¿Qué es un MLE?

En la clase pasada hicimos Estimación de Densidad de Probabilidad usando métodos Paramétricos y No Paramétricos.

En general es muy importante este proceso (estimar la distribución de probabilidad de un conjunto de datos).

En esta clase vamos a trabajar uno de los Frameworks o esquemas de trabajo mas comunes a la hora de hacer este proceso.

MLE - Estimación de Maxima Verosimilitud

Es una técnica que nos permite estimar densidades de probabilidad dentro de un esquema muy general.

¿Cuales son los elementos esenciales?

- Escoger una distribución: Teniendo solo una muestra de los datos.
- Escoger los parámetros de la distribución: Los que mejor ajustan la distribución de datos. Estos parámetros me van a permitir ajustar mejor o peor los datos.

Hay un problema general de Estimación de Maxima Verosimilitud que es; los datos obedecen a una distribución de probabilidad de una población enorme, pero nosotros no vamos a tener conocimiento de esa población enorme. Así que en general la distribución de probabilidad de los datos que tenemos, es diferente de la distribución de probabilidad del problema general, donde nosotros podríamos entender cual es su forma si nosotros tuviéramos conocimiento de todos los datos posibles.

Entonces:

- Escoger la distribución de probabilidad sobre una muestra de datos, esa es la restricción primera.
- Luego vamos a usar los parámetros que tiene una distribución para ajustar la distribución a nuestros datos. Esto suena a problema de Machine Learning, porque hay variables que hay que calibrar o tunear para ajustar mis datos.

Básicamente el MLE es un problema de optimización

Porque nosotros formulamos el esquema de trabajo dela siguiente manera:

- Yo tengo un conjunto de datos X (X en general representa un dataset), donde pueden haber muchos datos.
- Por otro lado yo tengo los parámetros de la distribución que estoy buscando ajustar a mis datos θ.

Entonces hay una probabilidad que yo obtengo de ajustar los datos a la distribución en concreto, a esta distribución se le suele escribir con la letra L [Verosimilitud - Likelihood].

$$P(X, \theta) = L(X, \theta)$$

Una vez que tengo una Densidad de Probabilidad de dados ciertos parámetros y dados un conjunto de puntos o dataset, entonces ¿como resuelvo el problema de optimización?

Como hay muchos conjuntos de parámetros que me van a permitir ajustar un conjunto de datos con diferentes grados de probabilidad, yo lo que voy a hacer es: de todas esas posibles combinaciones, escoger aquella cuya probabilidad es la maxima posible.

Entonces el máximo de la función L de verosimilitud dados los parámetros θ.

$$max L(X, \theta) \rightarrow max \prod_i \log P(X_i, \theta)$$

Otra hipótesis importante es que a veces esa distribución de probabilidad sobre el dataset se puede factorizar como el producto de varias probabilidades, donde cada probabilidad concierne a solo un data point X_i de todo el conjunto de datos. Entonces esta linea describe un producto de probabilidades.

Y aquí se presenta un problema común de siempre factorizar como un producto de varias probabilidades y es que como en general las probabilidades son números pequeños, cuando yo multiplico alas probabilidades, el numero que va resultando se hace todavía mas pequeño.

Computacionalmente esto no es bueno porque cuando trabajamos con maquinas de precision finita, como las PC's de hoy en dia: existe una precision limite debajo de la cual las maquinas ya no pueden calcular, entonces a eso se le llama **underflow**.

Cuando hacemos este tipo de problemas, trabajando con este tipo de filosofía el **underflow** puede ser muy común, así que la manera en que esto se resuelve es **Aplicando el logaritmo de las probabilidades** en lugar de las probabilidades en si mismas, porque al hacer el logaritmo, existe una propiedad de que el logaritmo de un producto es la suma de los logaritmos.

Entonces esto convierte el problema de multiplicaciones a sumas. Los Logaritmos convierten números pequeños, en números relativamente grandes y negativos, que son mas fáciles de manejar desde un punto de vista numérico.

> Por esto en un problema de Maxima Verosimilitud lo que hacemos es:
>
> $$max \log L(X, \theta) = max \sum_i \log P(X_i, \theta)$$
>
> Que es igual a calcular el máximo de sumar los logaritmos de las probabilidades individuales, donde cada probabilidad individual corresponde con un datapoint, dados los parámetros y se suma todo eso.

Este es el problema general de como se formula este esquema de trabajo para hallar la densidad de probabilidad que mejor ajusta a un cierto conjunto de datos.

Lo que haremos en la proxima clase, es aterrizar este esquema de trabajo a un caso muy particular que es la regresión lineal, que es uno de los ejemplos mas sencillos de aprendizaje de maquina. Este es uno de los ejemplos canónicos.

EXtras:

- MLE Youtube
- Probability concepts explained
- Maximum Likelihood for the normal distribution