Bayes en Machine Learning

El teorema de Bayes es muy util para saber como cuantificar la incertidumbre en situaciones donde hay conocimiento previo.

Por supuesto que el Teorema de Bayes no se queda afuera de la ciencia de datos. Los algoritmos utilizan el teorema de Bayes para procesos de optimización y aprendizaje.

MAP vs MLE

$$P(\overbrace{h}^{\text{hipótesis}} | \underbrace{D}_{\text{datos}}) = \frac{P(D|h)P(h)}{P(D)}$$

$$\max P(h|D) \rightarrow \max P(D|h)P(h)$$

Como sabemos en el area frecuentista trabajamos con el MLE o Maximum Likelihood Estimation [Estimación de Maxima Verosimilitud]. Que es un esquema basado en las probabilidades basado en el punto de vista frecuentista.

Existe un análogo para el area Bayesiana y se llama MAP (Maximum Aposteriori) - Maximizar la Probabilidad Posterior.

¿En que consiste?

Dado un conjunto de datos **D** ¿cual es la probabilidad de que yo conociendo el conjunto de datos, tenga una hipótesis que describa al conjunto de datos del modelo **h**? Por el teorema de Bayes sabemos que: La verosimilitud es la probabilidad de **D** dado la hipótesis **h**, por la probabilidad del modelo **h** entre la probabilidad de la evidencia (Probabilidad del conjunto de datos **D**).

En el caso del MLE, nosotros nos preocupábamos por optimizar la verosimilitud que es el análogo de $P(D|h)$. Ahora en el esquema Bayesiana nosotros optimizamos la Posteriori $P(h|D)$ por eso buscamos:

$$\max P(h|D) \rightarrow P(D|h)P(h)$$

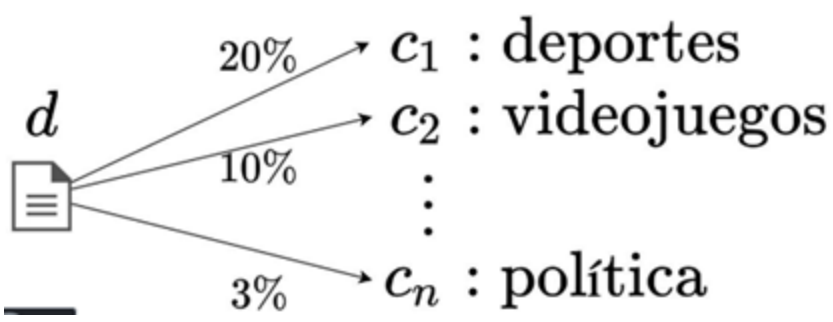
Y eso se traduce en el máximo del producto de 2 probabilidades y omitiendo la probabilidad de la evidencia $P(D)$.

¿Por que despreciamos la evidencia? Porque sin importar cual es el modelo que estamos usando la evidencia siempre va ser la misma y para reducir el problema de calcular 3 probabilidades a solo 2, se acostumbra a hacer el $máximo$ del producto de las probabilidades **Probabilidad prior x Verosimilitud**, y todos los problemas de optimización Bayesiana se reducen a eso. Aunque hay que poner atención a la hipótesis del modelamiento h y a los parámetros que posee.

Todo lo que diga Bayesiano en Machine Learning esta basado en este esquema, donde lo que hacemos **es hallar un máximo de las probabilidades posteriores**. Que se traduce a obtener parámetros a partir de los datos y no obtener los datos a partir de parámetros, entonces son como esquemas inversos pero ambos se usan para optimización.

¿Entonces aquí también entran en juego los problemas de clasificación? Si si entran. El clasificador más sencillo que se puede usar a partir de Bayes es el Naive-Bayes. La palabra "Naive o ingenuo" proviene de una hipótesis de **simplificación**.

Clasificador óptimo de Bayes



Un problema de clasificación como el que se menciono en clase pasadas, consiste en tener un dataset de documentos. Donde cada documentos puede pertenecer a una categoría:

- deportes
- juegos
- política

- etc.

Un clasificador probabilístico lo que hace es asignar una probabilidad de que el documento pertenezca a una de esas categorías. Donde luego se escoge la mayor de las probabilidades y decimos que ese documento pertenece a la categoría con mayor probabilidad.

Estos clasificadores también pueden estar basados en una filosofía Bayesiana.

¿Cómo seria?

Calcular la probabilidad

$$\max P(h|D) = \max P(D|h) \times P(h)$$

Lo que se traduce en calcular la probabilidad de diferentes hipótesis, es decir diferentes tipos de modelos h.

Lo que no se ve en la formulación es que las probabilidades son algo complicadas de calcular porque el **Espacio Muestral** es muy grande, sin embargo si no asumo ninguna simplificación sobre esas probabilidades decimos que tengo un **Modelo de Clasificación de Bayes Optimo - Optimal Bayes Classifier**.

En la realidad eso no es fácil de hacer porque el computo que se requiere es muy elevado, así que estamos limitados por eso.

El segundo paso que se suele hacer dada esa complicación es; que tenemos que factorizar la **Similitud** en producto de probabilidades mas sencillas, es decir ¿cual es la probabilidad de que ciertos datos sean descritos por una hipótesis de modelamiento?.

$$P(D|h) = P(\{x_1, x_2, x_3, \ldots, x_n\}|h)$$

Es decir de cada data point x_n

Yo voy a decir que esas probabilidades se pueden escribir como probabilidades independientes, es decir se puede factorizar como un producto de probabilidades completamente independientes.

$$P(D|h) = P(x_1|h)P(x_2|h)\ldots.$$

Cuando yo hago eso, las probabilidades que resultan son mucho mas sencillas porque yo lo que hago es iterar diferentes modelos sobre cada punto y eso reduce de manera considerable el espacio muestral, a esta hipótesis de simplificación es a la que llamamos **Naive Bayes**

Es así como Bayes tiene una participación en un modelo de clasificación. Pero no solamente sirven para clasificación sino que tienen otro tipo de aplicación.

Gracias a esto podemos ahondar en cursos posteriores sobre los tipos de algoritmos Bayesianos de clasificación, regresión lineal.

Este curso da las bases para tener un entendimiento y conocimiento mas amplio.