Machine Learning para Inteligencia Artificial

Máxima Verosimilitud

Universidad ORT Uruguay

26 de Marzo, 2025

La función de verosimilitud

- Datos $T = \{z_1, \ldots, z_N\}$
- Suponemos muestra iid de una distribución $p(z; \theta)$ parametrizada con θ
- La "probabilidad" de observar T es

$$p(T; \theta) = p(z_1; \theta) \cdot p(z_2; \theta) \cdots p(z_N; \theta)$$

- Para T dado, la verosimilitud de un parámetro θ es $L(\theta) = p(T; \theta)$
- La función de verosimilitud es $\theta \mapsto L(\theta)$
- **El principio de máxima verosimilitud**:

$$\boxed{\widehat{\theta} := \arg\max_{\theta} L(\theta)}$$

Menos el log de la verosimilitud

- Es más sencillo trabajar con (menos) el logaritmo de la verosimilitud

$$\ell(\theta) := -\ln L(\theta) = -\sum_{i=1}^{N} \ln p(z_i, \theta)$$

- Entonces, de forma equivalente tenemos

$$\widehat{\theta} = \arg\max_{\theta} L(\theta) = \arg\min_{\theta} \ell(\theta)$$

- Notar que $\widehat{\theta}$ depende de T

Ejemplo

- Consideremos nuevamente la densidad

$$p(z;\theta) = \frac{1}{2}(1+\theta z) \quad -1 \leq z \leq 1$$

  El parámetro θ también varía entre -1 y 1.

- La función de verosimilitud es

$$L(\theta) = \frac{1}{2^N}\prod_{i=1}^{N}(1+\theta z_i)$$

  que es un polinomio de grado N en θ.

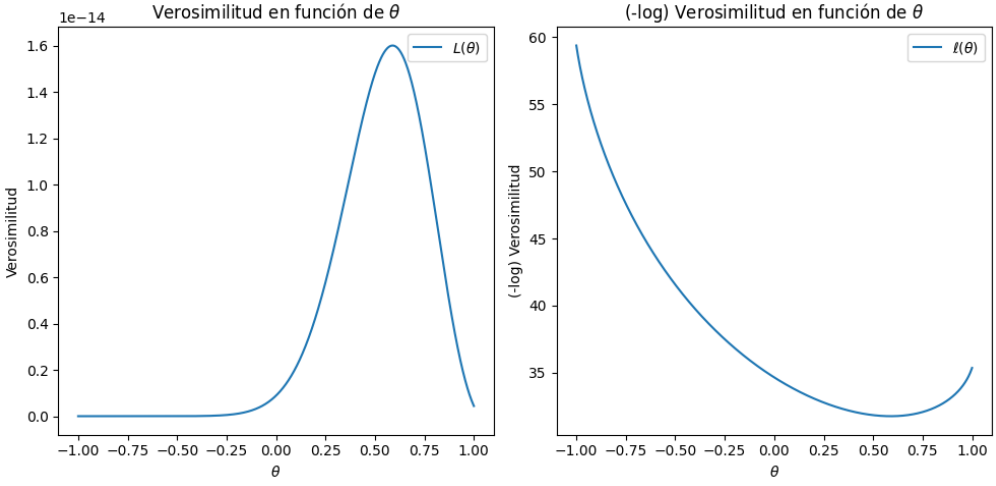
- $L'(\theta) = 0$ no tiene solución analítica (raíz de un polinomio de grado $N-1$)

Ejemplo: la muestra

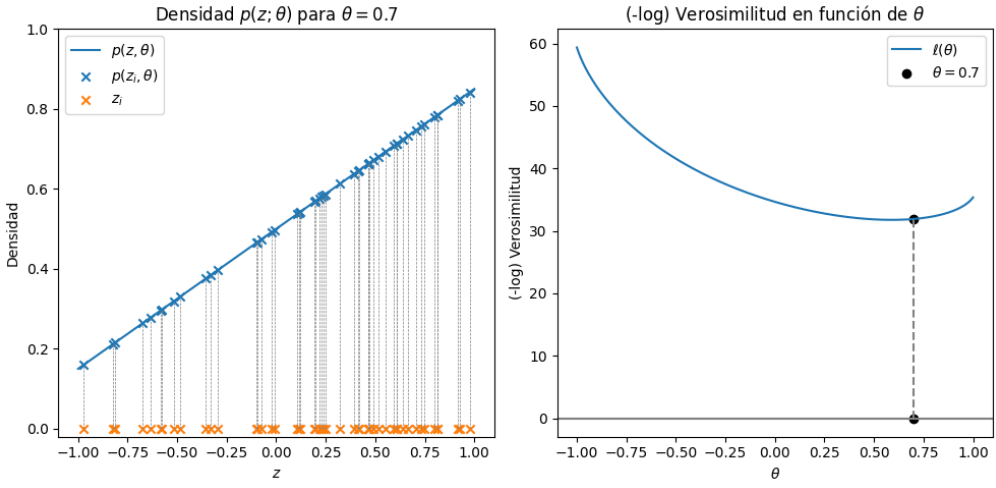
En este ejemplo la muestra (de tamaño $N = 50$) simulada con $\theta = 0.4$:

-0.81	-0.97	0.22	0.19	0.32	0.66	0.21	0.97	0.80	-0.51
-0.01	-0.57	-0.33	0.51	0.60	-0.82	0.10	0.24	0.59	0.73
-0.63	-0.48	-0.35	0.91	-0.67	0.74	0.23	-0.09	-0.02	0.63
0.19	-0.58	0.49	0.81	0.92	0.60	-0.29	0.46	0.81	0.46
0.41	-0.07	0.11	0.11	0.39	-0.09	0.46	0.55	0.41	0.70

Ejemplo: la verosimilitud

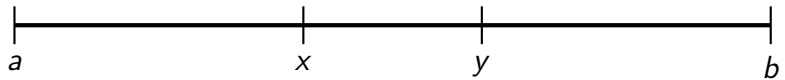


Ejemplo: la verosimilitud



Proportional Search

Dado un intervalo $[a, b]$



Subdividimos $[a, b]$ utilizando una proporción $\rho \in [0, 1/2]$:

$$x = a + \rho\,(b - a) \quad \text{e} \quad y = b - \rho\,(b - a)$$

- Si descartamos la parte superior, nos queda el intervalo $[a, y]$
- Si descartamos la parte inferior, nos queda el intervalo $[x, b]$

Proportional Search

1. Inicializar a y b que encierren al mínimo de la función $\ell(\theta)$
   Elegir proporción $\rho \in [0, 1/2]$
   Elegir tolerancia $\tau > 0$
2. Subdividir $x \leftarrow a + \rho\,(b - a)$ e $y \leftarrow b - \rho\,(b - a)$
3. Evaluar la función $\ell(x)$ y $\ell(y)$
4. Si $\ell(x) < \ell(y)$:
   4.1 $\hat{\theta} \leftarrow x$
   4.2 Reasignamos las variables $a \leftarrow a$, $b \leftarrow y$
5. Si $\ell(x) \geqslant \ell(y)$:
   5.1 $\hat{\theta} \leftarrow y$
   5.2 Reasignamos las variables $a \leftarrow x$, $b \leftarrow b$
6. Repetir desde 2. hasta que $|b - a| \leq \tau$. Devolver $\hat{\theta}$.

Descenso por Gradiente (derivada)

1. Inicializar θ
   Elegir tasa de aprendizaje $\alpha > 0$
   Elegir tolerancia $\tau > 0$

2. Calcular la derivada $\ell'(\theta)$

3. Actualizar el parámetro:
$$\theta \leftarrow \theta - \alpha\, \ell'(\theta)$$

4. Repetir desde 2. hasta que $|\ell'(\theta)| \leq \tau$ y devolver θ

Descenso por Gradiente Estocástico

Denotando $\ell_i(\theta) := -\ln p(z_i; \theta)$ tenemos que $\ell(\theta) = \sum_{i=1}^{N} \ell_i(\theta)$.

1. Inicializar θ
   Elegir tasa de aprendizaje $\alpha > 0$
   Elegir número de épocas E
2. Para cada época $e = 1, \ldots, E$ repetir:
   2.1 Ordenar aleatoriamente los datos
   2.2 Para cada dato z_i en el orden realizado en el paso 2.1 repetir:
      2.2.1 Calcular la derivada $\ell_i'(\theta)$
      2.2.2 Actualizar el parámetro: $\theta \leftarrow \theta - \alpha\, \ell_i'(\theta)$
3. Devolver θ

Bibliografía

- Everitt, Brian. Introduction to optimization methods and their application in statistics.(2012) Springer.