Ciencia de Datos

Práctico N°3: Estimación de Parámetros

Problema 1: La variable aleatoria X tiene distribución exponencial,

$$p(x|\theta) = \begin{cases} \theta\, e^{-\theta x} & \text{si } x \geq 0 \\ 0 & \text{en caso contrario} \end{cases}$$

a) Graficar $p(x|\theta)$ versus x para $\theta = 1$. Graficar $p(x|\theta)$ versus θ, $(0 \leq \theta \leq 5)$, para $x = 2$.
b) Suponiendo que n ejemplos $x_1, \cdots, x_n$ se generan de forma independiente de acuerdo a $p(x|\theta)$, mostrar que el estimador de máxima verosimilitud para θ viene dado por

$$\hat{\theta} = \left(\frac{1}{n} \sum_{k=1}^{n} x_k \right)^{-1} .$$

c) En el gráfico generado en el ítem (a) para $\theta = 1$, trazar una vertical en el valor del estimador de máxima verosimilitud correspondiente a un valor de n grande.

Problema 2: Suponer que la variable aleatoria X tiene distribución uniforme con parámetro θ,

$$p(x|\theta) \sim U(0, \theta) = \begin{cases} \dfrac{1}{\theta}, & \text{si } 0 \leq x \leq \theta, \\ 0, & \text{en caso contrario.} \end{cases}$$

a) Dados n ejemplos $\mathcal{D} = \{x_1, \ldots, x_n\}$ generados de manera independiente de acuerdo a $p(x|\theta)$, mostrar que el estimador de máxima verosimilitud para θ es máx$(\mathcal{D})$; esto es, el valor del máximo elemento de $\mathcal{D}$.
b) Se generan $n = 5$ datos con esta distribución y el máximo valor de esos puntos resulta máx$_k\{x_k\} = 0{,}6$. Graficar la verosimilitud $p(\mathcal{D}|\theta)$ en el rango $0 \leq \theta \leq 1$. Argumentar con palabras por que no es necesario conocer los otros 4 datos de la muestra.

Problema 3: Se prueba la salida de un sistema y las respuestas posibles son fallo o éxito. La variable aleatoria Y que cuenta el número k de fallos hasta obtener un éxito en la prueba $k + 1$ tiene distribución de probabilidad geométrica con parámetro p, la probabilidad de éxito en una prueba independiente,

$$P(Y = k|p) = (1 - p)^k\, p \text{ con } k = 0, 1, 2, \ldots .$$

Se realizan n series de pruebas independientes y se encuentra en cada una de ellas los valores $k_1, k_2, \ldots, k_n$. Calcular el estimador de máxima verosimilitud para el parámetro p a partir de los valores obtenidos.

Problema 4: La tabla adjunta reporta los datos empíricos tridimensionales correspondientes a tres clases independientes.
a) Calcular los valores de máxima verosimilitud $\hat{\mu}$ y $\hat{\sigma}^2$ de forma individual para cada una de las tres característica x_i de la categoria w_1 (problema unidimensional).
b) Calcular los valores de máxima verosimilitud $\hat{\mu}$ y $\hat{\Sigma}$ para cada una de las tres formas de apareamiento de a dos características para w_1 (problema bidimensional).

clase	w_1			w_2			w_3		
	x_1	x_2	x_3	x_1	x_2	x_3	x_1	x_2	x_3
1	0.42	-0.087	0.58	-0.4	0.58	0.089	0.83	1.6	-0.014
2	-0.2	-3.3	-3.4	-0.31	0.27	-0.04	1.1	1.6	0.48
3	1.3	-0.32	1.7	0.38	0.055	-0.035	-0.44	-0.41	0.32
4	0.39	0.71	0.23	-0.15	0.53	0.011	0.047	-0.45	1.4
5	-1.6	-5.3	-0.15	-0.35	0.47	0.034	0.28	0.35	3.1
6	-0.029	0.89	-4.7	0.17	0.69	0.1	-0.39	-0.48	0.11
7	-0.23	1.9	2.2	-0.011	0.55	-0.18	0.34	-0.079	0.14
8	0.27	-0.3	-0.87	-0.27	0.61	0.12	-0.3	-0.22	2.2
9	-1.9	0.76	-2.1	-0.065	0.49	0.0012	1.1	1.2	-0.46
10	0.87	-1.0	-2.6	-0.12	0.054	-0.063	0.18	-0.11	-0.49

c) Calcular los valores de máxima verosimilitud $\hat{\mu}$ y $\hat{\Sigma}$ usando las tres características x_i de la categoria w_1 (problema tridimensional).
d) Si se supone que las características son independientes entre sí el modelo gaussino es separable y la matriz Σ resulta diagonal, $\Sigma = \text{diag}(\sigma_1^2, \sigma_2^2, \sigma_3^2)$. Estimar por máxima verosimilitud la media y las componentes diagonales de Σ con los datos de las clases w_1 y w_2.
e) Comparar los resultados para la media de cada característica μ_i calculada en las formas previas. Explicar porqué son iguales o diferentes.
f) Comparar sus resultados para la varianza de cada característica σ_i^2 calculada de las formas previas. Explicar por que los resultados son iguales o diferentes.

Problema 5: Usando los datos de la tabla del problema anterior, construir modelos gaussianos de clasificación para cada uno de los ítems a continuación y calcular las tasas de error de clasificación en diferentes dimensiones.
a) Usar máxima verosimilitud para entrenar un dicotomizador gaussiano con los datos tridimensionales de las categorias w_1 y w_2. Integrar numéricamente para estimar la proporción del error.
b) Proyectar los datos sobre un subespacio bidimensional. Para cada uno de los tres subespacios definidos por $x_1 = 0$ ó $x_2 = 0$ ó $x_3 = 0$ entrenar un dicotomizador gaussiano. Integre numéricamente para estimar la proporción del error.
c) Proyectar ahora en subespacios unidimensionales, definidos por cada uno de los tres ejes. Entrenar un clasificador gaussiano e integre numéricamente para estimar la proporción del error.
d) Discutir el orden del rango de las tasas de error calculadas.
e) Suponiendo que se reestima la distribución en las diferentes dimensiones, el error de Bayes es mayor en los espacios proyectados?