# Ciencias de Datos 2025 - Trabajo Práctico 2

- **1.** Sea $\vec{x}$ un vector binario $d$-dimensional, es decir, tal que todas sus componentes toman los valores cero o uno, de acuerdo con una distribución de Bernoulli independiente con parámetro $\theta_i$ para la componente $i$-ésima (es decir, la probabilidad de que $x_i = 1$ es $\theta_i$).

  La función de verosimilitud de un vector $\vec{x}$ es:

$$P(\vec{x}|\vec{\theta}) = \prod_{i=1}^{d} \theta_i^{x_i}(1-\theta_i)^{1-x_i}$$

  Suponga que se tiene una muestra independiente de $n$ vectores binarios $\{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n\}$. Demuestre que el estimador de máxima verosimilitud para el vector de parámetros $\vec{\theta}$ es:

$$\hat{\vec{\theta}} = \frac{1}{n}\sum_{k=1}^{n} \vec{x}_k$$

- **2.** Realice una simulación de una muestra aleatoria $S = \{\vec{x}_1, \ldots, \vec{x}_N\}$ de una variable aleatoria $\vec{X}$ de dimensión $d = 7$ y tamaño $N = 1000$, donde cada componente $x_{k,i}$ sigue una distribución de Bernoulli con parámetros $\vec{\theta} = (\theta_1, \ldots, \theta_7)$. Estos parámetros deben generarse una única vez de forma aleatoria con valores tomados de una distribución uniforme en el intervalo $(0, 1)$.

  Calcule el vector suma de la muestra:

$$\vec{s} = \sum_{k=1}^{N} \vec{x}_k = (s_1, s_2, \ldots, s_7)$$

  y muestre que la verosimilitud de toda la muestra puede escribirse como:

$$P(S|\vec{\theta}) = \prod_{i=1}^{7} \theta_i^{s_i}(1-\theta_i)^{N-s_i}$$

- **3.** Suponga ahora que los parámetros $\theta_i$ siguen una distribución a priori uniforme para cada componente. Entonces, la distribución a posteriori de $\vec{\theta}$ dada la muestra $S$ es:

$$P(\vec{\theta}|S) = \prod_{i=1}^{d} \frac{(n+1)!}{s_i!(n-s_i)!}\theta_i^{s_i}(1-\theta_i)^{N-s_i}.$$

  Utilizando que

$$\int_0^1 \theta^m(1-\theta)^n d\theta = \frac{m!n!}{(m+n+1)!}$$

y que $x_i$ solo puede valer $0$ o $1$, demuestre la siguiente integral:

$$P(\vec{x}|S) = \int P(\vec{x}|\vec{\theta})\, P(\vec{\theta}|S)\, d\vec{\theta} = \prod_{i=1}^{d} \left(\frac{s_i + 1}{n + 2}\right)^{x_i} \left(1 - \frac{s_i + 1}{n + 2}\right)^{1-x_i}$$

Interprete el resultado y discuta su utilidad como clasificador bayesiano.

► **4.** Suponga que este modelo se utiliza para construir un clasificador de spam (es decir, un sistema que clasifica correos electrónicos como "spam" o "no spam"). Cada componente binaria $x_i$ indica la presencia ($x_i = 1$) o ausencia ($x_i = 0$) de una palabra clave específica en el cuerpo del correo.

Proponga un conjunto de 20 palabras que puedan actuar como predictores de spam y asígneles valores de $\theta_i$ que podrían ser razonables bajo la hipótesis de que el correo es spam.

Luego, escriba el texto de un correo y construya el vector binario correspondiente. Use los parámetros propuestos para determinar si el correo es spam.

Ejemplos de la talba a generar: $\theta_{ganaste} = 0{,}9$, $\theta_{promocion} = 0{,}8$.

Escriba una función que dado un texto, construya el vector binario y calcule la verosimilitud de que ese mensaje provenga de un correo no deseado (spam).