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

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: Para la estimación Bayesiana de la media μ desconocida de una distribución Gaussiana unidimensional de varianza σ^2 conocida, suponer como distribución a priori para el parámetro desconocido una distribución normal con parámetros (μ_0, σ_0), $p(\mu) \sim N(\mu_0, \sigma_0)$.
a) Calcular con todo detalle $p(\mu|\mathcal{D})$ y $p(x|\mathcal{D})$. Escribir la expresión de *dogmatismo*; es decir, el balance entre el conocimiento previo y el conocimiento empírico proveniente de los datos.
b) Graficar la densidad $p(x|\mathcal{D})$ obtenida por estimación bayesiana eligiendo los valores μ_0, σ_0 y σ y una muestra de entrenamiento $\mathcal{D} = \{x_1, \cdots, x_n\}$ generada conociendo el valor de μ.

Problema 4: Utilizando los datos empíricos tridimensionales correspondientes a tres clases independientes, dados en la siguiente tabla, construir modelos gaussianos de clasificación para cada uno de los ítems a continuación.
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).

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

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).
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, 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?

Problema 6: Naïve Bayes
a) Descargar el dataset Loan Data disponible en Kaggle. Indagar el diccionario de las columnas del dataset y en qué consiste el problema de predicción.
b) Estudiar el tutorial de `datacamp` para aprender a implementar un clasificador Naïve Bayes y cómo evaluar sus resultados usando Scikit-learn.
c) Estudiar las métricas de clasificación binaria derivadas de la matriz de confusión e interpretar el resultado del clasificador Naïve Bayes aplicado al dataset sobre cráditos.

Soluciones:

Problema 3

a) Conocida σ y dada una muestra independiente $\mathcal{D} = \{\S_\infty, \S_\in, \ldots, \S_\backslash\}$, se tiene que

$$p(x_i|\mu) = \frac{1}{\sqrt{2\,\pi}\sigma} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right).$$

La estimación a posteriori de μ dada la muestra $\mathcal{D}$ está dada por

$$p(\mu|\mathcal{D}) = \frac{p(\mathcal{D}|\mu)p(\mu)}{\displaystyle\int p(\mathcal{D}|\mu)\,p(\mu)\,d\mu}.$$

A continuación se usará la notación

$$\alpha = p(\mathcal{D})^{-1} = \left(\int p(\mathcal{D}|\mu)\,p(\mu)\,d\mu\right)^{-1}.$$

Por otra parte, como $\mathcal{D}$ es una muestra de observaciones independientes, su probabilidad es el producto de las probabilidades de los valores individuales individuales. Así resulta

$$p(\mu|\mathcal{D}) = \alpha\, p(\mathcal{D}|\mu)p(\mu) = \alpha\left(\prod_{i=1}^{n} p(x_i|\mu)\right)\,p(\mu) \text{ y entonces}$$

$$p(\mu|\mathcal{D}) = \alpha\left(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma}\, e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}\right)\,\frac{1}{\sqrt{2\pi}\sigma_0}\, e^{-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}} = \frac{\alpha}{(2\pi)^{(n+1)/2}\,\sigma^n\,\sigma_0}\, e^{-\displaystyle\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2}}\, e^{-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}}$$

Sumando los exponentes resulta

$$\begin{aligned}
-\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2} - \frac{(\mu-\mu_0)^2}{2\sigma_0^2} &= -\frac{1}{2}\left(\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{\sigma^2} + \frac{(\mu-\mu_0)^2}{\sigma_0^2}\right)\\
&= -\frac{1}{2}\left(\sum_{i=1}^{n}\frac{x_i^2}{\sigma^2} - 2\sum_{i=1}^{n}\frac{x_i\,\mu}{\sigma^2} + n\,\frac{\mu^2}{\sigma^2} + \frac{\mu^2}{\sigma_0^2} - 2\frac{\mu_0\,\mu}{\sigma_0^2} + \frac{\mu_0^2}{\sigma_0^2}\right)\\
&= -\frac{1}{2}\left(\sum_{i=1}^{n}\frac{x_i^2}{\sigma^2} + \frac{\mu_0^2}{\sigma_0^2}\right) - \frac{1}{2}\left(\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)\mu^2 - 2\left(\sum_{i=1}^{n}\frac{x_i}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}\right)\mu\right)\\
&= A + B + C
\end{aligned}$$

El término A no depende de μ por lo que e^A se absorve en la constante de normalización α. Por su lado, con los términos B y C se puede completar el cuadrado de un binomio. Para ello se definen las constantes

$$a = \frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}, \qquad b = \sum_{i=1}^{n}\frac{x_i}{\sigma^2} + \frac{\mu_0}{\sigma_0^2} = \frac{n}{\sigma^2}\,\overline{x_n} + \frac{\mu_0}{\sigma_0^2}, \quad \text{con } \overline{x_n} = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

siendo $\overline{x_n}$ la media muestral de los n datos observados. De esta forma, se puede escribir

$$\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)\mu^2 - 2\left(\sum_{i=1}^{n}\frac{x_i}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}\right)\mu = a\,\mu^2 - 2\,b\,\mu = a\left(\mu^2 - 2\,\frac{b}{a}\,\mu + \left(\frac{b}{a}\right)^2 - \left(\frac{b}{a}\right)^2\right) = \frac{(\mu-\frac{b}{a})^2}{1/a} - \frac{b^2}{a}.$$

El término b^2/a nuevamente se puede absorber en la normalización y así resulta que $p(\mu|\mathcal{D})$ es una densidad normal con media

$$\mu_n = \frac{b}{a} = \frac{n\,\overline{x_n}\,\sigma_0^2 + \mu_0\,\sigma^2}{n, \sigma_0^2 + \sigma^2}$$

y desviación estándar

$$\sigma_n = \frac{1}{a} = \frac{n\,\sigma_0^2 + \sigma^2}{\sigma^2\,\sigma_0^2}.$$

Conociendo $p(x|\mu)$ (modelo de los datos) y $p(\mu|\mathcal{D})$ puede ahora calcularse $p(x|\mathcal{D})$.

$$\begin{aligned}
p(x|\mathcal{D}) &= \int p(x|\mu)\,p(\mu|\mathcal{D})\,d\mu \\
&= \int \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \frac{1}{\sqrt{2\pi}\sigma_n} e^{-\frac{(\mu-\mu_n)^2}{2\sigma_n^2}}\,d\mu \\
&= \frac{1}{2\pi\,\sigma\,\sigma_n} \int \exp\left(-\frac{(x-\mu)^2}{2\,\sigma^2} - \frac{(\mu-\mu_n)^2}{2\,\sigma_n^2}\right) d\mu \\
&= \frac{1}{2\,\pi\,\sigma\,\sigma_n} \int \exp\left(-\frac{\sigma_n^2\,(x^2 - 2\mu x + \mu^2)}{2\,\sigma^2\,\sigma_n^2} - \frac{\sigma^2\,(\mu^2 - 2\mu\mu_n + \mu_n^2)}{2,\sigma_n^2\,\sigma^2}\right) d\mu \\
&= \frac{1}{2\,\pi\,\sigma\,\sigma_n} \int \exp\left(-\frac{(\sigma_n^2 + \sigma^2)\,\mu^2 - 2\,(\sigma_n^2\,x + \sigma^2\mu_n)\,\mu + \sigma_n^2\,x^2 + \sigma^2\,\mu_n^2}{2\,\sigma_n^2\,\sigma^2}\right) d\mu \\
&= \frac{1}{2\,\pi\,\sigma\,\sigma_n} \int \exp\left(-\frac{\left(\mu - \frac{\sigma_n^2\,x + \sigma^2\,\mu_n}{\sigma_n^2 + \sigma^2}\right)^2}{\frac{2\,\sigma_n^2\,\sigma^2}{\sigma_n^2 + \sigma^2}}\right) d\mu\ \exp\left(-\frac{\sigma_n^2\,x^2 + \sigma^2\,\mu_n^2 - \frac{(\sigma_n^2\,x + \sigma^2\,\mu_n)^2}{\sigma_n^2 + \sigma^2}}{2\,\sigma_n^2\,\sigma^2}\right)
\end{aligned}$$

El integrando de la integral respecto de μ es proporcional a una densidad normal con media y desvío estándar dado por

$$\mu_1 = \frac{\sigma_n^2\,x + \sigma^2\mu_n}{\sigma_n^2 + \sigma^2} \qquad \sigma_1 = \sqrt{\frac{\sigma_n^2\,\sigma^2}{\sigma_n^2 + \sigma^2}}\,.$$

Por lo tanto la integral es igual a $\sqrt{2\pi}\,\sigma_1$, lo cual es independiente del valor de x. El otro factor se compacta observando que el numerador puede reescribirse según

$$\begin{aligned}
&\sigma_n^2\,x^2 + \sigma^2\,\mu_n^2 - \frac{(\sigma_n^2\,x + \sigma^2\,\mu_n)^2}{\sigma_n^2 + \sigma^2} \\
= \ & \frac{1}{\sigma_n^2 + \sigma^2}\left(\sigma_n^4\,x^2 + \sigma_n^2\,\sigma^2\,\mu_n^2 + \sigma_n^2\,\sigma^2\,x^2 + \sigma^4\,\mu_n^2 - \sigma_n^4\,x^2 - 2\,\sigma_n^2\,\sigma^2\,\mu_n\,x - \sigma^4\,\mu_n^2\right) \\
= \ & \frac{\sigma_n^2\,\sigma^2}{\sigma_n^2 + \sigma^2}\left(x^2 - 2\,\mu_n\,x + \mu_n^2\right) = \frac{\sigma_n^2\,\sigma^2}{\sigma_n^2 + \sigma^2}(x - \mu_n)^2\,.
\end{aligned}$$

De estos resultados se concluye finalmente que

$$p(x|\mathcal{D}) = \mathcal{N}\left(\mu_n, \sqrt{\sigma_n^2 + \sigma^2}\right)\,.$$

Dogmatismo $= \dfrac{\sigma^2}{\sigma_0^2}$.