la verosimilitud (*likelihood*) de θ dada la muestra $\mathcal{X}$
  - Verosimilitud de θ dada la muestra: $\ell(\theta|\mathcal{X}) \equiv p(\mathcal{X}|\theta)$

Estimación paramétrica
Estimación por máxima verosimilitud

Maximum likelihood estimation [Alpaydin, 2010, Sec. 4.2]

Se asume que tenemos una muestra independiente e idénticamente distribuida $\mathcal{X} = \{x^t\}_{t=1}^N$. Las instancias de esta muestra se asocian a una familia de densidad de probabilidad conocida, $p(x|\theta)$, que se define mediante los parámetros θ: $x^t \sim p(x|\theta)$

- Verosimilitud (*likelihood*) de θ dada la muestra:

$$\ell(\theta|\mathcal{X}) \equiv p(\mathcal{X}|\theta) \approx \prod_{t=1}^{N} p(x^t|\theta) \qquad (19)$$

- Log-verosimilitud

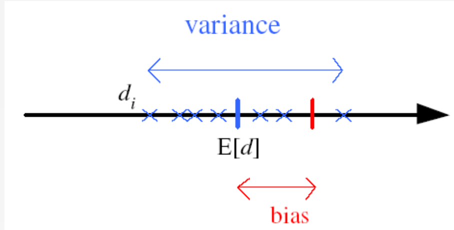
$$\mathcal{L}(\theta|\mathcal{X}) \equiv \log \ell(\theta|\mathcal{X}) = \sum_t \log p(x^t|\theta) \qquad (20)$$

- Estimación de θ por máxima verosimilitud (MLE) (*Maximum Likelihood Estimation*)

$$\hat{\theta} = \arg\max_\theta \mathcal{L}(\theta|\mathcal{X}) \qquad (21)$$

Evaluación del estimador: desviación y varianza

- Sea $\mathcal{X}$ la muestra
- Desconocemos θ
- sea $d = d(\mathcal{X})$ el estimador de θ
- ¿cuál es la calidad del estimador? ¿cuánto difiere el estimador de θ? $[d(\mathcal{X}) - \theta]^2$
  - Desviación: $b_\theta(d) = E[d] - \theta$
  - Varianza: $E[(d - E[d])^2]$



Error cuadrático medio:

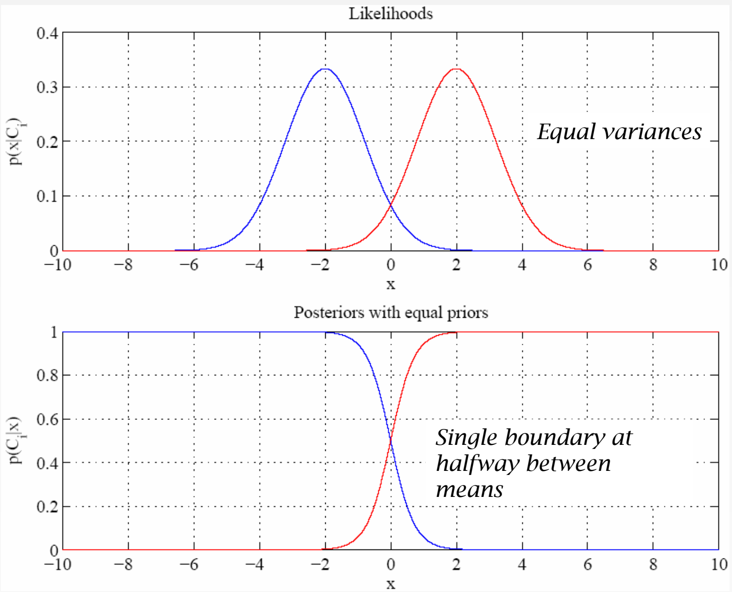
$$\begin{aligned} r(d,\theta) = E[(d-\theta)^2] &= (E[d]-\theta)^2 + E[(d-E[d])^2]] \\ &= \text{Desviación}^2 + \text{Varianza} \end{aligned}$$

Dilema bias-varianza: a medida que incrementamos la complejidad del estimador. . .

. . . disminuye la desviación (se ajusta mejor a los datos)

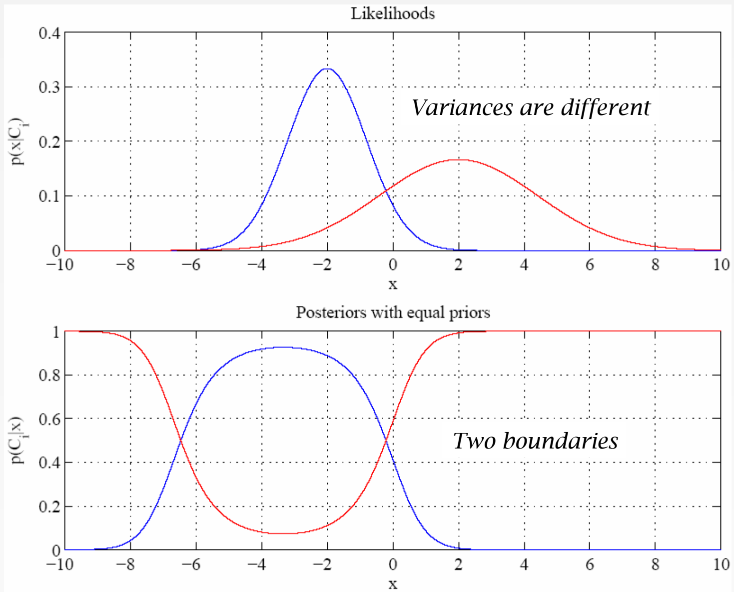
. . . pero aumenta la varianza (el ajuste varía más con los datos)

Evaluación del estimador: desviación y varianza



[Alpaydin, 2010, Sec 4.2.3]

Evaluación del estimador: desviación y varianza



Conclusiones

Modelos paramétricos:

- Se asume que los datos $\mathcal{X}$ se han generado, de forma aleatoria, por una función cuyo tipo es conocido pero que hay cierto error.
  - Ejemplo: $\mathcal{N}(\mu, \sigma^2)$ donde $\theta = \{\mu, \sigma^2\}$
- No se conoce la función, sólo el tipo o familia a la que pertenece (lineal, polinómica, etc.). Es decir, se conoce la función, salvo el valor de una serie de parámetros θ
- Estima θ usando $\mathcal{X}$
  - Criterio: *Maximum likelihood*

Bibliografia I

- Alpaydin, E. (2010). *Introduction to Machine Learning*. MIT Press.
- Duda, R. O., Hart, P. E., and Stork, D. G. (2000). *Pattern Classification*. Wiley-Interscience.
- James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning*, volume 6. Springer.
- Larose, D. T. (2006). *Data mining methods & models*. John Wiley & Sons.
- Orallo, J. H., Ramírez, M. J., and Ferri., C. (2004). *Introducción a la Minería de Datos*. Pearson Educación.
- Witten, I. H., Frank, E., and Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques*. The Morgan Kaufmann Series in Data Management Systems, 3rd edition.

Parte II

Apéndice

Índice

Distribuciones paramétricas elementales

Distribuciones paramétricas:

- Asume una forma para $p(x|\theta)$
  - Bernoulli
  - Multinomial
  - Distribución Gaussiana (normal)
- Estima θ usando $\mathcal{X}$

Distribuciones paramétricas elementales

Bernoulli

Distribución de Bernoulli: dos clases (estados) posibles (e.g. cara/cruz): $x \in \{0, 1\}$

$$P(x) = p^x(1-p)^{(1-x)} \qquad x \in \{0,1\} \tag{22}$$

$$\mathcal{L}(p|\mathcal{X}) = \log \prod_t p^{x^t}(1-p)^{(1-x)} \tag{23}$$

$$\tag{24}$$

Ejercicio: Estimar $\hat{p}$ por máxima verosimilitud:

$$\frac{\partial \mathcal{L}(p|\mathcal{X})}{\partial p} = 0 \Rightarrow \cdots \Rightarrow \hat{p} = \textit{ejercicio} \tag{25}$$

Distribuciones paramétricas elementales

Multinomial

Distribución de Multinomial: e.g. K experimentos de Bernoulli independientes

$$P(x_1, x_2, \cdots, x_K) = \prod_{i=1}^{K} p_i^{x_i} \qquad x \in \{0,1\} \tag{26}$$

$$\mathcal{L}(p_1 p_2, \ldots, p_K|\mathcal{X}) = \log \prod_t \prod_i p_i^{x_i^t} \tag{27}$$

$$\text{MLE: } \hat{p}_i = \sum_t \frac{x_i^t}{N} \tag{28}$$

Distribuciones paramétricas elementales

Gaussiana

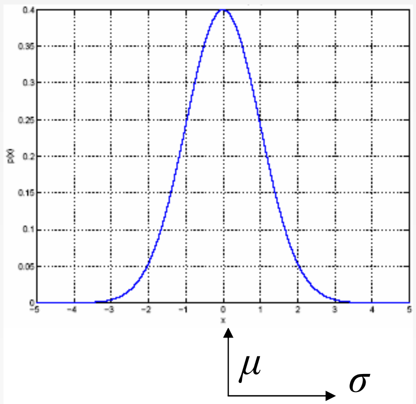


Figura: $N(0,1)$

Distribución Gaussiana:
[Alpaydin, 2010, Sec.4.2.3]

$$p(x) = \textit{ejercicio} \tag{29}$$

$$\mathcal{L}(\mu, \sigma|\mathcal{X}) = \textit{ejercicio} \tag{30}$$

$$\text{MLE:} = \tag{31}$$

$$\hat{\mu} = \textit{ejercicio} \tag{32}$$

$$\hat{\sigma}^2 = \textit{ejercicio} \tag{33}$$