Ejercicios clasificación

1 Clasificador óptimo

1. Consideramos tres poblaciones de mismas proporciones : 1/3. Supongamos que cada población tiene densidad f_i gaussianas de misma matriz de varianza $\Sigma = \mathrm{Id}_2$ y de esperanzas $\mu_1 = (0,0)$, $\mu_2 = (1,0)$ y $\mu_3 = (0,1)$. Describir las fronteras entre las diferentes clases.
2. Como visto en clase el error asintótico del clasificador del vecino más cercano L_{NN} (1-nearest neighbor) vale
$$L_{NN} = \mathbb{E}\left[2\eta(X)(1-\eta(X))\right]$$
y se mostró $L_{NN} \leq 2L^*$. Mostrar la cota más fuerte $L^* \leq L_{NN} \leq 2L^*(1-L^*)$.
3. Consideramos la regla de clasificación óptima g^* entre dos poblaciones de densidades
$$x \mapsto \lambda e^{-\lambda x}\mathbb{1}_{x\geq 0} \quad \text{con probabilidad } p$$
$$x \mapsto \frac{1}{\sqrt{2\pi}} e^{-(x-a)^2/2} \quad \text{con probabilidad } 1-p$$
para $a > 0$, $\lambda > 0$ y $p \in (0,1)$. Denotamos respectivamente sur regiones de clasificación R_1 y R_2. Mostrar que R_2 es un segmento. Escribir un código que implementa el calculo de R_2 en función de (a, λ, p).
4. La función de densidad exponencial en $\mathbb{R}^2$ está dada por
$$f_\lambda : x \in \mathbb{R}^2 \mapsto C(\lambda)e^{-\lambda\|x-a\|} \quad \text{por un } \lambda > 0 \text{ y un } a \in \mathbb{R}^2.$$
La constante $C(\lambda)$ es una constante de renormalización para asegurar que $\int f_\lambda = 1$. Consideramos la clasificación binaria entre dos clases de exponenciales en dimensión 2 (con dos λ y a diferentes). Mostrar que la separación entre esas clases es un circulo cuando los λ son diferentes y una recta si son iguales.

2 Clasificador Bayes ingenuo

1. Consideramos dos clases de palabras A y B. El conjunto de datos de entrenamiento es

Dato	Clase
aa	A
ab	A
ba	A
bb	B

   (a) Calcular las probabilidades $\mathbb{P}(A)$ y $\mathbb{P}(B)$.

   (b) Calcular las estimaciones de las probabilidades $\mathbb{P}(a|A)$, $\mathbb{P}(b|A)$, $\mathbb{P}(a|B)$ y $\mathbb{P}(b|B)$. Que problema surge ?

   (c) Mostrar que las probabilidades precedentes son suficientes para implementar el clasificador Bayes ingenuo.

   (d) Para solucionar el problema del inciso (b), proponemos usar la suavización
$$\mathbb{P}_s(E) = \frac{\text{“casos positivos de E”} + \alpha}{\text{“total de casos”} + \alpha d}$$
   Calcular las $\mathbb{P}_s(a|A)$, $\mathbb{P}_s(b|A)$, $\mathbb{P}_s(a|B)$ y $\mathbb{P}_s(b|B)$.

   (e) Proponemos los datos de testing siguientes

Palabras :
aaba
a
bbba
bccbba
bbbb

Dar, para cada palabra, el resultado de la clasificación por el método Bayes ingenuo basado en las probabilidades $\mathbb{P}_s$

2. Proponemos los datos de entrenamiento siguientes :

Clase	Tamaño (T)	Peso (P)
1	S	1
1	S	2
1	M	1
1	L	2
1	S	1
1	M	2
2	M	3
2	L	2
2	M	1
2	L	2
2	L	3
2	L	2

   (a) Identificar las probabilidades condicionales necesarias para poder implementar el clasificador empírico óptimo.
   (b) Estimarlas con los datos de entrenamiento. Que problema tenemos ?
   (c) Estimar las leyes empíricas de T y P en cada clase.
   (d) Calcular el riesgo del clasificador Bayes ingenuo basado en las leyes empíricas obtenidas.

3. Consideramos el problema de clasificación donde los datos son de la forma $X = (D, C)$ con $D \in \{0, 1\}$ una variable binaria y C una variable continua en $\mathbb{R}$ asociados a sus etiquetas $Y \in \{0, 1\}$.
   (a) Escribir la verosimilitud de una observación (d, c, y) denotando $p(c|d, y)$ la densidad condicional de C dado D y Y.
   (b) Simplificar la formula anterior cuando se hace el supuesto del clasificador Bayes ingenuo (que supondremos para los incisos siguientes).
   (c) Dar el estimador de máximo de verosimilitud para la probabilidad $\mathbb{P}(D = 1|Y = y)$ basado en un conjunto de datos $(d_i, c_i, y_i)_{i \leq N}$.
   (d) Si supongamos adicionalmente que la distribución condicional de C dado Y es gaussiana, describir los estimadores de máximo de verosimilitud a considerar.
   (e) Escribir un codigo en Python, R o pseudo-code que permite construir el clasificador Bayes ingenuo sobre el conjunto de datos de entrenamiento $(d_i, c_i, y_i)_{i \leq N}$.

3 Discriminación lineal

1. Nos interesamos al discriminante de Fisher
   (a) Escribir una función (en Python, R o pseudo-code) que permite simular N variables gaussianas de media μ y varianza Σ. Visualizar los datos de dos grupos equilibrados con $N_1 = N_2 = 100$, $\mu_1 = (1, 0.5)$, $\mu_2 = (-1, -0.5)$ y

$$\Sigma_1 = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix} \qquad \Sigma_2 = \begin{pmatrix} 1 & -0.5 \\ -0.5 & 1 \end{pmatrix}$$

   (b) Implementar un script que calcula el vector a de proyección asociado al discriminante de Fisher y encontrar el valor óptimo de a_0. Calcular el número de datos mal clasificados.
   (c) Considerar una dirección a aleatoria y calcular el a_0 optimo correspondiente. Guardar el número n_E de datos mal clasificados por este clasificador. Repetir el experimento 200 veces y hacer un histograma de n_E y compararlo al n_E dado en el caso del discriminante de Fisher.

2. Mostrar que tenemos $L^* = L = 0$ (un clasificador lineal óptimo es globalmente óptimo) cuando las dos clases se distribuyen en conjuntos A y B disjuntos convexos y cerrados.

3. Consideramos la clasificación (en dos clases) sobre datos del plano $(x_i, y_i)_i$ dada por la función XOR. Mostrar que existe ninguno clasificador lineal que pueda acercarse del clasificador óptimo teórico.