Exploración y análisis de datos

Febrero 2023

GOBIERNO DE ESPAÑA
VICEPRESIDENCIA PRIMERA DEL GOBIERNO
MINISTERIO DE ASUNTOS ECONÓMICOS Y TRANSFORMACIÓN DIGITAL
SECRETARÍA DE ESTADO DE DIGITALIZACIÓN E INTELIGENCIA ARTIFICIAL
red.es
Centro de Referencia Nacional en Comercio Electrónico y Marketing CRN Digital
UNIÓN EUROPEA
Barrabés
The Valley

"El FSE invierte en tu futuro"

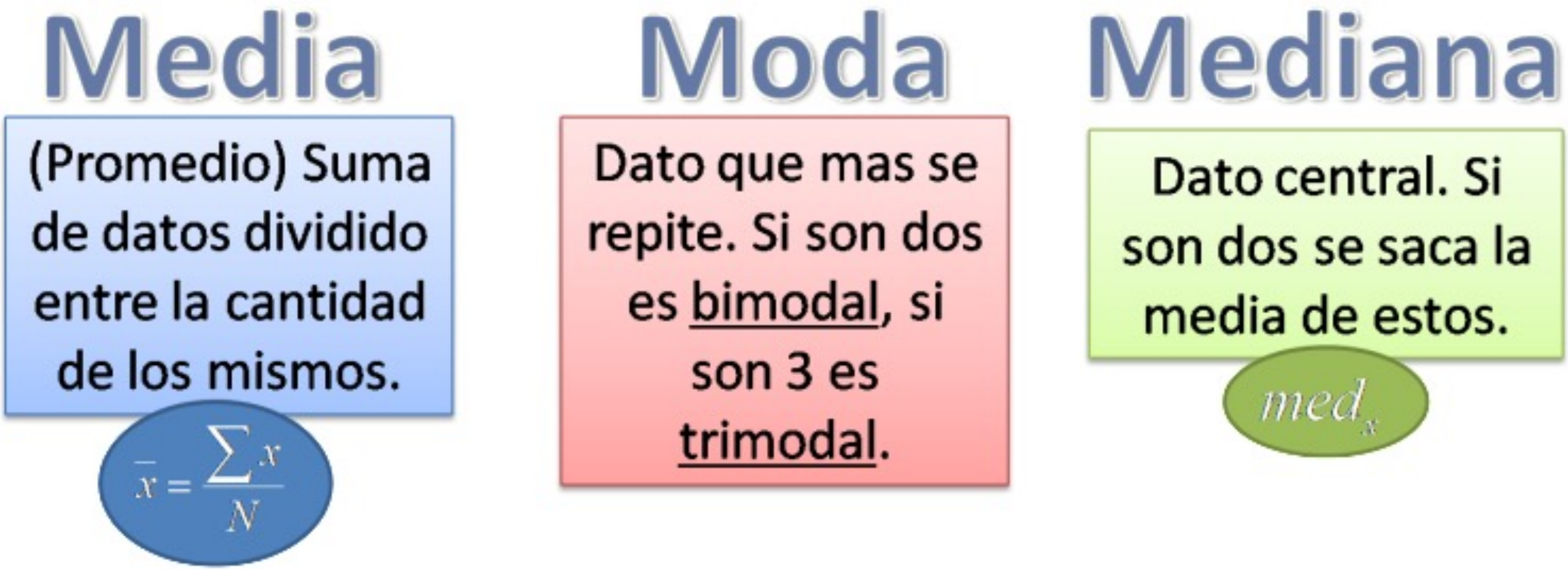
Fondo Social Europeo

Índice

Estadística descriptiva

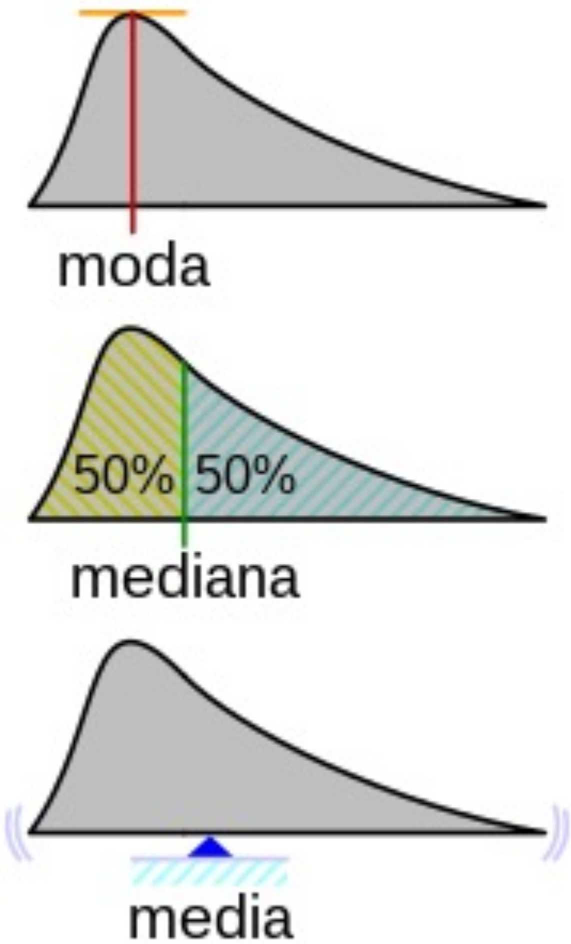
Estadística descriptiva

Moda, media aritmética y mediana



Estadística descriptiva

Moda, media aritmética y mediana



Estadística descriptiva

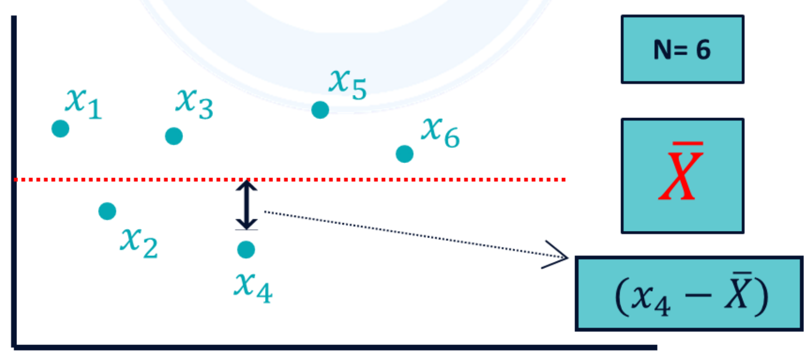
Varianza y desviación estándar

VARIANZA

$$\sigma^2 = \frac{\sum_1^N (x_i - \bar{X})^2}{N}$$

- **X** → Variable
- $\mathbf{x_i}$ → Observación número i de la variable X.
- **N** → Número de observaciones.
- $\bar{\mathbf{X}}$ → Es la media de la variable X.

Es una medida de dispersión que representa la variabilidad de una serie de datos respecto a su media.



Estadística descriptiva

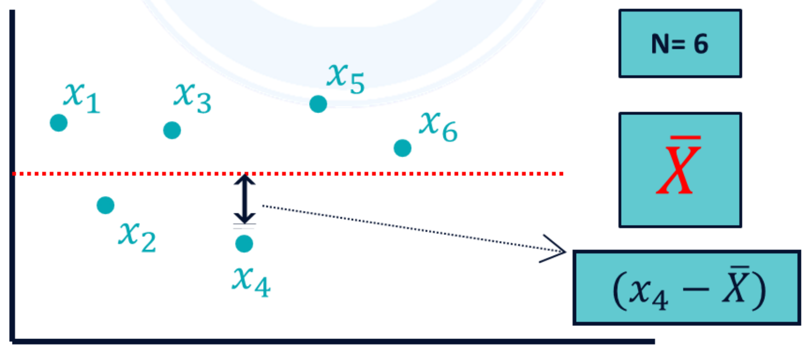
Varianza y desviación estándar

DESVIACIÓN ESTÁNDAR

$$\sigma = \sqrt{\frac{\sum_1^N (x_i - \bar{X})^2}{N}}$$

- **X** → Variable
- $\mathbf{x_i}$ → Observación número i de la variable X.
- **N** → Número de observaciones.
- $\bar{\mathbf{X}}$ → Es la media de la variable X.

También conocida como desviación típica σ es una medida que ofrece información sobre la dispersión media de una variable.



Conceptos básicos de probabilidad

Conceptos básicos de probabilidad

Teoría de la probabilidad

“La probabilidad de que una moneda saque cara es 0.5”.
Interpretaciones:

- **Frecuentista**: Las probabilidades representan frecuencias de eventos a largo plazo. “Si lanzamos la moneda muchas veces, esperamos obtener cara en la mitad de las veces.”
- **Bayesiana**: La probabilidad se utiliza para cuantificar nuestra incertidumbre sobre algo. Más relacionado con información que con ensayos repetidos. “Hay igual probabilidad de obtener cara que cruz en la siguiente tirada.”

Probability theory is nothing but common sense reduced to calculation. — Pierre Laplace, 1812

Conceptos básicos de probabilidad

Interp. Bayesiana

Puede modelar la incertidumbre de eventos que no tienen frecuencias a largo plazo.

- Calcular la probabilidad de que un mensaje que recibir en el correo sea spam.
- Calcular la distribución de probabilidad de la localización de un objetivo en una pantalla radar.

Variables aleatorias discretas

Variables aleatorias discretas

Definiciones básicas

Problema binario:

- $p(A)$: Probabilidad de que el evento A sea verdad.
- $0 \leq p(A) \leq 1$
- $p(\bar{A}) = 1 - p(A)$
- $p()$: Función de masa de probabilidad

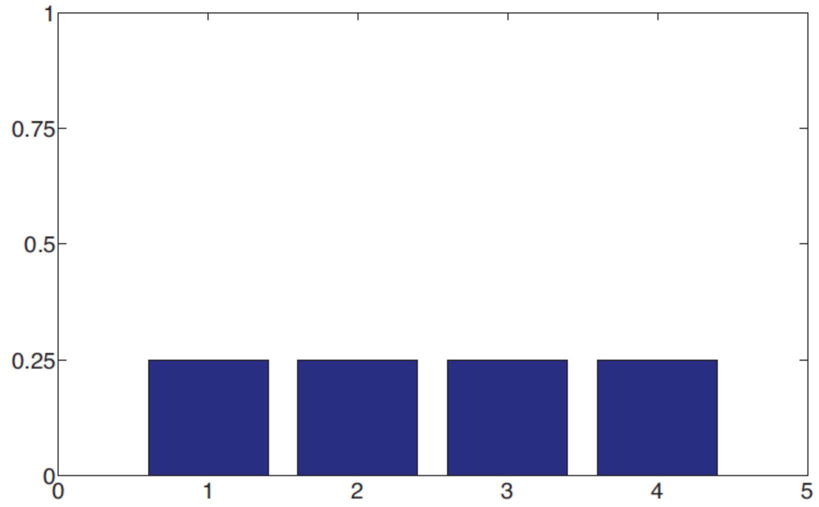
Variable aleatoria discreta: X

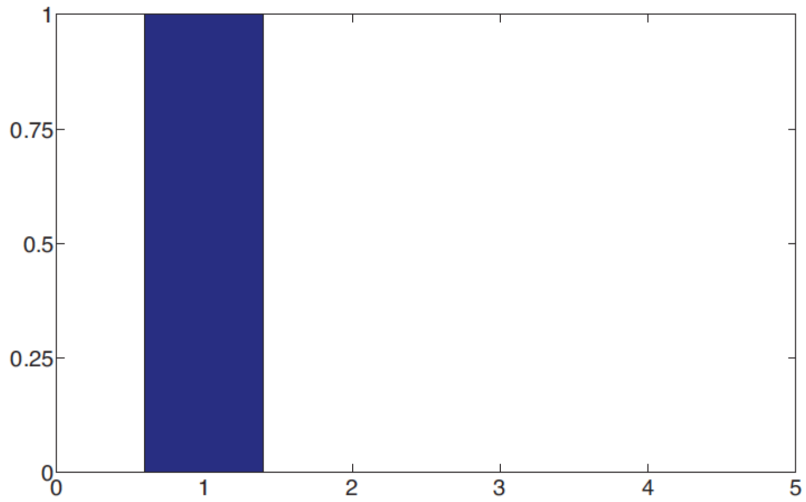
- $p(X = x)$: Probabilidad de que el evento X sea igual x.

$$\sum_{x \in \mathcal{X}} p(x) = 1.$$

Variables aleatorias discretas

Ejemplos





Reglas básicas de probabilidad

Reglas básicas de probabilidad

Distribuciones conjuntas y marginales

$$\begin{aligned} p(A \vee B) &= p(A) + p(B) - p(A \wedge B) \\ &= p(A) + p(B) \end{aligned}$$

Probabilidad conjunta:

$$p(A, B) = p(A \wedge B) = p(A|B)p(B)$$

$$p(X_{1:D}) = p(X_1)p(X_2|X_1)p(X_3|X_2, X_1)p(X_4|X_1, X_2, X_3) \ldots p(X_D|X_{1:D-1})$$

Distribución marginal:

$$p(A) = \sum_b p(A, B) = \sum_b p(A|B = b)p(B = b)$$

Reglas básicas de probabilidad

Probabilidad condicional y regla de Bayes

Probabilidad condicional:

$$p(A|B) = \frac{p(A,B)}{p(B)} \text{ if } p(B) > 0$$

Regla de Bayes:

$$p(X = x|Y = y) = \frac{p(X = x, Y = y)}{p(Y = y)} = \frac{p(X = x)p(Y = y|X = x)}{\sum_{x'} p(X = x')p(Y = y|X = x')}$$

Ejemplo práctico

Te vas a hacer un test (x) para cáncer de pecho (y), específicamente una mamografía.

- Si el test es positivo, cuál es la probabilidad de tener cáncer?

Depende de cómo de fiable sea el test.

Sensibilidad del test: 80%

$$p(x = 1|y = 1) = 0.8$$

- Podrías pensar que tienes el 80% de probabilidad de tener cáncer pero esto no es verdad!

No puedes ignorar

- la probabilidad a priori de tener cáncer de pecho.

$$p(y = 1) = 0.004$$

- La probabilidad de un falso positivo / falsa alarma

$$p(x = 1|y = 0) = 0.1$$

$$\begin{aligned} p(y = 1|x = 1) &= \frac{p(x = 1|y = 1)p(y = 1)}{p(x = 1|y = 1)p(y = 1) + p(x = 1|y = 0)p(y = 0)} \\ &= \frac{0.8 \times 0.004}{0.8 \times 0.004 + 0.1 \times 0.996} = 0.031 \end{aligned}$$

Reglas básicas de probabilidad

Independencia incondicional

Variables aleatorias independientes incondicionalmente

$$X \perp Y \iff p(X,Y) = p(X)p(Y)$$

Theorem 2.2.1. *$X \perp Y|Z$ iff there exist function g and h such that*

$$p(x,y|z) = g(x,z)h(y,z)$$

for all x, y, z such that $p(z) > 0$.

Útil para construir modelos probabilísticos: Clasificador Naive Bayes, modelos de Markov, modelos gráficos,…

Variables aleatorias continuas

Variables aleatorias continuas

Funciones de distribución y de densidad de probabilidad

$F(q) = p(X \leq q)$: Función de distribución acumulada de X (cdf)

- Monótona creciente
- $p(a < X \leq b) = F(b) - F(a)$

$\mathrm{f}(x) = \frac{d}{dx} F(x)$: Función de densidad de probabilidad (pdf)

- Probabilidad de que una variable continua esté comprendida en un intervalo finito:

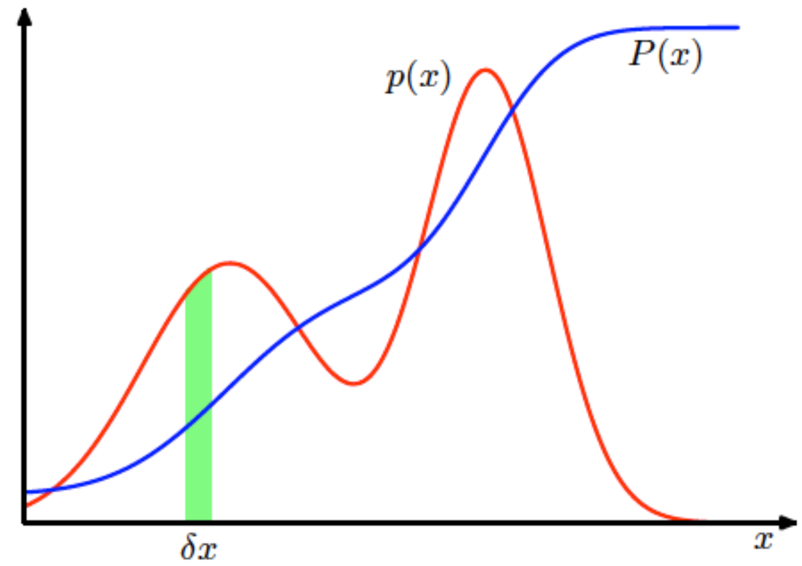
$$P(a < X \leq b) = \int_a^b f(x)dx$$

Variables aleatorias continuas

Propiedades de la función de densidad de probabilidad

$$p(x) \geqslant 0$$

$$\int_{-\infty}^{\infty} p(x)\,\mathrm{d}x = 1$$



Variables aleatorias continuas

Función cuantil, media y varianza

- $F^{-1}(\alpha)$: Función de distribución acumulada inversa. **Función cuantil**.

Es el valor de x_α tal que $\mathrm{P}(X \leq x_\alpha) = \alpha$

- **Mediana**: $F^{-1}(0.5)$, con la mitad de la masa de probabilidad a la izquierda y la otra mitad a la derecha.
- **Media**:

$$\mathbb{E}[X] \triangleq \int_{\mathcal{X}} x \; p(x)dx$$

- **Varianza**:

$$\begin{aligned} \operatorname{var}[X] &\triangleq \mathbb{E}\left[(X-\mu)^2\right] = \int (x-\mu)^2 p(x)dx \\ &= \int x^2 p(x)dx + \mu^2 \int p(x)dx - 2\mu \int xp(x)dx = \mathbb{E}\left[X^2\right] - \mu^2 \end{aligned}$$

- **Desviación estándar**:

$$\operatorname{std}[X] \triangleq \sqrt{\operatorname{var}[X]}$$

$$\mathbb{E}\left[X^2\right] = \mu^2 + \sigma^2$$

Operador esperanza

Operador esperanza

Variables discretas y continuas

$$\mathbb{E}[f] = \int p(x) f(x) \, \mathrm{d}x.$$

$$\mathbb{E}[f] = \sum_x p(x) f(x)$$

Operador esperanza

Varianza y covarianza

- **Varianza**: Proporciona una medida de cuanta variabilidad hay en x alrededor de su valor medio.

$$\mathrm{var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2$$

- **Covarianza**: Extiende el concepto a dos variables aleatorias y expresa hasta qué punto x e y varían juntas. Si son independientes, la covarianza es nula.

$$\begin{aligned}\mathrm{cov}[x, y] &= \mathbb{E}_{x,y}\left[\{x - \mathbb{E}[x]\}\{y - \mathbb{E}[y]\}\right] \\ &= \mathbb{E}_{x,y}[xy] - \mathbb{E}[x]\mathbb{E}[y]\end{aligned}$$

Operador esperanza

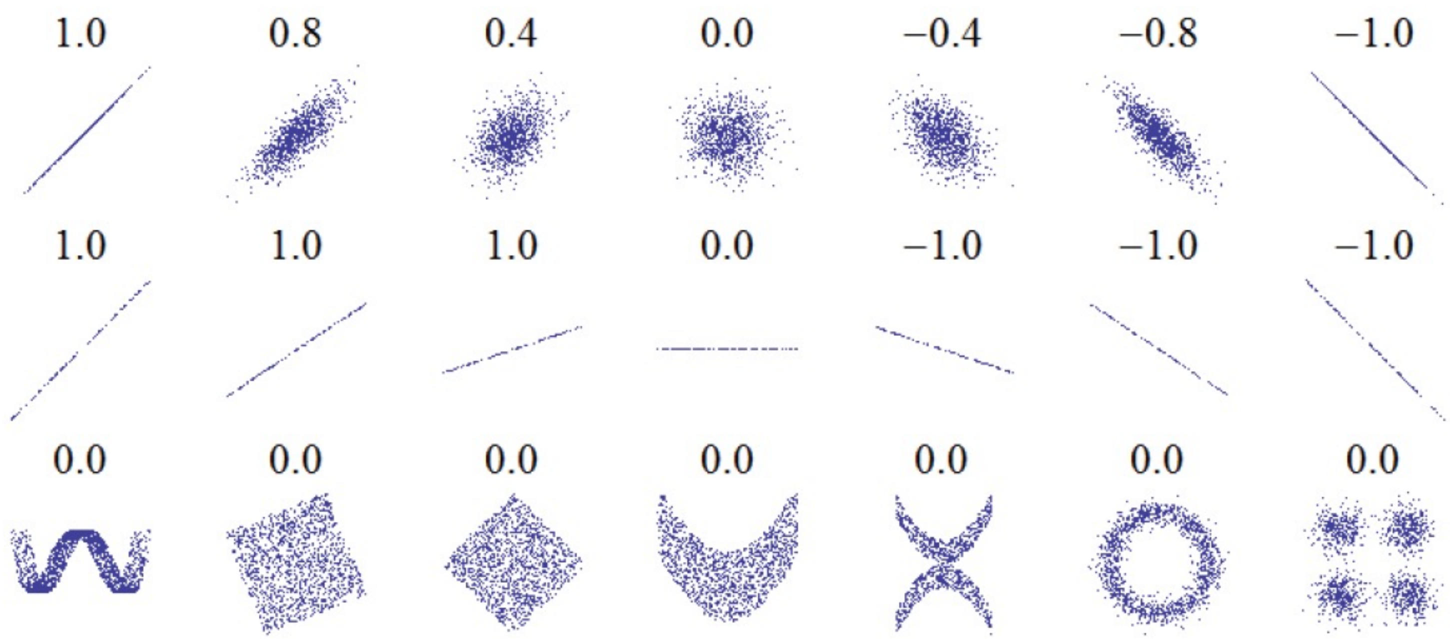
Coeficiente de correlación (Pearson) y matriz de correlación

$$\operatorname{corr}[X,Y] \triangleq \frac{\operatorname{cov}[X,Y]}{\sqrt{\operatorname{var}[X]\operatorname{var}[Y]}}$$

$$\mathbf{R} = \begin{pmatrix} \operatorname{corr}[X_1,X_1] & \operatorname{corr}[X_1,X_2] & \cdots & \operatorname{corr}[X_1,X_d] \\ \vdots & \vdots & \ddots & \vdots \\ \operatorname{corr}[X_d,X_1] & \operatorname{corr}[X_d,X_2] & \cdots & \operatorname{corr}[X_d,X_d] \end{pmatrix}$$

Operador esperanza

Coeficiente de correlación (Pearson) y matriz de correlación



El coeficiente de correlación refleja el ruido y la dirección de la relación lineal, pero no la pendiente de esa relación ni aspectos no lineales.
Incorrelación no implica independencia.

Operador esperanza

Coeficiente de correlación (Pearson) y matriz de correlación

	date	weight	resp_1	resp_2	resp_3	resp_4	resp	feature_0	feature_1	feature_2
date	1.00	0.02	0.02	0.01	0.01	-0.02	-0.03	0.01	-0.05	-0.01
weight	0.02	1.00	-0.01	-0.01	-0.02	-0.02	-0.02	-0.02	-0.05	0.06
resp_1	0.02	-0.01	1.00	0.92	0.80	0.49	0.58	0.04	0.03	0.02
resp_2	0.01	-0.01	0.92	1.00	0.91	0.55	0.68	0.04	0.04	0.01
resp_3	0.01	-0.02	0.80	0.91	1.00	0.77	0.80	0.03	0.04	0.01
resp_4	-0.02	-0.02	0.49	0.55	0.77	1.00	0.95	-0.06	0.00	-0.04
resp	-0.03	-0.02	0.58	0.68	0.80	0.95	1.00	-0.06	0.00	-0.04
feature_0	0.01	-0.02	0.04	0.04	0.03	-0.06	-0.06	1.00	0.03	0.03
feature_1	-0.05	-0.05	0.03	0.04	0.04	0.00	0.00	0.03	1.00	0.84
feature_2	-0.01	0.06	0.02	0.01	0.01	-0.04	-0.04	0.03	0.84	1.00

Aplicaciones de la probabilidad al enfoque Bayesiano

Aplicaciones de la probabilidad al enfoque Bayesiano

Deshielo del Ártico

- No es un evento que se repita (frecuentista).
- Tenemos una idea de cómo de rápido el hielo polar se está derritiendo.
- Si recibimos información de un satélite sobre nuevos diagnósticos, podríamos revisar nuestra opinión sobre cómo de rápido se produce el deshielo.
- Nuestra evaluación del problema provocará nuevas acciones sobre por ejemplo la emisión de gases.

Queremos cuantificar la incertidumbre y hacer revisiones precisas cuando tenemos nuevas evidencias.

Aplicaciones de la probabilidad al enfoque Bayesiano

Enfoque Bayesiano en Machine Learning

Usar la teoría de la probabilidad para describir la incertidumbre en los parámetros del modelo, w, o en la elección del modelo en sí.

- Distribución de **probabilidad a priori**: p(w)

Captura nuestras suposiciones sobre w antes de observar los datos.

- El efecto de observar los datos se expresa a través de la probabilidad condicional, p(D|w). Es una función de los parámetros y se llama ***likelihood*** *function*. Expresa cómo de probable es el conjunto de datos para diferentes configuraciones de los parámetros.
- Así podemos evaluar la incertidumbre en los parámetros, w, después de observar lso datos en forma de probabilidad a posteriori, p(w|D).

$$p(\mathbf{w}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathbf{w})p(\mathbf{w})}{p(\mathcal{D})}$$

$$p(\mathcal{D}) = \int p(\mathcal{D}|\mathbf{w})p(\mathbf{w})\,\mathrm{d}\mathbf{w}$$

Aplicaciones de la probabilidad al enfoque Bayesiano

Enfoque Bayesiano en Machine Learning

Usar la teoría de la probabilidad para describir la incertidumbre en los parámetros del modelo, w, o en la elección del modelo en sí.

$$\text{posterior} \propto \text{likelihood} \times \text{prior}$$

Aplicaciones de la probabilidad al enfoque Bayesiano

Enfoque Bayesiano en Machine Learning

Interpretación frecuentista: **w** se considera un vector de parámetros fijos cuyo valor se determina a través de un estimador y los errores se obtienen considerando la distribución **de posibles conjuntos de datos** D.

Interpretación Bayesiana: **Sólo hay un conjunto de datos** D (el observado) y la incertidumbre en los parámetros se expresa a través de una **distribución de probabilidad sobre w**.

Aplicaciones de la probabilidad al enfoque Bayesiano

Enfoque Bayesiano en Machine Learning

Elegimos los valores de los parámetros que optimizan un criterio.

Estimador "Maximum likelihood": se seleccionan los parámetros **w** que maximizan la función likelihood, p(D|**w**)

Esto corresponde a elegir los valores de w para los cuales la probabilidad de los datos observados es maximizada.

Función error: ***Negative log likelihood***, es monótona decreciente. Maximizar la *likelihood* es equivalente a minimizar el error.

Algunas distribuciones discretas comunes

Algunas distribuciones discretas

Bernoulli

Variables aleatorias binarias.
Ejemplo: lanzar una moneda y obtener cara

$$p(x = 1|\mu) = \mu \qquad\qquad p(x = 0|\mu) = 1 - \mu$$

$$\mathrm{Bern}(x|\mu) = \mu^x (1 - \mu)^{1-x}$$

$$\begin{aligned} \mathbb{E}[x] &= \mu \\ \mathrm{var}[x] &= \mu(1 - \mu) \end{aligned}$$

Algunas distribuciones discretas

Bernoulli

Aplicación en ML, asumiendo que las observaciones son independientes e idénticamente distribuidas.

$$p(\mathcal{D}|\mu) = \prod_{n=1}^{N} p(x_n|\mu) = \prod_{n=1}^{N} \mu^{x_n}(1-\mu)^{1-x_n}$$

Algunas distribuciones discretas

Binomial

Cuando el tamaño del conjunto de datos es N.
Ejemplo: Lanzar N monedas y obtener m caras.

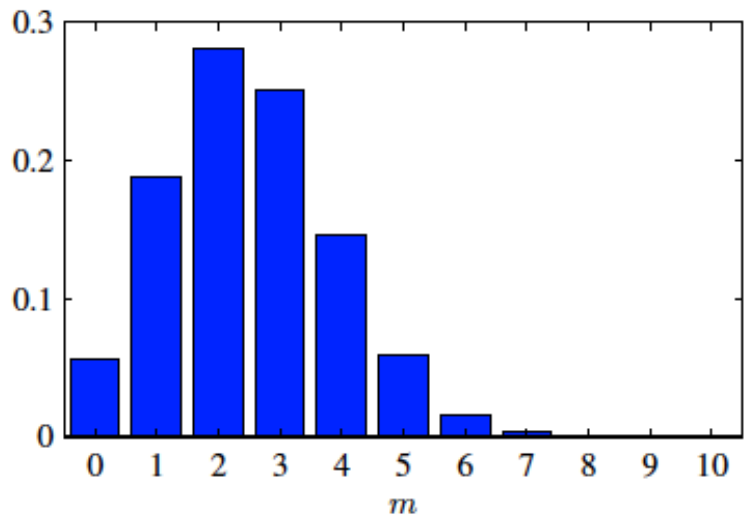
$$\mathrm{Bin}(m|N,\mu) = \binom{N}{m}\mu^m(1-\mu)^{N-m}$$

$$\binom{N}{m} \equiv \frac{N!}{(N-m)!m!}$$

Algunas distribuciones discretas

Binomial

Ejemplo: N=10 y μ=0.25



Algunas distribuciones discretas

Multinomial

Cuando el tamaño del conjunto de datos es N y la variable discreta puede tomar uno de K estados posibles mutuamente exclusivos.

$$\mu_k = P(x_k = 1)$$

Ejemplo: lanzar N veces un dado

One-hot encoding

$$\mathbf{x} = (0, 0, 1, 0, 0, 0)^{\mathrm{T}}$$

$$\textstyle\sum_{k=1}^{K} x_k = 1$$

$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_k^{x_k}$$

$$\textstyle\sum_k \mu_k = 1$$

$$\mu_k \geqslant 0$$

Algunas distribuciones discretas

Multinomial

Otra notación

$$\mathrm{Mu}(\mathbf{x}|n,\boldsymbol{\theta}) \triangleq \binom{n}{x_1 \ldots x_K} \prod_{j=1}^{K} \theta_j^{x_j}$$

vs

$$\mathrm{Mu}(\mathbf{x}|1,\boldsymbol{\theta}) = \prod_{j=1}^{K} \theta_j^{\mathbb{I}(x_j=1)}$$

$$\binom{n}{x_1 \ldots x_K} \triangleq \frac{n!}{x_1!x_2!\cdots x_K!}$$

x_j es las veces que aparece la cara j del dado

Algunas distribuciones discretas

Poisson

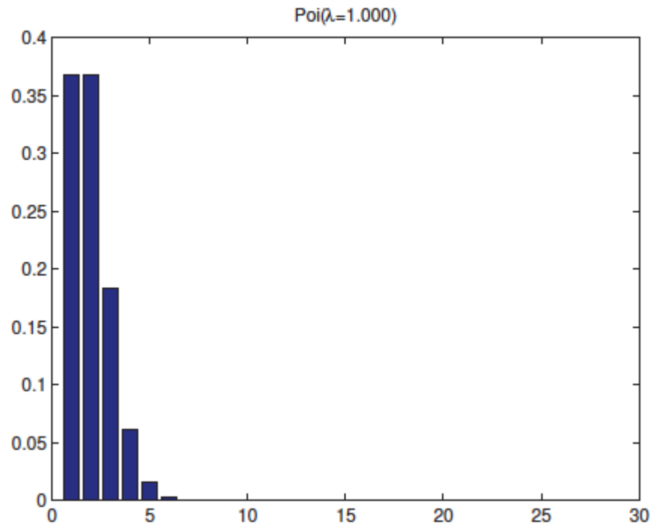
Para modelar cuentas de eventos raros como accidentes de tráfico.

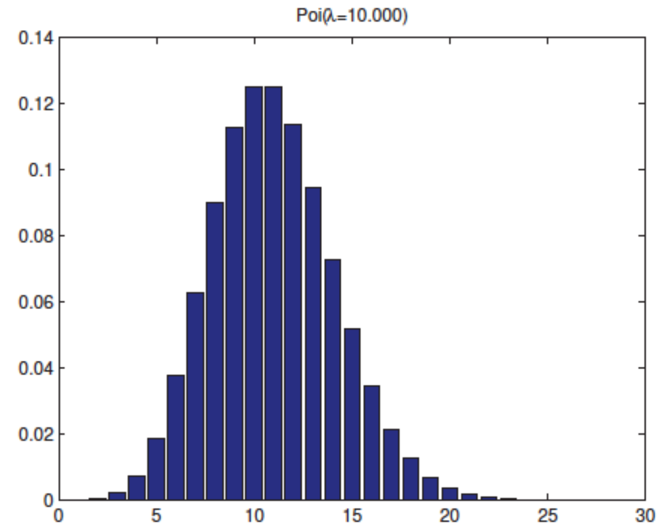
$X \in \{0, 1, 2, \ldots\}$

$$\mathrm{Poi}(x|\lambda) = e^{-\lambda} \frac{\lambda^x}{x!} \qquad \lambda > 0$$

Algunas distribuciones discretas

Poisson





Algunas distribuciones continuas comunes

Algunas distribuciones continuas

Gaussian

De una variable única o multidimensional
Parámetros: media y varianza/covarianza

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

Algunas distribuciones continuas

Gaussian

Teorema del límite central
Sujeto a ciertas condiciones, la suma de un conjunto de variables aleatorias, que también es en sí una variable aleatoria, tiene una distribución que se convierte cada vez más en Gaussiana cuando el número de términos en la suma aumenta.

Algunas distribuciones continuas

T-student

$$\mathcal{T}(x|\mu,\sigma^2,\nu) \quad \propto \quad \left[1+\frac{1}{\nu}\left(\frac{x-\mu}{\sigma}\right)^2\right]^{-\left(\frac{\nu+1}{2}\right)}$$

$$\text{mean} = \mu, \text{mode} = \mu, \ \text{var} = \frac{\nu\sigma^2}{(\nu-2)}$$

Algunas distribuciones continuas

Laplace

Parámetros: localización, μ y escala, b.

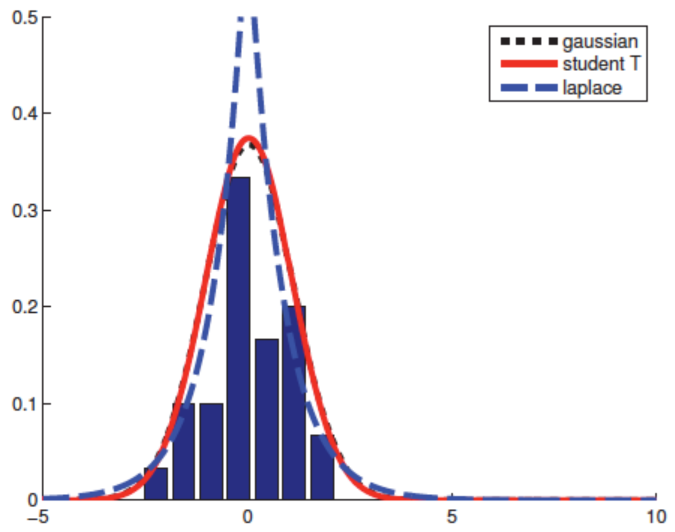
$$\mathrm{Lap}(x|\mu, b) \triangleq \frac{1}{2b} \exp\left(-\frac{|x-\mu|}{b}\right)$$

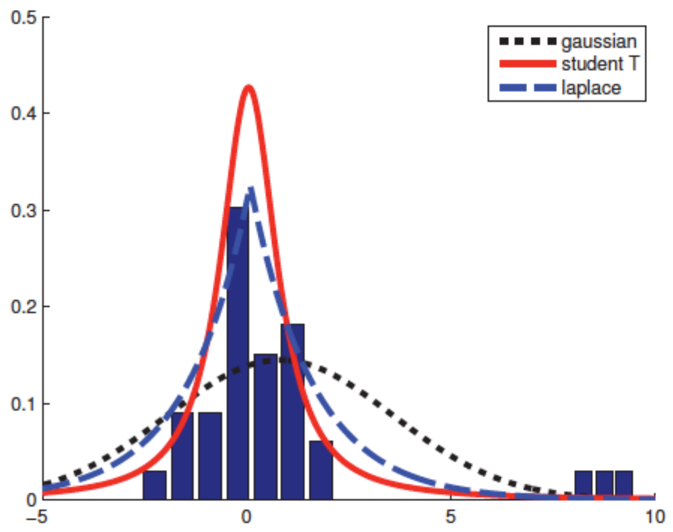
$$\text{mean} = \mu, \ \text{mode} = \mu, \ \text{var} = 2b^2$$

Algunas distribuciones continuas

Gaussiana, T-student y Laplace

La Gaussiana es más sentible a *outliers*.





Algunas distribuciones continuas

Gamma

Para datos reales positivos
Parámetros: forma, α, y escala, λ. Mayores que 0.

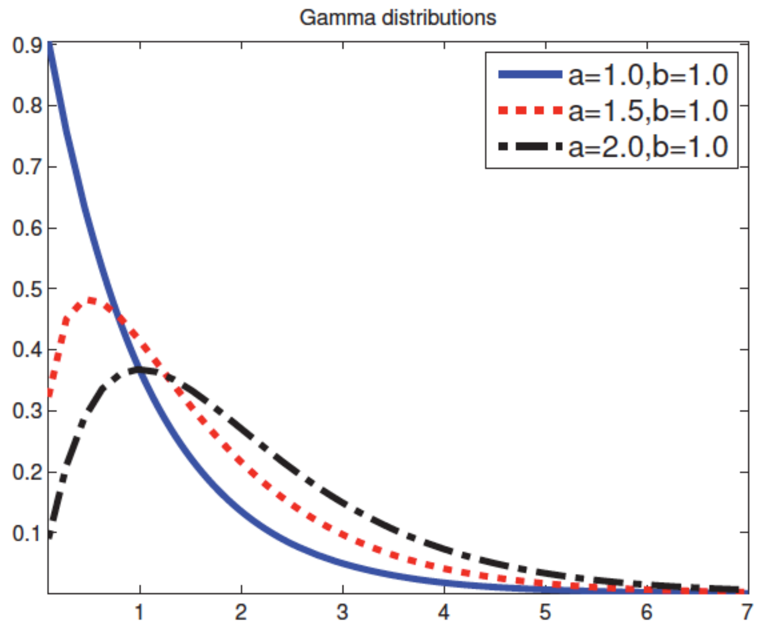
$$\mathrm{Ga}(T|\mathrm{shape}=a,\mathrm{rate}=b) \triangleq \frac{b^a}{\Gamma(a)}T^{a-1}e^{-Tb}$$

$$\Gamma(x) \triangleq \int_0^\infty u^{x-1}e^{-u}du$$

$$\mathrm{mean}=\frac{a}{b},\ \mathrm{mode}=\frac{a-1}{b},\ \mathrm{var}=\frac{a}{b^2}$$

Algunas distribuciones continuas

Gamma



Algunas distribuciones continuas

Beta

Parámetros: a y b. Mayores que 0.

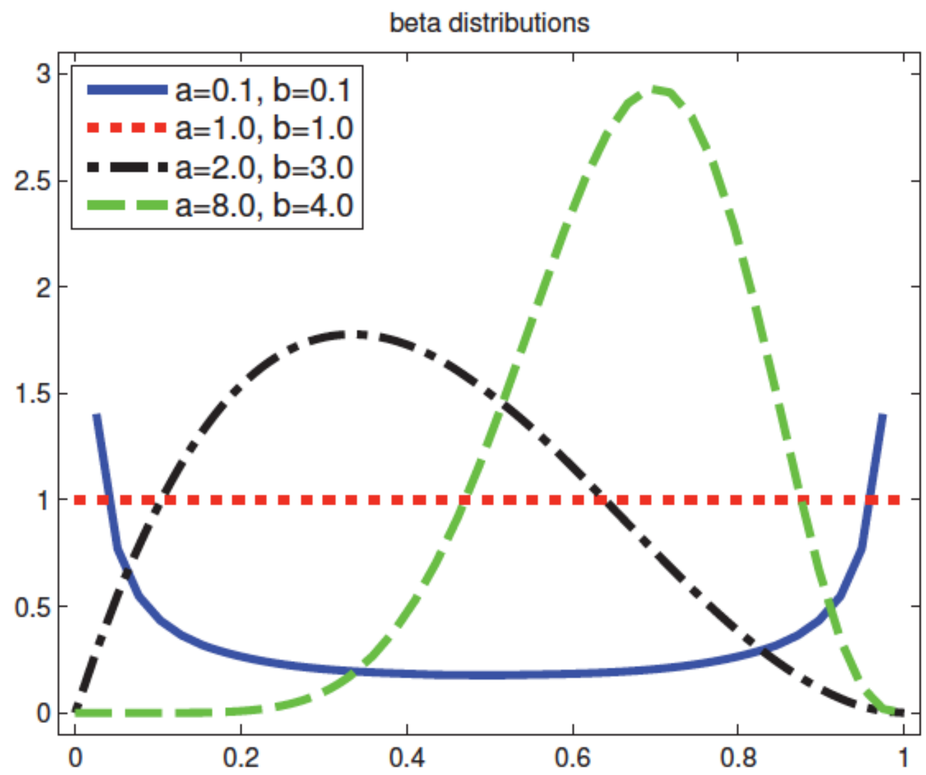
$$\mathrm{Beta}(x|a,b) = \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1}$$

$$B(a,b) \triangleq \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

$$\text{mean} = \frac{a}{a+b},\ \text{mode} = \frac{a-1}{a+b-2},\ \text{var} = \frac{ab}{(a+b)^2(a+b+1)}$$

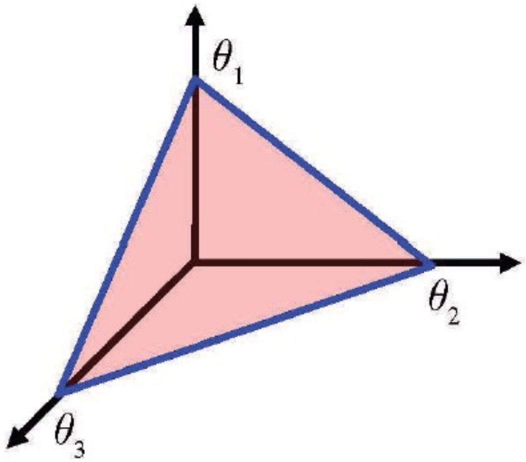
Algunas distribuciones continuas

Beta



Algunas distribuciones continuas

Dirichlet



Generalización multivariable de la distribución Beta.
Simplex de probabilidad.

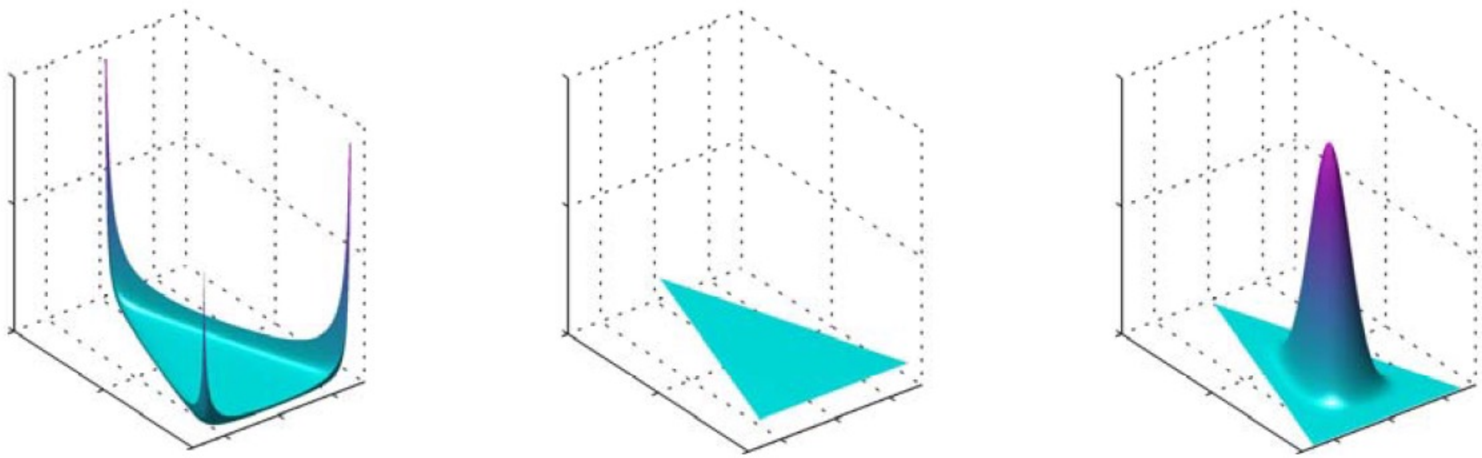
$$S_K = \{\mathbf{x} : 0 \leq x_k \leq 1, \sum_{k=1}^{K} x_k = 1\}$$

$$\mathrm{Dir}(\mathbf{x}|\boldsymbol{\alpha}) \triangleq \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^{K} x_k^{\alpha_k - 1} \mathbb{I}(\mathbf{x} \in S_K)$$

Algunas distribuciones continuas

Dirichlet

$$\alpha_k = 0.1, 1\ y\ 10$$



Supuesto básico en ML

los datos $=$ señal $+$ ruido (Gaussiana)
es decir, $y = f(x) + \epsilon$ donde ϵ es el 'error irreducible'

Aunque el ruido (ϵ) es aleatorio, si se tiene un número grande de muestras finalmente se aproxima a una distribución Gaussiana.

En matemáticas eso se llama el "teorema del límite central"

Práctica

Practiquemos sobre estadística y conceptos básicos de probabilidad.

Notebook:
5. Estadística descriptiva y probabilidad

Duración estimada: 1h

Contacto

Correo: a.cobo.aguilera@gmail.com

LinkedIn: Aurora Cobo Aguilera
GitHub: AuroraCoboAguilera
Google Scholar: Aurora Cobo Aguilera

GOBIERNO DE ESPAÑA

VICEPRESIDENCIA PRIMERA DEL GOBIERNO

MINISTERIO DE ASUNTOS ECONÓMICOS Y TRANSFORMACIÓN DIGITAL

SECRETARÍA DE ESTADO DE DIGITALIZACIÓN E INTELIGENCIA ARTIFICIAL

red.es

Centro de Referencia Nacional en Comercio Electrónico y Marketing

CRN Digital

UNIÓN EUROPEA

"El FSE invierte en tu futuro"

Fondo Social Europeo

Barrabés

The Valley