Probabilidad y estadística para la Inteligencia Artificial: Máxima verosimilitud y mínimos cuadrados

Camilo Enrique Argoty Pulido

Especialización en Inteligencia Artificial

16 de septiembre de 2025

Estadísticos y estimadores

Muestras aleatorias

Un conjunto X_1, X_2,..., X_n de v.a.'s independientes pero que se distribuyen igual (es decir, tienen la misma f.d.p.) se denomina **muestra aleatoria de tamaño** n sobre X, donde X es una v.a. que se distribuye como las variables que integran la muestra.

Muestras aleatorias

Un conjunto X_1, X_2,..., X_n de v.a.'s independientes pero que se distribuyen igual (es decir, tienen la misma f.d.p.) se denomina **muestra aleatoria de tamaño** n sobre X, donde X es una v.a. que se distribuye como las variables que integran la muestra.

Usualmente se entiende una muestra aleatoria como un conjunto de datos que se toman de mediciones en la realidad.

Muestras aleatorias

Un conjunto X_1, X_2,..., X_n de v.a.'s independientes pero que se distribuyen igual (es decir, tienen la misma f.d.p.) se denomina **muestra aleatoria de tamaño** n sobre X, donde X es una v.a. que se distribuye como las variables que integran la muestra.

Usualmente se entiende una muestra aleatoria como un conjunto de datos que se toman de mediciones en la realidad.

El sentido de la definición es que una muestra, más allá de ser un conjunto de datos, entiende dichos datos como los resultados de un mismo experimento realizado varias veces.

Estadísticos y estimadores

Un **estadístico** es una función calculada a partir de muestras aleatorias.

Estadísticos y estimadores

Un **estadístico** es una función calculada a partir de muestras aleatorias.

Una forma intuitiva de entender un estadístico, es verlo como una función que calcula cosas a partir de datos recopilados experimentalmente.

Estadísticos y estimadores

Un **estadístico** es una función calculada a partir de muestras aleatorias.

Una forma intuitiva de entender un estadístico, es verlo como una función que calcula cosas a partir de datos recopilados experimentalmente.

Un estadístico que se usa para aproximar un parámetro de una v.a. X (usando muestras sobre X) se denomina **estimador (puntual)** de θ.

Estadísticos y estimadores

Un **estadístico** es una función calculada a partir de muestras aleatorias.

Una forma intuitiva de entender un estadístico, es verlo como una función que calcula cosas a partir de datos recopilados experimentalmente.

Un estadístico que se usa para aproximar un parámetro de una v.a. X (usando muestras sobre X) se denomina **estimador (puntual)** de θ.

Si θ es un parámetro de una v.a. X, y $\tilde{\theta}(X_1, X_2, \dots, X_n)$ es un estimador para θ, el **sesgo** de $\tilde{\theta}$ es:

Estadísticos y estimadores

Un **estadístico** es una función calculada a partir de muestras aleatorias.

Una forma intuitiva de entender un estadístico, es verlo como una función que calcula cosas a partir de datos recopilados experimentalmente.

Un estadístico que se usa para aproximar un parámetro de una v.a. X (usando muestras sobre X) se denomina **estimador (puntual)** de θ.

Si θ es un parámetro de una v.a. X, y $\tilde{\theta}(X_1, X_2, \ldots, X_n)$ es un estimador para θ, el **sesgo** de $\tilde{\theta}$ es:

$$E[(\tilde{\theta} - \theta)] = E[\tilde{\theta}] - \theta$$

Estadísticos y estimadores

Un **estadístico** es una función calculada a partir de muestras aleatorias.

Una forma intuitiva de entender un estadístico, es verlo como una función que calcula cosas a partir de datos recopilados experimentalmente.

Un estadístico que se usa para aproximar un parámetro de una v.a. X (usando muestras sobre X) se denomina **estimador (puntual)** de θ.

Si θ es un parámetro de una v.a. X, y $\tilde{\theta}(X_1, X_2, \ldots, X_n)$ es un estimador para θ, el **sesgo** de $\tilde{\theta}$ es:

$$E[(\tilde{\theta} - \theta)] = E[\tilde{\theta}] - \theta$$

Mientras que el **error cuadrático medio** de $\tilde{\theta}$ es:

Estadísticos y estimadores

Un **estadístico** es una función calculada a partir de muestras aleatorias.

Una forma intuitiva de entender un estadístico, es verlo como una función que calcula cosas a partir de datos recopilados experimentalmente.

Un estadístico que se usa para aproximar un parámetro de una v.a. X (usando muestras sobre X) se denomina **estimador (puntual)** de θ.

Si θ es un parámetro de una v.a. X, y $\tilde{\theta}(X_1, X_2, \ldots, X_n)$ es un estimador para θ, el **sesgo** de $\tilde{\theta}$ es:

$$E[(\tilde{\theta} - \theta)] = E[\tilde{\theta}] - \theta$$

Mientras que el **error cuadrático medio** de $\tilde{\theta}$ es:

$$ECM(\tilde{\theta}) = E[(\tilde{\theta} - \theta)^2]$$

Estadísticos y estimadores

Un **estadístico** es una función calculada a partir de muestras aleatorias.

Una forma intuitiva de entender un estadístico, es verlo como una función que calcula cosas a partir de datos recopilados experimentalmente.

Un estadístico que se usa para aproximar un parámetro de una v.a. X (usando muestras sobre X) se denomina **estimador (puntual)** de θ.

Si θ es un parámetro de una v.a. X, y $\tilde{\theta}(X_1, X_2, \ldots, X_n)$ es un estimador para θ, el **sesgo** de $\tilde{\theta}$ es:

$$E[(\tilde{\theta} - \theta)] = E[\tilde{\theta}] - \theta$$

Mientras que el **error cuadrático medio** de $\tilde{\theta}$ es:

$$ECM(\tilde{\theta}) = E[(\tilde{\theta} - \theta)^2]$$

De aquí, si el sesgo de un estimador es 0, se dice que el estimador es **insesgado**, mientras que si minimiza el ECM, se dice de **mínimo error cuadrático medio**.

La media muestral

Sea X_1, X_2,...,X_n una muestra sobre X.

La media muestral

Sea X_1, X_2,...,X_n una muestra sobre X.

La **media muestral** de X_1, X_2,...,X_n es el estimador:

La media muestral

Sea X_1, X_2,...,X_n una muestra sobre X.

La **media muestral** de X_1, X_2,...,X_n es el estimador:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

La media muestral

Sea X_1, X_2,...,X_n una muestra sobre X.

La **media muestral** de X_1, X_2,...,X_n es el estimador:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Sea $\mu = E[X]$. Por la linealidad de la esperanza, se tiene que:

La media muestral

Sea X_1, X_2,...,X_n una muestra sobre X.

La **media muestral** de X_1, X_2,...,X_n es el estimador:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Sea $\mu = E[X]$. Por la linealidad de la esperanza, se tiene que:

$$E[\bar{X}] = E\Big[\frac{1}{n}\sum_{i=1}^{n} X_i\Big] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu$$

La media muestral

Sea X_1, X_2,...,X_n una muestra sobre X.

La **media muestral** de X_1, X_2,...,X_n es el estimador:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Sea $\mu = E[X]$. Por la linealidad de la esperanza, se tiene que:

$$E[\bar{X}] = E\Big[\frac{1}{n}\sum_{i=1}^{n} X_i\Big] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu$$

Por lo anterior, la media muestral es un estimador insesgado.

La varianza muestral

La media muestral es una v.a. en sí, la cual tiene su propia varianza. Por el **Teorema Central del límite** (que veremos más adelante):

La varianza muestral

La media muestral es una v.a. en sí, la cual tiene su propia varianza. Por el **Teorema Central del límite** (que veremos más adelante):

$$Var(\bar{X}) = \frac{\sigma^2}{n},$$

La varianza muestral

La media muestral es una v.a. en sí, la cual tiene su propia varianza. Por el **Teorema Central del límite** (que veremos más adelante):

$$Var(\bar{X}) = \frac{\sigma^2}{n},$$

donde σ^2 es $Var(X)$.

La varianza muestral

La media muestral es una v.a. en sí, la cual tiene su propia varianza. Por el **Teorema Central del límite** (que veremos más adelante):

$$Var(\bar{X}) = \frac{\sigma^2}{n},$$

donde σ^2 es $Var(X)$.

A primera vista, se pensaría que el mejor estimador para σ^2 debería ser:

La varianza muestral

La media muestral es una v.a. en sí, la cual tiene su propia varianza. Por el **Teorema Central del límite** (que veremos más adelante):

$$Var(\bar{X}) = \frac{\sigma^2}{n},$$

donde σ^2 es $Var(X)$.

A primera vista, se pensaría que el mejor estimador para σ^2 debería ser:

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

La varianza muestral

Sin embargo, cuando se calcula el sesgo se obtiene:

La varianza muestral

Sin embargo, cuando se calcula el sesgo se obtiene:

$$
\begin{aligned}
E\Big[\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2\Big] &= \frac{1}{n}\Big(\sum_{i=1}^{n}[(X_i-\mu)^2]-n(\bar{X}-\mu)^2\Big) = \\
&= \frac{1}{n}\Big(\sum_{i=1}^{n}E[(X_i-\mu)^2]-nE[(\bar{X}-\mu)^2]\Big) = \\
&= \frac{1}{n}(\sum_{i=1}^{n}\sigma^2-n\frac{\sigma^2}{n}) = \frac{(n-1)}{n}\sigma^2 \neq \sigma^2
\end{aligned}
$$

La varianza muestral

Por lo anterior, si se desea un estimador insesgado para la varianza, debe definirse la **varianza muestral** como:

La varianza muestral

Por lo anterior, si se desea un estimador insesgado para la varianza, debe definirse la **varianza muestral** como:

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

Estimación de parámetros por máxima verosimilitud

La función de verosimilitud

Sean X_1, X_2,..., X_n un conjunto de v.a.'s iid (independientes e idénticamente distribuidas), es decir una muestra aleatoria.

La función de verosimilitud

Sean X_1, X_2,..., X_n un conjunto de v.a.'s iid (independientes e idénticamente distribuidas), es decir una muestra aleatoria.

Supóngase que f.d.p la distribución de la muestra es $f(x, \theta)$, donde θ es un parámetro.

La función de verosimilitud

Sean X_1, X_2,..., X_n un conjunto de v.a.'s iid (independientes e idénticamente distribuidas), es decir una muestra aleatoria.

Supóngase que f.d.p la distribución de la muestra es $f(x, \theta)$, donde θ es un parámetro.

Si se toma θ como una variable y tendiendo en cuenta que las variables X_1, X_2,..., X_n son independientes, entonces la densidan conjunta de la muestra puede verse como una función sobre θ:

La función de verosimilitud

Sean X_1, X_2,..., X_n un conjunto de v.a.'s iid (independientes e idénticamente distribuidas), es decir una muestra aleatoria.

Supóngase que f.d.p la distribución de la muestra es $f(x, \theta)$, donde θ es un parámetro.

Si se toma θ como una variable y tendiendo en cuenta que las variables X_1, X_2,..., X_n son independientes, entonces la densidan conjunta de la muestra puede verse como una función sobre θ:

$$
\begin{aligned}
L(\theta) &= f_{X_1, X_2, \ldots, X_n}(x_1, x_2, \ldots, x_n; \theta) = \\
&= f_{x_1}(x_1, \theta) \cdot f_{x_2}(x_2, \theta) \cdots f_{x_n}(x_n, \theta) = \\
&= \Pi_{i=1}^{n} f_{x_i}(x_i, \theta)
\end{aligned}
$$

La función de verosimilitud

Sean X_1, X_2,..., X_n un conjunto de v.a.'s iid (independientes e idénticamente distribuidas), es decir una muestra aleatoria.

Supóngase que f.d.p la distribución de la muestra es $f(x, \theta)$, donde θ es un parámetro.

Si se toma θ como una variable y tendiendo en cuenta que las variables X_1, X_2,..., X_n son independientes, entonces la densidan conjunta de la muestra puede verse como una función sobre θ:

$$\begin{aligned} L(\theta) &= f_{X_1, X_2, \ldots, X_n}(x_1, x_2, \ldots, x_n; \theta) = \\ &= f_{x_1}(x_1, \theta) \cdot f_{x_2}(x_2, \theta) \cdots f_{x_n}(x_n, \theta) = \\ &= \Pi_{i=1}^{n} f_{x_i}(x_i, \theta) \end{aligned}$$

Esta función se denomina **función de verosimilitud** para θ.

Estimador de máxima verosimilitud

Como su nombre lo indica, la función de verosimilitud da cuenta de qué tan creíble es un posible valor del parámetro θ, para unos resultados experimentales dados.

Estimador de máxima verosimilitud

Como su nombre lo indica, la función de verosimilitud da cuenta de qué tan creíble es un posible valor del parámetro θ, para unos resultados experimentales dados.

Por lo tanto, si se busca un estimador, una muy buena alternativa es buscar el estimador que sea más creíble a la luz de los datos experimentales, esto es, un estimador que maximice dicha función de verosimilitud.

Estimador de máxima verosimilitud

Como su nombre lo indica, la función de verosimilitud da cuenta de qué tan creíble es un posible valor del parámetro θ, para unos resultados experimentales dados.

Por lo tanto, si se busca un estimador, una muy buena alternativa es buscar el estimador que sea más creíble a la luz de los datos experimentales, esto es, un estimador que maximice dicha función de verosimilitud.

Este estimador se denomina el **Estimador de Máxima Verosimilitud**:

Estimador de máxima verosimilitud

Como su nombre lo indica, la función de verosimilitud da cuenta de qué tan creíble es un posible valor del parámetro θ, para unos resultados experimentales dados.

Por lo tanto, si se busca un estimador, una muy buena alternativa es buscar el estimador que sea más creíble a la luz de los datos experimentales, esto es, un estimador que maximice dicha función de verosimilitud.

Este estimador se denomina el **Estimador de Máxima Verosimilitud**:

$$EMV_\theta(X_1, X_2, \ldots, X_n) = argmax_\theta(L(\theta))$$

La función LL (logaritmo de verosimilitud)

En la mayoría de los casos, maximizar $L(\theta)$ suele ser tedioso y/o difícil. Por lo tanto, suele tomarse el logaritmo de la función de verosimilitud el cual, dado que el logaritmo es creciente, nos va a llevar al mismo argumento máximo.

La función LL (logaritmo de verosimilitud)

En la mayoría de los casos, maximizar $L(\theta)$ suele ser tedioso y/o difícil. Por lo tanto, suele tomarse el logaritmo de la función de verosimilitud el cual, dado que el logaritmo es creciente, nos va a llevar al mismo argumento máximo.

Example

El tiempo de realización en minutos de una determinada tarea dentro de un proceso industrial es una variable aleatoria con función de densidad:

La función LL (logaritmo de verosimilitud)

En la mayoría de los casos, maximizar $L(\theta)$ suele ser tedioso y/o difícil. Por lo tanto, suele tomarse el logaritmo de la función de verosimilitud el cual, dado que el logaritmo es creciente, nos va a llevar al mismo argumento máximo.

Example

El tiempo de realización en minutos de una determinada tarea dentro de un proceso industrial es una variable aleatoria con función de densidad:

$$f(x) = \frac{x}{\theta^2} e^{-\frac{x}{\theta}} \text{ si } x > 0$$

La función LL (logaritmo de verosimilitud)

En la mayoría de los casos, maximizar $L(\theta)$ suele ser tedioso y/o difícil. Por lo tanto, suele tomarse el logaritmo de la función de verosimilitud el cual, dado que el logaritmo es creciente, nos va a llevar al mismo argumento máximo.

Example

El tiempo de realización en minutos de una determinada tarea dentro de un proceso industrial es una variable aleatoria con función de densidad:

$$f(x) = \frac{x}{\theta^2} e^{-\frac{x}{\theta}} \text{ si } x > 0$$

donde $\theta > 0$.

La función LL (logaritmo de verosimilitud)

En la mayoría de los casos, maximizar $L(\theta)$ suele ser tedioso y/o difícil. Por lo tanto, suele tomarse el logaritmo de la función de verosimilitud el cual, dado que el logaritmo es creciente, nos va a llevar al mismo argumento máximo.

Example

El tiempo de realización en minutos de una determinada tarea dentro de un proceso industrial es una variable aleatoria con función de densidad:

$$f(x) = \frac{x}{\theta^2} e^{-\frac{x}{\theta}} \text{ si } x > 0$$

donde $\theta > 0$.
Encontrar un EMV para θ.

Regresión y mínimos cuadrados

La esperanza condicional como regresión

Se entiende como **regresión** una función sobre una variable aleatoria X con la que se busca aproximar otra variable aleatoria Y.

La esperanza condicional como regresión

Se entiende como **regresión** una función sobre una variable aleatoria X con la que se busca aproximar otra variable aleatoria Y.

Si se busca una función $\tilde{Y}$ calculada en X que aproxime la v.a. Y y que minimice el ECM (Error cuadrático medio):

La esperanza condicional como regresión

Se entiende como **regresión** una función sobre una variable aleatoria X con la que se busca aproximar otra variable aleatoria Y.

Si se busca una función $\tilde{Y}$ calculada en X que aproxime la v.a. Y y que minimice el ECM (Error cuadrático medio):

$$E[(\tilde{Y} - Y)^2]$$

La esperanza condicional como regresión

Se entiende como **regresión** una función sobre una variable aleatoria X con la que se busca aproximar otra variable aleatoria Y.

Si se busca una función $\tilde{Y}$ calculada en X que aproxime la v.a. Y y que minimice el ECM (Error cuadrático medio):

$$E[(\tilde{Y} - Y)^2]$$

Si $\tilde{Y} = h(X)$, para alguna función (medible) h, Entonces,

La esperanza condicional como regresión

Se entiende como **regresión** una función sobre una variable aleatoria X con la que se busca aproximar otra variable aleatoria Y.

Si se busca una función $\tilde{Y}$ calculada en X que aproxime la v.a. Y y que minimice el ECM (Error cuadrático medio):

$$E[(\tilde{Y} - Y)^2]$$

Si $\tilde{Y} = h(X)$, para alguna función (medible) h, Entonces,

$$E[(h(X) - Y)^2] \leq E[(g(X) - Y)^2]$$

La esperanza condicional como regresión

Se entiende como **regresión** una función sobre una variable aleatoria X con la que se busca aproximar otra variable aleatoria Y.

Si se busca una función $\tilde{Y}$ calculada en X que aproxime la v.a. Y y que minimice el ECM (Error cuadrático medio):

$$E[(\tilde{Y} - Y)^2]$$

Si $\tilde{Y} = h(X)$, para alguna función (medible) h, Entonces,

$$E[(h(X) - Y)^2] \leq E[(g(X) - Y)^2]$$

para toda función medible g.

La esperanza condicional como regresión

Se entiende como **regresión** una función sobre una variable aleatoria X con la que se busca aproximar otra variable aleatoria Y.

Si se busca una función $\tilde{Y}$ calculada en X que aproxime la v.a. Y y que minimice el ECM (Error cuadrático medio):

$$E[(\tilde{Y} - Y)^2]$$

Si $\tilde{Y} = h(X)$, para alguna función (medible) h, Entonces,

$$E[(h(X) - Y)^2] \leq E[(g(X) - Y)^2]$$

para toda función medible g.
Esto ocurre cuando:

La esperanza condicional como regresión

Se entiende como **regresión** una función sobre una variable aleatoria X con la que se busca aproximar otra variable aleatoria Y.

Si se busca una función $\tilde{Y}$ calculada en X que aproxime la v.a. Y y que minimice el ECM (Error cuadrático medio):

$$E[(\tilde{Y} - Y)^2]$$

Si $\tilde{Y} = h(X)$, para alguna función (medible) h, Entonces,

$$E[(h(X) - Y)^2] \leq E[(g(X) - Y)^2]$$

para toda función medible g.
Esto ocurre cuando:

$$\tilde{Y} = E(Y|X = x)$$

donde

La esperanza condicional como regresión

Se entiende como **regresión** una función sobre una variable aleatoria X con la que se busca aproximar otra variable aleatoria Y.

Si se busca una función $\tilde{Y}$ calculada en X que aproxime la v.a. Y y que minimice el ECM (Error cuadrático medio):

$$E[(\tilde{Y} - Y)^2]$$

Si $\tilde{Y} = h(X)$, para alguna función (medible) h, Entonces,

$$E[(h(X) - Y)^2] \leq E[(g(X) - Y)^2]$$

para toda función medible g.
Esto ocurre cuando:

$$\tilde{Y} = E(Y|X = x)$$

donde

$$E(Y|X = x) = \int y f_{(Y|X)(y|x)dy}$$

La esperanza condicional como regresión

Se entiende como **regresión** una función sobre una variable aleatoria X con la que se busca aproximar otra variable aleatoria Y.

Si se busca una función $\tilde{Y}$ calculada en X que aproxime la v.a. Y y que minimice el ECM (Error cuadrático medio):

$$E[(\tilde{Y} - Y)^2]$$

Si $\tilde{Y} = h(X)$, para alguna función (medible) h, Entonces,

$$E[(h(X) - Y)^2] \leq E[(g(X) - Y)^2]$$

para toda función medible g.
Esto ocurre cuando:

$$\tilde{Y} = E(Y|X = x)$$

donde

$$E(Y|X = x) = \int y f_{(Y|X)(y|x) dy}$$

Así, la esperanza condicional, es la variable $h(X)$ para alguna función medible h que mejor aproxima a Y en el sentido de ECM.

Minimos cuadrados para regresión lineal

Si la función $h(x)$ que se busca es de la forma $h(x) = \beta_0 + \beta_1 x$, ya $h(X)$ no es $E(Y|X)$ salvo cuando X se distribuye normal.

Minimos cuadrados para regresión lineal

Si la función $h(x)$ que se busca es de la forma $h(x) = \beta_0 + \beta_1 x$, ya $h(X)$ no es $E(Y|X)$ salvo cuando X se distribuye normal.

De aquí,

$$\beta_1 = \frac{Cov(X,Y)}{Var(X)}$$

Minimos cuadrados para regresión lineal

Si la función $h(x)$ que se busca es de la forma $h(x) = \beta_0 + \beta_1 x$, ya $h(X)$ no es $E(Y|X)$ salvo cuando X se distribuye normal.

De aquí,

$$\beta_1 = \frac{Cov(X,Y)}{Var(X)}$$

y

Minimos cuadrados para regresión lineal

Si la función $h(x)$ que se busca es de la forma $h(x) = \beta_0 + \beta_1 x$, ya $h(X)$ no es $E(Y|X)$ salvo cuando X se distribuye normal.

De aquí,

$$\beta_1 = \frac{Cov(X,Y)}{Var(X)}$$

y

$$\beta_0 = E[X] - \beta_1 E[Y]$$