MLE en MAchine Learning

Haremos un ejemplo y particular.

¿Cómo se usa el MLE en Machine Learning? En general el Machine Learning consiste en ajustar densidades a datos. Desde un punto de vista probabilístico, todo se resume a esa frase, ya sea que tengamos problemas supervisados como: clasificación, regresión o problemas no supervisados: clusterización.

Regresión lineal con MLE

$$y = \underbrace{m}_{\text{pendiente}} x + \overbrace{b}^{\text{intercepto}} = \underbrace{b_0}_{\text{weight}} x + \overbrace{b_1}^{bias}$$

$$P(y|x) \rightarrow \max \sum_i \log P(y_i|x_i; \underbrace{h}_{\text{modelo}})$$

Vemos que es la ecuación de una recta esta en la anterior diapositiva, y la regresión lineal consiste en que tengo un conjunto de datos sobre los cuales yo tengo una intuición de que eso debería seguir un modelo lineal (Osea una recta). Es decir una variable $y = mx + b$, en las escuelas se enseña de este modo, sabemos que m es la pendiente y b es la intersección.

Luego cuando uno ve esto en ciencia de datos, nos damos cuenta que se escribe como $y = b_0x + b_1$, donde antes b_0 era la **pendiente** y aquí se llama **peso (weight)** y b_1 era la **intersección** y ahora es **bias**.

Entonces es lo mismo, solo decimos que hay estas similitudes en la matemática convencional y el Machine Learning. Dado un problema yo debo encontrar la ecuación de la recta que mejor ajuste a los datos, aplicando MLE significa que haremos:

$$P(y|x) \rightarrow max \sum_i \log P(y_i|x_i; h)$$

Fíjate que hay notación de probabilidad condicional. La probabilidad de que dados unos datos x tenga el resultado y, esto lo traduzco como el máximo de la suma de los logaritmos de las probabilidades de cada pareja de puntos (x, y) de mi conjunto de datos y esto es calculado en cada una de las probabilidades suponiendo h. ¿Quién es h? Es el modelo que yo voy a ajustar.

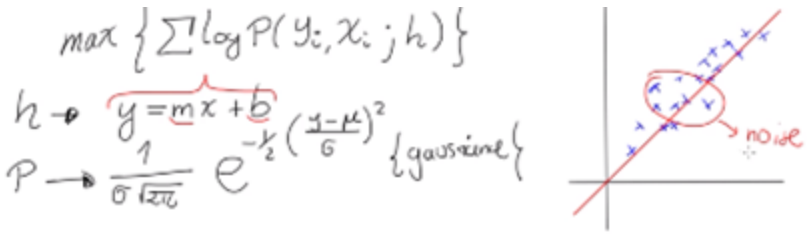
En general como MLE sirve para mas cosas h no necesariamente tiene que ser un modelo lineal, pero en este caso h va a ser nuestra hipótesis de que el modelo que voy a buscar ajustar es justamente una **linea recta**.

$$P(y|x) \rightarrow \max \sum_i \log P(y_i|x_i; \underbrace{h}_{\text{modelo}})$$

$$h \rightarrow y = b_0x + b_1$$

$$P \rightarrow \frac{1}{\sigma\sqrt{2\pi}} \exp -\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2$$

Tenemos un conjunto de datos que al cual yo quiero ajustar la recta en color rojo, nosotros ya intuimos que la recta debe tener la ecuación $y = mx + b$ y lo que busco es determinar los coeficientes que mejor ajustan con los datos



Los datos tienen una desviación con respecto de la recta, a esa desviación la conocemos como ruido **noise**. La hipótesis fundamental es que las desviaciones o ruido respecto de la recta (roja) $y = mx + b$ siguen una forma de distribución Gaussiana. Por eso nuestra hipótesis de modelamiento probabilístico es que la probabilidad va a seguir una forma Gaussiana.

$$P \to \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2\right]$$

Entonces ya determinamos aquí el primer punto de MLE que escogimos nuestra distribución P sera una Gaussiana.

$$\text{MLE} \to \text{dist P} \to \text{Gaussiana}$$

El esquema tradicional mediante el cual uno aprende el problema de la **Regresión Lineal** es con los mínimos cuadrados.

¿Que quiere decir mínimos cuadrados? Cuando tenemos el dato real, sabemos que en hay un punto (x_i, y_i) y que cuando tomamos el valor x_i y le aplicamos la función teórica (la ecuación de la recta roja) es decir; $y = mx_i + b$ el y resultante es diferente de y_i (Aquí los datos reales).

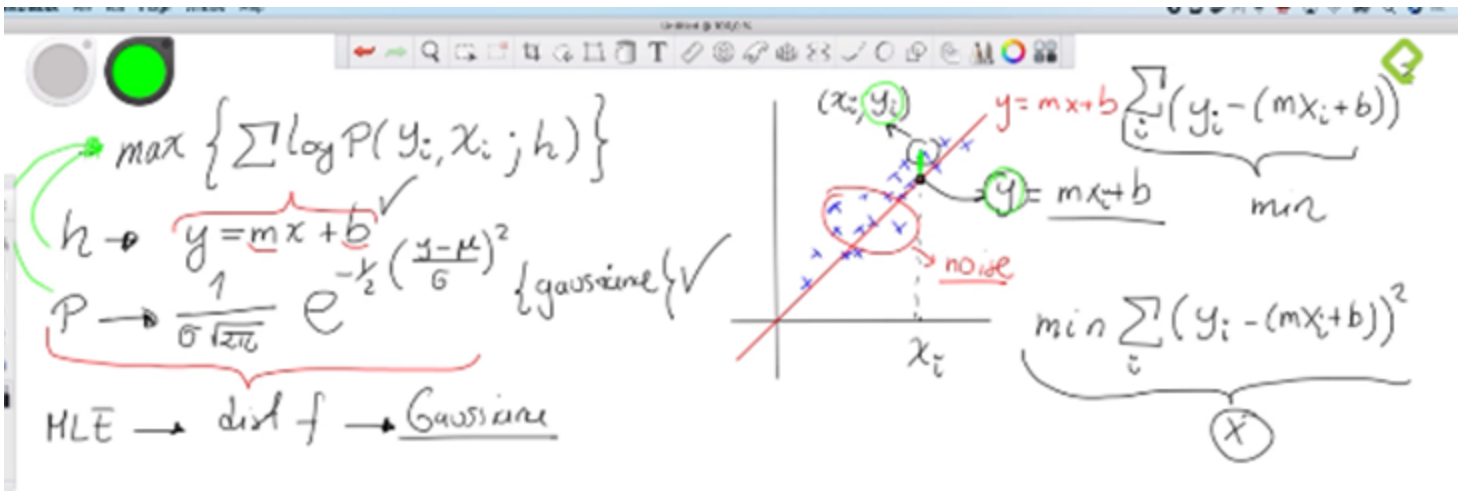
Estos valores son diferentes y eso determina un valor de error para cada punto de mi conjunto y error se calcula como

$$\text{error} = y_i - (mx_i + b)$$

Como esta diferencia a veces da positiva y aveces negativa, entonces estos errores individuales se suelen elevar al cuadrado $(y_i - (mx_i + b))^2$ y luego se suman todos los errores $\sum_i (y_i - (mx_i + b))^2$, en general lo que se busca es minimizar la suma.

En un esquema tradicional lo que se busca con el problema de regresión lineal es minimizar es decir; encontrar el mínimo de la suma de todos esos errores cuadráticos, en resumen eso se conoce como el método de mínimos cuadrados. Y claro, lo mas usual es entender la regresión lineal desde el punto de vista de los mínimos cuadrados.

Lo que veremos aquí es que MLE es totalmente equivalente a lo que se ve con la regresión lineal y el método de mínimos cuadrados, para entender que también es un problema probabilístico con ese esquema.



¿Cómo funciona?

Nosotros tenemos nuestra hipótesis de probabilidad de que el ruido de los datos y una distribución Gaussiana y una hipótesis de modelamiento de que el modelo debe ser un modelo lineal $y = mx + b$. Entonces ¿cómo se combinan ambas cosas?

Estos 2 entes h y P van a entrar a la ecuación de estimación de maxima verosimilitud $\max\{\sum \ln P(Y_i, X_i; h)\}$.

Resulta que dentro de nuestra función de probabilidad $\frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2\right]$ yo tengo la variable que implica al ruido X y el promedio μ o media donde la tendencia tiene que ser la tendencia de la recta que yo estoy buscando calcular es decir que la resta entre $X - \mu$ es totalmente equivalente a la resta de $(y_i - (mx_i + b))$.

Entonces matemáticamente esto ¿como va a resultar?

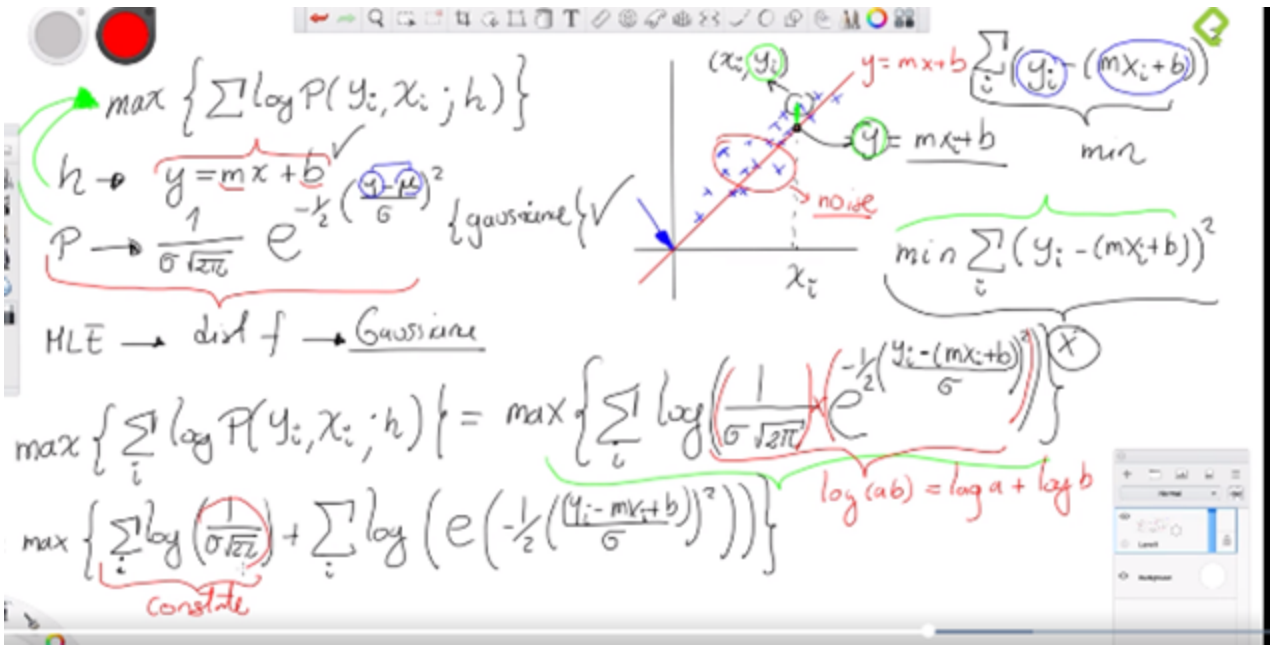
Entonces el $\max\{\sum \ln P(Y_i, X_i; h)\}$ se va a reducir a calcular

$$\max\left\{\sum \ln P\left[\frac{1}{\sigma\sqrt{2\pi}}\exp^{-\frac{1}{2}\left(\frac{Y_i-(mX_i+b)}{\sigma}\right)^2}\right]\right\}$$

Hasta aquí queda formulado el problema y parece no haber una relación entre la ecuación anterior y $\min\ \sum_i \left(Y_i-(mX_i+b)\right)^2$.

Si es similar, veamos $\log(ab) = \log a + \log b$. ¿Cómo se convierte esto?

$$\max\left\{\sum_i \log\left(\frac{1}{\sigma\sqrt{2\pi}}\right)+\sum_i \log\left(e^{-\frac{1}{2}\left(\frac{Y_i-(mX_i+b)}{\sigma}\right)^2}\right)\right\}$$



Analizando

Cuando yo calculo el máximo, el termino de la izquierda es constante, no hace la diferencia realmente ¿porque? porque $\sigma, \pi, 2$ son constantes. Entonces podemos despreciar el termino porque no va a generar una diferencia significativa. Por lo que se puede reducir a calcular solamente el máximo del segundo termino:

$$\max\left\{\sum_i \log\left(e^{-\frac{1}{2}\left(\frac{Y_i-(mX_i+b)}{\sigma}\right)^2}\right)\right\}$$

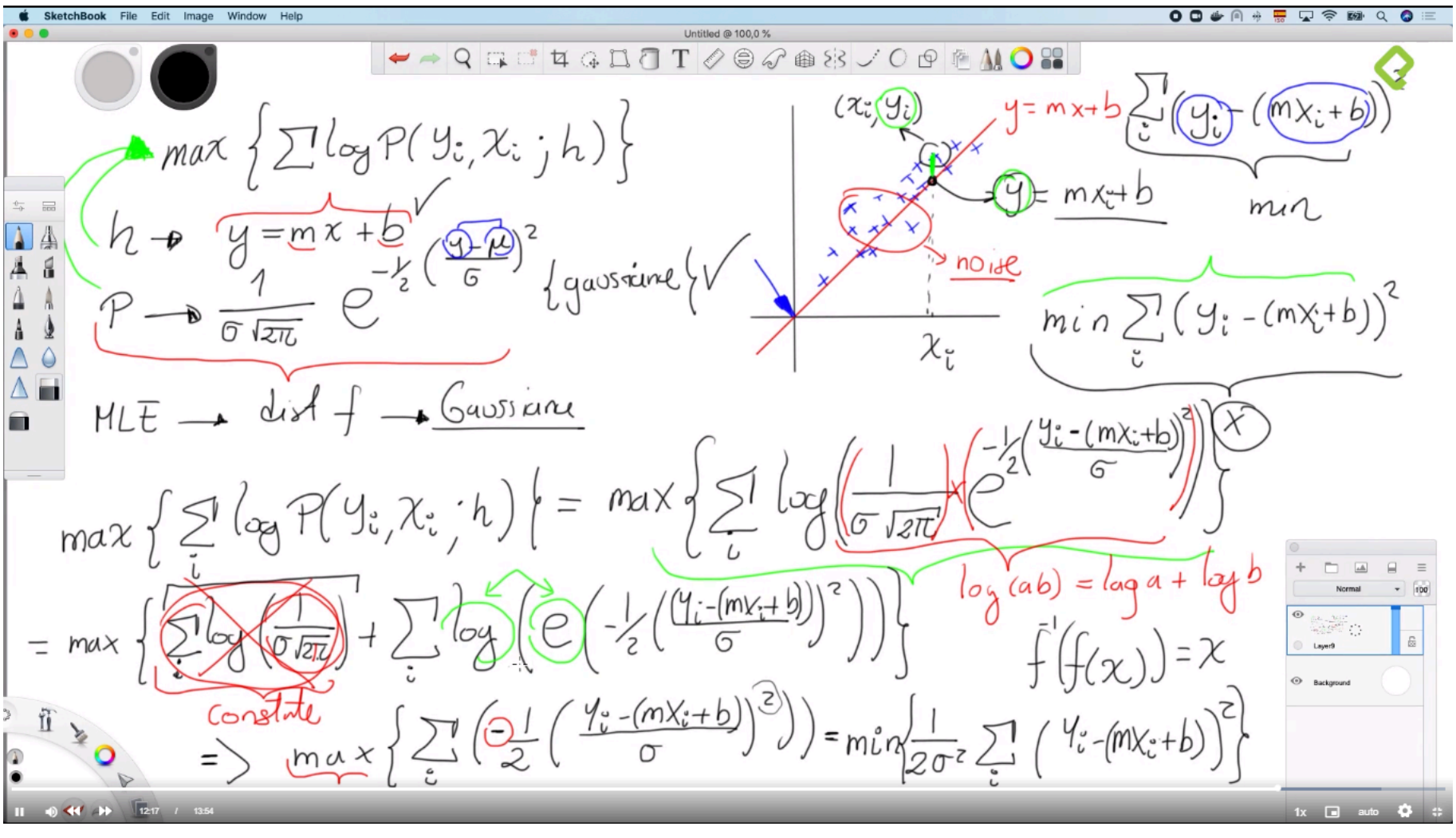
* OJO: log es igual al logaritmo natural

Entonces hay que observar que la función e y log son complementarias y al realizar la operación te devuelven el mismo valor $f^{-1}\left(f(x)\right) = x$, es decir las operaciones son inversas. Entonces esto resulta en lo siguiente:

$$\max\left\{\sum_i \left(-\frac{1}{2}\left(\frac{Y_i-(mX_i+b)}{\sigma}\right)^2\right)\right\}$$

Ahora hay otra situación y es el signo $-$ que antecede a $\frac{1}{2}$, el signo convierte el $\max$ en $\min$ si yo lo quisiera quitar, entonces quedaría como; el máximo de esos números negativos es equivalente al mínimo de esos mismos números pero en valor absoluto y luego las constantes las puedo sacar de la sumatoria y quedaría como:

$$\min\left\{\frac{1}{2\sigma^2}\sum_i \left(Y_i-(mX_i+b)\right)^2\right\}$$



Oh sorpresa nos encontramos con que el MLE también consiste en calcular el $mínimo$ de la **suma de los errores cuadráticos**, es decir la misma expresión de **Mínimos Cuadrados**

MLE

$$\min \left\{ \frac{1}{2\sigma^2} \sum_i \left(Y_i - (mX_i + b)\right)^2 \right\}$$

Mínimos cuadrados

$$\min \left\{ \sum_i (y_i - (mx_i + b))^2 \right\}$$

Esto quiere decir que también el **método de mínimos cuadrados** es en realidad también una **Estimación de Maxima Verosimilitud** y aquí esta demostrado hasta cierto punto y con cierto formalismo que son el mismo método.

Resumen

Así pues hemos aplicado un razonamiento matemático para entender que el problema de regresión lineal es también **la estimación de una densidad de probabilidad** sobre un conjunto de datos y por lo tanto también es un problema eminentemente probabilístico donde teníamos una hipótesis de maxima verosimilitud, sobre un logaritmo de distribuciones de probabilidad que estaban condicionadas por una hipótesis de modelamiento h que correspondía con la ecuación de una recta $mx + b$ y esto sujeto a que la distribución del error de los datos entre los datos reales y la linea recta a la cual debo ajustar sigue una distribución Gaussiana o distribución normal P.

El resultado nos indica que cuando vemos este problema como un error de minimizacion de de mínimos cuadrados es totalmente equivalente, y entonces es como entendemos un ejemplo donde la probabilidad juega un papel muy importante en un problema de Machine Learning puntual.

Extras:

- MLE
- MLE 2