Facultad de Matemática, Astronomía y Física y Computación
Universidad Nacional de Córdoba

Ciencia de Datos

Práctico N°4: Técnicas no parámetricas

Problema 1:

Siendo $p(x) \sim U(0, a)$ la distribución uniforme sobre $[0, a]$ y $\phi(x) = \exp(x)$ con $x > 0$, el kernel exponencial,

a) mostrar que la esperanza del estimador basado en ventana de Parzen exponencial, de arista h_n, resulta

$$\overline{p}_n(x) = \begin{cases} 0 & x < 0\,, \\ \frac{1}{a}\left(1 - e^{-x/h_n}\right) & 0 \le x \le a\,, \\ \frac{1}{a}\left(e^{a/h_n} - 1\right) e^{-x/h_n} & a \le x\,. \end{cases}$$

b) Graficar esta curva con $a = 1$, y usando los valores $h_n = 1, 1/4$ y $1/16$.

c) ¿Cuán pequeño tiene que ser h_n para obtener menos del 1 % de desvio sobre el 99 % del rango $0 < x < a$?

d) Calcular h_n para la condicion anterior si a=1 y graficar $\overline{p}_n(x)$ en el rango $0 \le x \le 0{,}05$.

Problema 2: Estudiar la *regla del vecino más cercano* en la sección 4.5 sobre del libro *Pattern Classification*, R.O. Duda, P.E. Hart, and D.G. Stork, Wiley 2nd ed (2001).

Denotando con $P_n(e)$ la probabilidad de error para la regla del vecino más cercano con muestras de tamaño n y

$$P = \lim_{n\to\infty} p_n(e)\,,$$

probar que se cumplen las desigualdades

$$P^* \le P \le P^*\left(2 - \frac{c}{c-1}P^*\right),$$

donde P^* es el error de Bayes y c el número de clases.

Problema 3: Estudiar la implementación de $k-$nearest neighbors provista por scikit-learn.

a) Aplicar $k-$nearest neighbors para clasificar el iris dataset.

b) Comparar con el resultado de naïve Bayes. Discutir las matrices de confusión resultantes.

Problema 4: Estudiar la implementación de Kernel Density Estimation provista en el libro *Python Data Science Handbook*, Jake VanderPlas, O'Reilly Media (2016).

a) Identicar los priors y la función de pérdida codificada.

b) Usar este clasificador con Kernel gaussiano con la hand-written digits database, usando el ancho de banda default. Comparar con el valor estimado usando `GridSearchCV`. ¿Qué exactitud (accuracy) se obtiene usando el bandwidth estimado y cuál es la exactitud usando el bandwidth default?

c) Encontrar el ancho de banda óptimo usando `GridSearchCV` para clasificadores con kernels `exponencial` y `epanechnikov` para la database digits. Comparar el valor de accuracy obtenido con el bandwidth default.

Soluciones:

Problema 1:

a) Para una muestra de tamaño n, el estimador de la densidad viene dado por

$$\hat{p}_n(x) = \frac{1}{n}\sum_i \frac{1}{V_n}\phi\left(\frac{x - x_i}{h_n}\right),$$

donde x_i son las muestras que caen dentro de la región de volumen V_n centrada en x. Esta región es, para el caso general, un hipercubo de lado h_n y por lo tanto $V_n = h_n^d$, donde d es la dimensión del espacio.

Suponiendo que las muestras x_i provienen de una distribución unidimensional uniforme $U(0, a)$, y que $\phi(x) = e^{-x}$ para $x > 0$, entonces el valor esperado de $\hat{p}_n(x)$ resulta

$$\begin{aligned}
E[\hat{p}_n(x)] &= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h^1}E\left[\exp\left(-\frac{x - x_i}{h_n}\right) I_{(0,x)}(x_i)\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}\int_0^a \exp\left(-\frac{x-u}{h_n}\right) I_{(0,x)}(u)\,\frac{1}{a}\,du,
\end{aligned}$$

donde como la sumatoria calcula siempre el mismo valor se puede reemplazar por n y función indicadora está incluida porque la densidad exponencial es nula para valores negativos de x.

- Si $x < 0$, entonces para todo $u \in (0, a)$ se cumple $x - u < 0$. Por lo tanto $I_{(0,x)}(u) = 0$ y el valor esperado resulta nulo.
- Si $0 < x < a$, entonces la integral es nula para $u > x$.
- Si $x \geq a$, entonces para todo $u \in (0, a)$ se cumple $x - a > 0$. Luego $I_{(0,x)}(u) = 1$.

En el caso $0 < x < a$,

$$\begin{aligned}
E\left[\hat{p}_n(x)\right] &= \frac{1}{a\,h_n}\int_0^x \exp\left(-\frac{x-u}{h_n}\right) du \\
&= \frac{1}{a\,h_n}\exp\left(-\frac{x}{h_n}\right) h_n \left[\exp\left(\frac{u}{h_n}\right)\right|_0^x \\
&= \frac{1}{a}\,(1 - e^{-x/h_n}).
\end{aligned}$$

En particular, para $h_n \to 0$ el valor esperado tiende a la distribución uniforme en el intervalo $(0, a)$.
En el xaso $x \geq a$,

$$\begin{aligned}
E\left[\hat{p}_n(x)\right] &= \frac{1}{a\,h_n}\int_0^a \exp\left(-\frac{x-u}{h_n}\right) du \\
&= \frac{1}{a\,h_n}\exp\left(-\frac{x}{h_n}\right) h_n \left[\exp\left(\frac{u}{h_n}\right)\right]_0^a \\
&= \frac{1}{a}\,e^{-x/h_n}\,(e^{a/h_n} - 1).
\end{aligned}$$

b) La varianza del estimador está dado por

$$\begin{aligned}
\mathrm{Var}(\hat{p}) &= E\left[(\hat{p} - p)^2\right] = \frac{1}{a}\int_0^a (1 - \hat{p}(u))^2\,du \\
&= \frac{1}{a^2}\int_0^a \left(1 - (1 - e^{-u/h_n})\right)^2 du = \frac{1}{a^2}\int_0^a e^{-2u/h_n}\,du \\
&= \frac{-h_n}{2\,a^2}\left[e^{-2u/h_n}\right]_0^a = \frac{h_n}{2\,a^2}\left(1 - e^{-2a/h_n}\right).
\end{aligned}$$

El sesgo (bias) sobre $0 < x < a$ es

$$\hat{p}(x) - p(x) = \frac{1}{a}\left(1 - (1 - e^{-x/h_n}\right) = \frac{1}{a}\, e^{-x/h_n}.$$

El sesgo es decreciente en x. Entonces, para que el sesgo sea menor que el 1 % en el 99 % del intervalo $(0, a)$ se tiene que cumplir $x \geq 0{,}01\, a$ y

$$\frac{1}{a}\, e^{-x/h_n} < 0{,}01, \; -\frac{x}{h_n} < \ln(0{,}01\, a), \; -\frac{x}{\ln(0{,}01\, a)} > h_n,$$

dado que $\ln(0{,}01\, a) < 0$. Así, se tiene que

$$h_n < \frac{0{,}01}{-\ln(0{,}01\, a)}.$$

Para el caso $a = 1$ resulta aproximadamente $h_n < 0{,}0022$.

Problema 2: Denotando con $P_n(e)$ la probabilidad de error para la regla del vecino más cercano y muestras de tamaño n, cuando una muestra x', que es vecino más cercano de x, se la clasifica en una clase errónea resulta

$$P_n(e) = \sum_{x,x'} P_n(e|x, x')\, p(x, x').$$

Sean n muestras correspondientes a c clases distintas. Cada muestra es un par (x', θ_j), donde θ_j significa que x' corresponde a la clase j. Así un error ocurre si x' es el vecino más cercano a x y θ_j es distinto de la clase θ de x. Suponiendo que las muestras son todas independientes entre sí, se tiene que

$$P(\theta, \theta_j|x, x') = P(\theta|x)\, P(\theta_j|x').$$

Sumando sobre todas la posibles clases de x ($1 \leq i \leq c$), la probabilidad del error es uno menos la probabilidad de los aciertos,

$$P(e|x, x') = 1 - \sum_{i=1}^{c} P(\theta = i, \theta_j = i|x, x') = 1 - \sum_{i=1}^{c} P(\theta|x)\, P(\theta_j = i|x'),$$

$$P(e|x) = \int P(e|x, x')\, p(x'|x)\, dx' = \int \left(1 - \sum_{i=1}^{c} P(\omega_i|x)\, P(\omega_i|x')\right) p(x'|x)\, dx'.$$

Para $n \to \infty$, x' está muy próximo a x y puede suponerse que la densidad $p(x'|x)$ converge a una delta de Dirac, $p(x'|x) \mapsto \delta(x - x')$. Así resulta,

$$P(e|x) = 1 - \sum_{i=1}^{c} \left(P(\omega_i|x)\right)^2.$$

Es evidente que el error de Bayes cumple $P^* \leq P$. Para calcular la cota superior, sea x y su clase ω_m. Se quiere minimizar

$$\sum_{i=1}^{c} P^2(\omega_i|x) = P^2(\omega_m|x) + \sum_{i \neq m}^{c} P^2(\omega_i|x).$$

El segundo término se minimiza si todas las probabilidades en la sumatoria son iguales, esto es

$$P(\omega_i|x) = \frac{1 - P(\omega_m|x)}{c - 1}.$$

Como $1 - P(\omega_m|x) = P^*(e|x)$ se tiene que

$$\sum_{i=1}^{c} P^2(\omega_i|x) \geq (1 - P^*(e|x))^2 + \frac{P^{*2}(e|x)}{c-1}.$$

Por otra parte,

$$\mathrm{Var}(P^*(e|x)) = \int (P^*(e|x) - P^*)^2 \, p(x)\, dx = \int P^{*2}(e|x)\, p(x)\, dx - P^{*2} \geq 0,$$

de donde se obtiene que

$$\int P^{*2}(e|x)\, p(x)\, dx \geq P^{*2}.$$

De todo lo anterior, resulta que para un número finito de muestras,

$$P^* \leq P \leq P^* \left(2 - \frac{c}{c-1}\, P^* \right).$$

FaMAF 2023