Modelos Estadísticos para la Regresión y la Clasificación

Práctico 3 - Estimación de máxima verosimilitud

Micaela Long

Instituto de Matemática y Estadística Prof. Rafael Laguardia (IMERL)
Facultad de Ingeniería, Universidad de la República, Uruguay

23 de agosto de 2024

Práctico 3
Probabilidad

Material para hacer práctico 3 (disponible en EVA):

- Teóricos 19/8 y 21/8 (misma presentación)

Algunas definiciones...

1. Un **estimador** $\hat{\theta}$ es una función de los datos, que se utiliza para aproximar un parámetro poblacional θ desconocido.

2. Una **estimación** es un valor numérico obtenido al aplicar el estimador $\hat{\theta}$ a una muestra específica.

3. Ejemplo: dada X variable aleatoria, si queremos estimar $\theta = \mathbb{E}(X)$ podemos usar como **estimador** el promedio de la muestra:

$$\hat{\theta}_n = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = T(X_1, \cdots, X_n)$$

Dada una muestra $x_1, \cdots, x_n$, una **estimación** de $\mathbb{E}(X)$ es

$$(\hat{\theta}_n)_{obs} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

1. **Sesgo de un estimador:**
$$\text{Sesgo}(\hat{\theta}_n) = \mathbb{E}(\hat{\theta}_n) - \theta$$

2. Un estimador es **insesgado** si
$$\text{Sesgo}(\hat{\theta}_n) = 0$$

3. Un estimador es **asintóticamente insesgado** si
$$\lim_{n\to\infty} \text{Sesgo}(\hat{\theta}_n) = 0$$

4. **Varianza de un estimador:**
$$\text{Var}(\hat{\theta}_n) = \mathbb{E}\left[\left(\hat{\theta}_n - \mathbb{E}(\hat{\theta}_n)\right)^2\right]$$

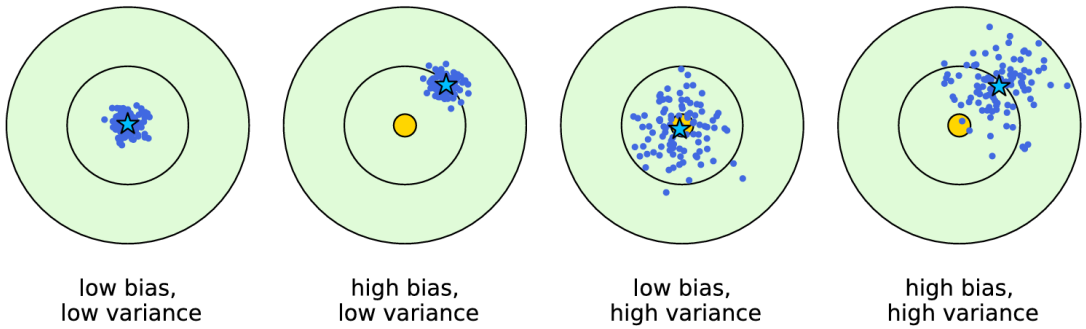

Figure 3: The classic dartboard analogy for explaining bias and variance.

Figura: A Unified Theory of Diversity in Ensemble Learning, Wood et al, 2023

1. **Error cuadrático medio (ECM)**

$$ECM(\hat{\theta}_n) = \mathbb{E}(\hat{\theta}_n - \theta)^2$$

2. **Descomposición sesgo varianza**

$$ECM(\hat{\theta}_n) = \text{Sesgo}(\hat{\theta}_n)^2 + \text{Var}(\hat{\theta}_n)$$

   Si el estimador es insesgado tenemos:

$$ECM(\hat{\theta}_n) = \text{Var}(\hat{\theta}_n)$$

   Para estos estimadores minimizar el ECM es minimizar la varianza.

3. La **cota inferior de Cramér-Rao** proporciona una cota inferior teórica para la varianza de cualquier estimador insesgado de un parámetro desconocido:

$$\text{Var}(\hat{\theta}) \geq \frac{1}{I(\theta)} \quad \text{con} \quad I(\theta) = -\mathbb{E}\left[\frac{\partial^2 \log f_\theta(x)}{\partial \theta^2}\right]$$

   donde $I(\theta)$ es la información de Fisher del parámetro θ (mide la cantidad de información que proporcionan los datos sobre θ).

4. Un estimador es **eficiente** si es insesgado y tiene varianza mínima (alcanza la cota de CR).

5. Un estimador es **asintóticamente eficiente** si en el límite alcanza la varianza mínima.

- Queremos estimador insesgado con varianza mínima (que alcance la cota de CR).
- Si podemos factorizar

$$\frac{\partial \log f_\theta(x)}{\partial \theta} = I(\theta)(g(x) - \theta)$$

entonces $\hat{\theta} = g(x)$ es el estimador MVU y su varianza es $\frac{1}{I(\theta)}$

- **Observación:** El estimador MVU puede no existir.

Un estimador $\hat{\theta}_n$ de θ es **consistente** si $\forall \epsilon > 0$

$$\lim_{n \to \infty} \mathbb{P}(|\hat{\theta}_n - \theta| < \epsilon) = 1$$

Es decir, para cualquier $\epsilon > 0$, la probabilidad de que $\hat{\theta}_n$ esté a una distancia mayor que ϵ de θ tiende a cero a medida que n crece.

Estimador de máxima verosimilitud (MLE)

La idea central es encontrar el valor del parámetro que maximiza la función de verosimilitud:

$$\mathcal{L}_n(\theta) = \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n \mid \theta) = \prod_{i=1}^{n} \mathbb{P}(X_i = x_i \mid \theta)$$

que es una medida de **qué tan probable es que los datos observados hayan sido generados por un modelo con un cierto valor del parámetro**.

Formalmente

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta} \mathcal{L}_n(\theta)$$

Si $\mathcal{L}_n(\theta_1) > \mathcal{L}_n(\theta_2)$ entonces es más probable que los datos hayan sido generados por un modelo que tiene como parámetro a θ_1, que por un modelo que tenga como parámetro a θ_2.

Ejercicio 1: Estimación por máxima verosimilitud

Sea $X_1, \ldots, X_n$ un muestreo aleatorio de una distribución discreta con recorrido $\{0, 1, 2, 3\}$. Supongamos que el parámetro θ solo puede tomar los valores $\theta = 0$ y $\theta = 1$. La función de probabilidad puntual para $\theta = 0$ y $\theta = 1$ es:

X	$\theta = 0$	$\theta = 1$
$X = 0$	0,1	0,2
$X = 1$	0,3	0,4
$X = 2$	0,3	0,3
$X = 3$	0,3	0,1

Se tiene una muestra con $n = 6$ y los datos son $0, 3, 1, 2, 0, 3$. Hallar el estimador de máxima verosimilitud de θ.

Ejercicio 1

Función de verosimilitud (que queremos maximizar):

$$\mathcal{L}(\theta) = \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n \mid \theta) = \prod_{i=1}^{n} \mathbb{P}(X_i = x_i \mid \theta)$$

para calcularla utilizamos los valores de la tabla

X	$P(X \mid \theta = 0)$	$P(X \mid \theta = 1)$
0	0,1	0,2
1	0,3	0,4
2	0,3	0,3
3	0,3	0,1

y que $(x_1, \cdots, x_6) = (0, 3, 1, 2, 0, 3)$.

$$\mathcal{L}(0) = 0{,}1 \times 0{,}3 \times 0{,}3 \times 0{,}3 \times 0{,}1 \times 0{,}3 = 0{,}1^2 \times 0{,}3^4 = 0{,}000081$$

$$\mathcal{L}(1) = 0{,}2 \times 0{,}1 \times 0{,}4 \times 0{,}3 \times 0{,}2 \times 0{,}1 = 0{,}2^2 \times 0{,}1^2 \times 0{,}4 \times 0{,}3 = 0{,}000048$$

Como $\mathcal{L}(0) > \mathcal{L}(1)$, el estimador de máxima verosimilitud es $\hat{\theta} = 0$.

Observación: Si tuvieramos más observaciones

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} \mathbb{P}(X_i = x_i \mid \theta)$$

sería más costosa (computacionalmente) de calcular, podría dar números de magnitud muy chica (estamos multiplicando probabilidades).

Por esto muchas veces consideramos

$$\log(\mathcal{L}(\theta)) = \log\left(\prod_{i=1}^{n} \mathbb{P}(X_i = x_i \mid \theta)\right) = \sum_{i=1}^{n} \log(\mathbb{P}(X_i = x_i \mid \theta))$$

teniendo en cuenta que $\log(x)$ es una función monótona creciente.

Y en lugar de maximizar lo anterior, minimizamos

$$-\log(\mathcal{L}(\theta))$$

Ejercicio 2: Bayes vs. Máxima verosimilitud

Se tiene 3 monedas con probabilidad de cara p igual a 0.4, 0.5 y 0.6, respectivamente. Beto toma una de las monedas y se la da a Ana. Después de lanzar la moneda 100 veces, Ana obtiene cara 53 veces.

1. Hallar una estimación de p basada en el método de máxima verosimilitud.
2. Si se sabe que Beto elige las monedas con probabilidad 0.1, 0.4, y 0.5 respectivamente. ¿Cambiarías tu estimación?

Ejercicio 2
Parte 1

$$\mathcal{L}_n(p) = \mathbb{P}(X_1 = x_1, \ldots, X_{100} = x_n \mid p)$$

Nuestros datos son 53 caras en $n = 100$ lanzamientos:

$$\mathcal{L}_{100}(0{,}4) = \mathbb{P}(X_1 = x_1, \ldots, X_{100} = x_{100} \mid p = 0,4) = \binom{100}{53} 0{,}4^{53}(1 - 0{,}4)^{47} = 0{,}00256$$

es la probabilidad de sacar 53 caras (y 47 números), por la cantidad de formas de sacar 53 caras en 100, dado que la probabilidad de sacar cara es 0,4.

Observación: $\mathcal{L}_n(p)$ es una binomial!

De la misma forma

$$\mathcal{L}_{100}(0{,}5) = \binom{100}{53} 0{,}5^{53}(1 - 0{,}5)^{47} = 0{,}0666$$

$$\mathcal{L}_{100}(0{,}6) = \binom{100}{53} 0{,}6^{53}(1 - 0{,}6)^{47} = 0{,}0292$$

Luego el estimador de máxima verosimilitud es $\hat{p} = 0{,}5$

Ejercicio 2
Parte 2

Para la parte 2, tenemos las probabilidades a priori de elegir cada moneda. En el enfoque bayesiano, p es una variable aleatoria:

$$p = \begin{cases} 0{,}4 & \text{con probabilidad } 0{,}1 \\ 0{,}5 & \text{con probabilidad } 0{,}4 \\ 0{,}6 & \text{con probabilidad } 0{,}5 \end{cases}$$

o en otras palabras

$$\mathbb{P}(p = 0{,}4) = 0{,}1$$
$$\mathbb{P}(p = 0{,}5) = 0{,}4$$
$$\mathbb{P}(p = 0{,}6) = 0{,}5$$

Lo que queremos maximizar es la probabilidad a posteriori, que recordando la **fórmula de Bayes** es:

$$\mathbb{P}(p = p_i|x) = \frac{\overbrace{\mathbb{P}(x \mid p = p_i)}^{\text{verosimilitud } \mathcal{L}(p_i)} \mathbb{P}(p = p_i)}{\sum_{j=1}^{3} \mathbb{P}(x \mid p = p_j)\mathbb{P}(p = p_j)}$$

Calcular $\mathbb{P}(p = 0{,}4|x), \mathbb{P}(p = 0{,}5|x), \mathbb{P}(p = 0{,}6|x)$ *y ver si el estimador es distinto al MLE!*

Ejercicio 3: Momentos vs. Máxima Verosimilitud

Sea $X_1, \ldots, X_n$ un muestreo i.i.d. de una distribución normal $N(\theta, \theta)$.

1. Calcular los estimadores de momentos $\hat{\theta}_M$ y de máxima verosimilitud $\hat{\theta}_{MLE}$ de θ
2. Nos enfocamos en esta parte en el estimador de máxima verosimilitud.

   a) Probar que $\hat{\theta}_{MLE}$ es sesgado pero asíntóticamente insesgado

   b) Probar que $\mathrm{Var}(\hat{\theta}_{MLE})$ alcanza la cota de Crámer-Rao.

Estimador de momentos

- **Momento**: esperanza de una potencia de una variable aleatoria.
- Momento de primer orden es $\mathbb{E}(X)$, el momento de segundo orden es $\mathbb{E}(X^2)$, etc.
- Estimador de momentos se basa en la idea de hacer coincidir los **momentos muestrales** (medidas estadísticas calculadas a partir de los datos) con **los momentos poblacionales** (valores esperados derivados de la distribución teórica) para encontrar el parámetro desconocido.

$$\mathbb{E}(X) = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\mathrm{Var}(X) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

- En nuestro ejercicio $\theta = \mathbb{E}(X) = \mathrm{Var}(X)$.

Ejercicio 4: Sesgo de un estimador

Un estimador $\hat{\theta}$ de un parámetro θ tiene distribución normal. Se sabe además que $\text{ECM}(\hat{\theta}) = 8$ y $P(\hat{\theta} \leq \theta) = 0{,}8413$. Hallar el sesgo de $\hat{\theta}$.

Sugerencia: Usar

$$X \sim \mathcal{N}(0,1) \Rightarrow \mathbb{P}(X \leq 1) = 0{,}8413$$

Viernes 30/8:

- Retomamos práctico 3.