# Problem set 3

Estimación no paramétrica
Aprendizaje Automático

## 1. P1

Genere 1000 puntos en el eje $x$ a partir de la siguiente pdf:

$$p(x) = \frac{1}{3} \cdot \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left(-\frac{x^2}{2\sigma_1^2}\right) + \frac{2}{3} \cdot \frac{1}{\sqrt{2\pi\sigma_2^2}} \exp\left(-\frac{(x^2-2)^2}{2\sigma_2^2}\right)$$

con $\sigma_1^2 = \sigma_2^2 = 0{,}2$.

Calcule una aproximación a esta pdf con una ventana de Parzen, siendo $h = 0{,}1$ (grafique el resultado). Repita el experimento con:

(a) $h = 0{,}01$ y $N = 1000$.

(b) $h = 0{,}1$ y $N = 20,000$.

## 2. P2

Si consideramos $N$ puntos: $x_1, x_2, \ldots, x_N \in \Re^d$, la estimación de $p(x)$ usando los $k$ vecinos más próximos es:

$$p(x) \approx \frac{k}{NV(x)}$$

1. Escoja un valor para $k$.
2. Encuentre las distancias entre $x$ y todos los puntos $x_i$.
3. Encuentre los $k$ vecinos más próximos a $x$.
4. Calcule el volumen $V(x)$ en el que se encuentran los $k$ vecinos más próximos.
5. Calcule $p(x)$.

Nota. Si la distancia euclidiana es utilizada y esta distancia entre el $k$ vecino más lejano y $x$ es $\rho$, el volumen $V(x)$ es:

$$(d = 1).\ V(x) = 2\rho$$
$$(d = 2).\ V(x) = \pi\rho^2$$
$$(d = 3).\ V(x) = \frac{4}{3}\pi\rho^3$$

Genere los datos $X$ del ejemplo anterior y calcule $p(x)$ con $k = 21$. Repita el problema para $k = 5, 100$ y $N = 5,000$.

## 3. P3

Considere un problema de clasificación bidimensional donde los vectores son generados a partir de dos clases equiprobables $\omega_1$ y $\omega_2$. Las clases son modeladas con dos distribuciones gaussianas con medias: $\mu_\mathbf{1} = [0\ 0]^t$ y $\mu_\mathbf{2} = [1\ 2]^t$ y las matrices de covarianza $\Sigma_1 = \Sigma_2 = \begin{bmatrix} 0{,}8 & 0{,}2 \\ 0{,}2 & 0{,}8 \end{bmatrix}$.

(a) Genere dos conjuntos de datos, $X_1$ y $X_2$, con $1,000$ y $5,000$ puntos respectivamente.

(b) Tome $X_1$ como el conjunto de entrenamiento y clasifique $X_2$ usando un clasificador KNN con $k = 3$ y distancias euclidianas.

(c) Calcule el error de clasificación.

## 4. P4

Podemos hacer algunos cuestionamientos al algoritmo de clasificación de los vecinos más próximos:

(a) El uso de distancias euclidianas no es siempre adecuado. Muestre un ejemplo donde esto suceda y proponga una solución para el caso.

(b) ¿Se necesitan siempre todos los datos de entrenamiento para realizar una clasificación? ¿Existe(n) alguna(s) manera(s) de reducir la complejidad del algoritmo?