Clase 4: Máxima verosimilitud

MDS7104 Aprendizaje de Máquinas

Felipe Tobar

Iniciativa de Datos e Inteligencia Artificial
Universidad de Chile

27 de marzo de 2022

UNIVERSIDAD
DE CHILE

Enfoque de máxima verosimilitud

- El criterio MC asume que ningún modelo es el modelo correcto y por lo tanto se busca un modelo *aproximado* a los datos tal que la discrepancia entre el modelo candidato y los datos sea mínima.
- El enfoque de máxima verosimilitud es radicalmente distinto ya que consiste encontrar el modelo que con mayor probabilidad ha generado *exactamente* los datos observados.
- Debido a la naturaleza aleatoria de los datos, para implementar este concepto es necesario considerar modelos probabilísticos, de forma de poder calcular la probabilidad de que los datos $\mathcal{D}$ hayan sido generados por un modelo en particular. De esta forma, se elegirá el modelo que maximice dicha probabilidad.

Para el caso del problema de regresión, se considerarán distintos modelos que relacionen la variable de salida como una variable aleatoria y a través de una distribución condicional (a la entrada x y el parámetro del modelo θ) de la forma

$$y|x, \theta \sim p(y|x, \theta),$$

donde enfatizamos que y es la única variable aleatoria y tanto el parámetro θ como la entrada x son cantidades fijas (el parámetro θ es desconocido y la entrada x es conocida u *observada*).

Independencia en el modelo de regresión

- Usualmente asumiremos que los datos $\{y_i\}_{i=1}^N$ generados a partir de las entradas $\{x_i\}_{i=1}^N$, son **condicionalmente independientes** dado el modelo: si conociésemos el modelo (i.e., si conociésemos θ), entonces para dos entradas x_i, x_j independientes, las salidas correspondientes y_i, y_j son independientes.
- Por otra parte, los valores generados por el modelo $\{y_i\}_{i=1}^N$ **no son independientes** ya que si lo fueran, la predicción de una observación nueva $y_\star$ en base a una secuencia de observaciones $\{y_i\}_{i=1}^N$ estaría dada por

$$p(y_\star|\{y_i\}_{i=1}^N) \overset{\text{(prob. cond.)}}{=} \frac{p(y_\star, \{y_i\}_{i=1}^N)}{p(\{y_i\}_{i=1}^N)} \overset{\text{(indep.)}}{=} \frac{p(y_\star), p(\{y_i\}_{i=1}^N)}{p(\{y_i\}_{i=1}^N)} = p(y_\star).$$

  Es decir, las observaciones pasadas $\{y_i\}_{i=1}^N$ no aportarían para la predicción. Por el contrario, como nuestro supuesto es de **independencia condicional** la expresión correcta es la siguiente:

$$p(y_\star|\{y_i\}_{i=1}^N, \theta) = p(y_\star|\theta),$$

  lo cual quiere decir que las observaciones pasadas no son útiles para predecir el futuro **solo si conozco el modelo**. Esto es evidente, pues si conozco el modelo, no necesito datos para saber de $y_\star$.

Caso particular: regresión lineal

En particular, en el caso de la regresión lineal podemos considerar el siguiente **modelo generativo**:

$$y = a^\top x + b + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2),$$

el cual consta de una parte determinística (afín en x) y una parte aleatoria caracterizada por la variable aleatoria ϵ.

El modelo probabilístico anterior puede expresarse mediante la siguiente densidad condicional

$$y|x \sim p(y|x, \theta) = \mathcal{N}(y; a^\top x + b, \sigma_\epsilon^2),$$

donde $\theta = (a, b, \sigma_\epsilon^2)^\top$ representa todos los parámetros del modelo.

- El supuesto de independencia condicional está garantizado al imponer que las realizaciones de $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$ sean iid.
- Es posible aprender el modelo desde múltiples observaciones, pues intuitivamente todas las observaciones aportan evidencia sobre el parámetro en común θ.
- Si, por el contrario, cada observación tuviese su propio parámetro, el aprendizaje no sería posible.

Nota: estamos assumiendo la existencia de un patrón subjacente

Función de verosimilitud

Sean $\{y_i\}_{i=1}^N$ observaciones generadas por un **modelo generativo** definido mediante la densidad de probabilidad $y \sim p(y|\theta)$, donde $\theta \in \Theta$ es el parámetro (desconocido) del modelo.
La dependencia de las observaciones cuando el modelo es desconocido permite introducir la siguiente definición:

Definition (verosimilitud)

Se define $L : \Theta \to [0, \infty)$ como la densidad de probabilidad de los datos observados condicional al parámetro θ, es decir,

$$\theta \mapsto L(\theta) := p(\{y_i\}_{i=1}^N|\theta).$$

Dicho valor de L se denomina *verosimilitud* del modelo $p(y|\theta)$ o, equivalentemente, del parámetro θ.

En algunos casos, consideraremos las notaciones $L_{\mathbf{y}}(\theta)$ o $L(\theta|\mathcal{D})$ para enfatizar que la verosimilitud es tomada con respecto a las observaciones $\mathbf{y}$ o al conjunto $\mathcal{D}$.

Es importante enfatizar que la función $L(\theta)$ **no es una densidad de probabilidad**, ya que si bien $p(\{y_i\}_{i=1}^N|\theta)$ integra 1 cuando se integra con respecto a los datos y, no necesariamente integra 1 cuando se integra con respecto a θ.

Ejemplo: verosimilitud de un modelo gaussiano

Consideremos un modelo gaussiano definido por

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

y un conjunto de observaciones $\mathbf{y} = \{y_i\}_{i=1}^N$ iid. La verosimilitud de $\theta = (\mu, \sigma^2)^\top$ está dada por:

$$\begin{aligned} L(\theta) = p(\mathbf{y}|\mu, \sigma^2) &\overset{\text{(iid)}}{=} \prod_{i=1}^N p(y_i|\mu, \sigma^2) = \prod_{i=1}^N \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(y_i-\mu)^2}{2\sigma^2}\right) \\ &= \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left(\frac{-\sum_{i=1}^N (y_i-\mu)^2}{2\sigma^2}\right) \\ &= \dots \\ &= \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left(\frac{-(\bar{s}-\bar{y}^2)}{2\sigma^2/N}\right) \exp\left(\frac{-(\mu-\bar{y})^2}{2\sigma^2/N}\right), \end{aligned}$$

donde $\bar{y} = \frac{1}{N}\sum_{i=1}^N y_i$ es el promedio de las observaciones y $\bar{s} = \sum_{i=1}^N y_i^2/N$ es el promedio de los cuadrados de las observaciones.

Elección del modelo de acuerdo a la máxima verosimilitud

Definida la verosilimitud, surge la pregunta natural de cómo elegir θ a partir de la función de verosimilitud. Si bien una respuesta natural es elegir el θ que maximice $L(\theta)$, esta elección tiene una justificación mucho más detallada.

Sea $D(p_1, p_2)$ alguna medida de discrepancia entre dos modelos p_1, p_2. Luego, se elegirá el $\hat{\theta}$ tal que $p(y|\hat{\theta})$ es lo más *cercano* posible al modelo real $p(y|\theta)$ con respecto a la discrepancia D, es decir, el que minimiza la expresión

$$D(p(y|\theta), p(y|\hat{\theta})).$$

Recordar: y es la variable, la notación alternativa es $D(p(\cdot|\theta), p(\cdot|\hat{\theta}))$

Desafortunadamente, notemos que formular y resolver este problema no es posible en el caso general, pues la expresión de arriba depende del parámetro real θ, el cual no conocemos, con lo que no podríamos resolver dicho problema de optimización.

Sin embargo, veamos que podemos considerar una métrica que ofrece una alternativa para optimizar la discrepancia entre el modelo real y el aproximado, independientemente de que no conozcamos el valor de θ.

Motivación de la máxima verosimilitud

La medida de discrepancia que se utilizará se deriva de la teoría de la información. Se define la divergencia de Kullback-Leibler entre el modelo real $p = p(y|\theta)$ y el aproximado $q = p(y|\hat{\theta})$ como

$$\mathrm{KL}(p(y|\theta), p(y|\hat{\theta})) := \int_y \log\left(\frac{p(y|\theta)}{p(y|\hat{\theta})}\right) p(y|\theta)\mathrm{d}y.$$

Si bien no es posible calcular dicha integral, observemos que esta es una esperanza con respecto a la densidad $p(y|\theta)$, por lo que se puede utilizar una aproximación de Monte Carlo usando las N observaciones en D:

$$\mathrm{KL}(p(y|\theta), p(y|\hat{\theta})) \approx \mathrm{KL}_N(p(y|\theta), p(y|\hat{\theta})) = \frac{1}{N}\sum_{i=1}^{N} \log\left(\frac{p(y_i|\theta)}{p(y_i|\hat{\theta})}\right),$$

donde las muestras $\forall i, y_i \sim p(y|\theta)$ iid.

Motivación de la máxima verosimilitud

Observemos que el minimizante de la expresión anterior $\hat{\theta}_N$ verifica:

$$\begin{aligned}
\hat{\theta}_N &= \arg\min_{\hat{\theta}} \sum_{i=1}^{N} \log\left(\frac{p(y_i|\theta)}{p(y_i|\hat{\theta})}\right) \\
&= \arg\min_{\hat{\theta}} \sum_{i=1}^{N} \log p(y_i|\theta) - \sum_{i=1}^{N} \log p(y_i|\hat{\theta}) \\
&= \arg\max_{\hat{\theta}} \sum_{i=1}^{N} \log p(y_i|\hat{\theta}) \\
&= \arg\max_{\hat{\theta}} \prod_{i=1}^{N} p(y_i|\hat{\theta}).
\end{aligned}$$

Es decir, si las muestras son condicionalmente independientes, entonces la expresión anterior implica que $\hat{\theta}_N$ es también el maximizante de la función de verosimilitud. En efecto,

$$\hat{\theta}_N = \arg\max_{\hat{\theta}} \prod_{i=1}^{N} p(y_i|\hat{\theta}) = \arg\max_{\hat{\theta}} p(\mathbf{y}|\hat{\theta}) = \arg\max_{\hat{\theta}} L_{\mathbf{y}}(\theta).$$

Por lo tanto, el parámetro que minimiza la KL corresponde al estimador de máxima verosimilitud (EMV) $\Rightarrow$ **aprendemos no solo parámetros sino que modelos.**

EMV para el modelo lineal gaussiano (1)

Para el modelo lineal con ruido gaussiano $y = a^\top x + b + \epsilon$

$$y|x \sim p(y|x, \theta) = \mathcal{N}(y; a^\top x + b, \sigma_\epsilon^2),$$

la verosimilitud viene dada por:

$$L_{\mathbf{y}}(\theta) = \prod_{i=1}^{N} \mathcal{N}(y_i; a^\top x_i + b, \sigma_\epsilon^2) = \frac{1}{(2\pi\sigma_\epsilon^2)^{N/2}} \exp\left(\frac{-\sum_{i=1}^{N}(y_i - a^\top x_i - b)^2}{2\sigma_\epsilon^2} \right).$$

Usualmente, se trabaja con el logaritmo de la verosimilitud, referido como *log-verosimilitud*, $l(\theta) = \log L(\theta)$, por su facilidad de interpretación y optimización.

De este modo, la log-verosimilitud del modelo lineal gaussiano está dada por

$$l(\theta) = \underbrace{-N \log \sqrt{2\pi\sigma_\epsilon^2}}_{\text{dispersión}} + \underbrace{\frac{-1}{2\sigma_\epsilon^2} \sum_{i=1}^{N}(y_i - a^\top x_i - b)^2}_{\text{ajuste}}.$$

EMV para el modelo lineal gaussiano (2)

En particular, el estimador de máxima verosimilitud para los parámetros de la parte lineal (i.e., ignorando σ_ϵ^2) está dado por:

$$[a^{\mathrm{MV}}, b^{\mathrm{MV}}] = \arg\min_{a,b} \sum_{i=1}^{N} (y_i - a^\top x_i - b)^2.$$

Observemos que es posible identificar esta última expresión como el costo de MC, es decir, el estimador de máxima verosimilitud es el minimizante del mismo costo que el estimador de MC. Consecuentemente, ambos estimadores son iguales y por lo tanto:

$$[\hat{a}, \hat{b}] = [a^{\mathrm{MV}}, b^{\mathrm{MV}}] = [a^{\mathrm{MC}}, b^{\mathrm{MC}}] = \left(\tilde{X}^\top \tilde{X}\right)^{-1} \tilde{X}^\top Y.$$

Además, recordemos que luego de determinar el estimador con criterio de MC, es posible calcular la varianza de los errores (error cuadrático medio) de nuestro modelo mediante

$$\text{Varianza} = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{a}^\top x_i - \hat{b})^2.$$

EMV para el modelo lineal gaussiano (3)

Por otra parte, en el contexto de máxima verosimilitud, la varianza es un parámetro del modelo por lo que puede ser calculado maximizando la log-verosimilitud al igual que para los parámetros a y b:

$$\sigma^2_{\mathrm{MV}} = \arg\max \left(-\frac{N}{2} \log(\sigma^2_\epsilon) + \frac{-1}{2\sigma^2_\epsilon} \sum_{i=1}^{N} (y_i - a^\top x_i - b)^2 \right).$$

Dado que ya se optimizó sobre los parámetros a y b, solo falta aplicar la condición de primer orden sobre σ^2_ϵ:

$$\frac{\partial l(\theta)}{\partial \sigma^2_\epsilon} = -\frac{N}{2\sigma^2_\epsilon} + \frac{1}{2(\sigma^2_{\mathrm{MV}})^2} \sum_{i=1}^{N} (y_i - \hat{a}^\top x_i - \hat{b})^2 = 0 \Rightarrow \sigma^2_\epsilon = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{a}^\top x_i - \hat{b})^2.$$

Con lo cual se obtiene la misma expresión de la varianza que al usar mínimos cuadrados.

Además, observemos que el estimador de MV de la varianza del ruido depende de los parámetros a y b, pero no al revés.