DigitalHouse > Coding School

DATA SCIENCE

UNIDAD 3
MÓDULO 4

Naïve Bayes

Octubre 2017

OBJETIVOS DE LA CLASE

1. **Entender el funcionamiento del algoritmo Naive Bayes a bajo nivel**

2. **Entender en qué casos se utiliza: como Benchmark y como clasificador eficiente**

3. **Implementar el modelo y evaluar la performance**

Naive Bayes

- Es una familia de clasificadores simples basados en la aplicación del Teorema de Bayes.

- Se basa en la idea de combinar la información de distintas variables independendientes que tampoco se interfieren entre sí.

- Podemos ver un problema de clasificación de la siguiente forma, donde L son las labels y features es la matriz de fatures-:

$$P(L \mid \text{features}) = \frac{P(\text{features} \mid L)P(L)}{P(\text{features})}$$

- Si queremos utilizar Naive Bayes para comparar la probabilidad de que un caso pertenezca a una u otra clase, podemos plantear:

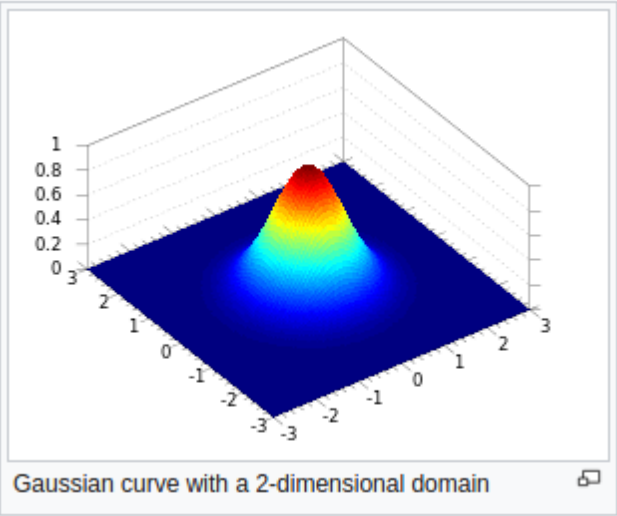
$$\frac{P(L_1 \mid \text{features})}{P(L_2 \mid \text{features})} = \frac{P(\text{features} \mid L_1)}{P(\text{features} \mid L_2)} \frac{P(L_1)}{P(L_2)}$$

- ¿Cómo calculamos la probabilidad de ocurrencia de los features para cada clase? $P(\text{features} \mid L_i)$

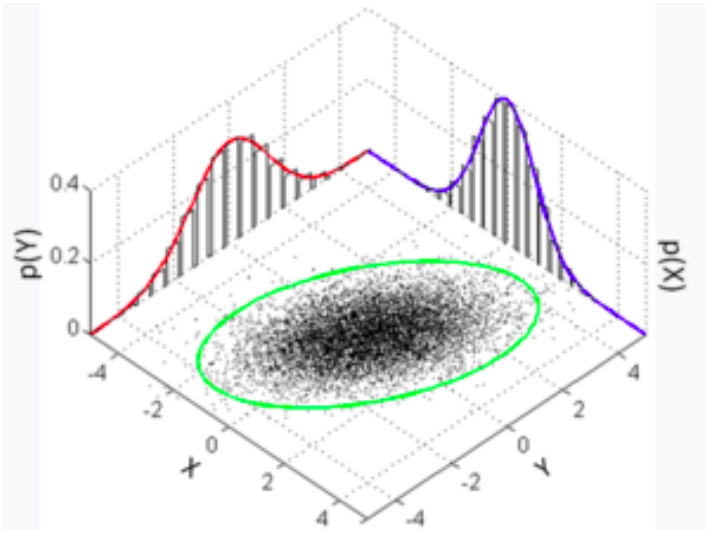
- Tenemos que suponer un "**modelo generativo**", es decir, un determinado proceso generador de datos cuyos parámetros vamos a estimar en función de los datos de entrenamiento.

¿Qué podríamos asumir acerca de los datos que nos ayude a modelar el proceso que los generó?

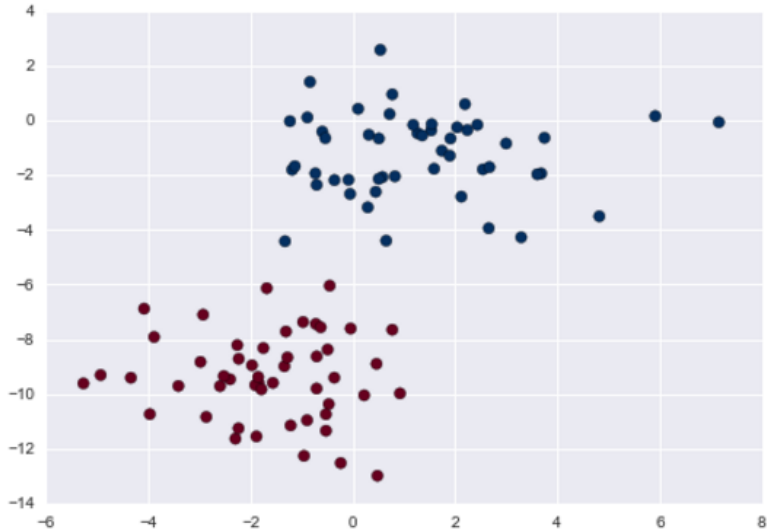
Uno de los supuestos más sencillos que se puede hacer, es que cada una de las clases proviene de un proceso Gaussiano multidimensional de generación de datos.

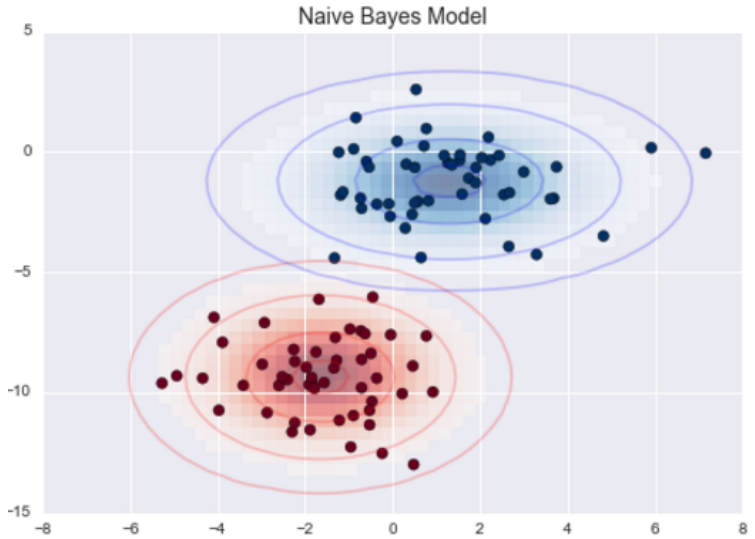


Gaussian curve with a 2-dimensional domain



Un modelo de tipo “Gaussian Naive Bayes” toma datos como los que se ven abajo y calcula una distribución para cada una de las clases. Lo bueno, es que obtenemos para cada nuevo punto que querramos clasificar un valor de probabilidad de cada una de las clases.





En Naive Bayes supusimos que cada par de features es totalmente independiente.
El teorema de Bayes indica:

$$P(y \mid x_1, \ldots, x_n) = \frac{P(y)P(x_1, \ldots x_n \mid y)}{P(x_1, \ldots, x_n)}$$

Y necesitamos asumir independencia de la siguiente forma:

$$P(x_i|y, x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) = P(x_i|y),$$

Para poder resolver el problema de esta forma:

$$P(y \mid x_1, \ldots, x_n) = \frac{P(y) \prod_{i=1}^{n} P(x_i \mid y)}{P(x_1, \ldots, x_n)}$$

$$P(y \mid x_1, \ldots, x_n) \propto P(y) \prod_{i=1}^{n} P(x_i \mid y)$$

$$\Downarrow$$

$$\hat{y} = \arg\max_y P(y) \prod_{i=1}^{n} P(x_i \mid y),$$

La solución del problema, dados estos supuestos es elegir la clase que tiene la mayor probabilidad de ocurrencia dado que observamos cada uno de los features INDIVIDUALES.

Los distintos algoritmos de Naive Bayes difieren en la distribución que supo $P(x_i \mid y)$.

- Gaussian Naive Bayes: Supone **distribución Gaussiana** multidimensional
- Naive Bayes Multinomial: Supone **distribución multinomial**.

Los clasificadores basados en Naive Bayes hacen fuertes supuestos sobre los datos, así que en general no van a tener tan buena performance como otros modelos de clasificación más complejos.

Dicho esto, tienen las siguientes ventajas:

- Son algoritmos muy rápidos tanto para entrenar como para predecir
- Brindan una predicción probabilística (tenemos probabilidades para cada clase)
- Son sencillos de interpretar
- No requieren “tunear” ningún hiperparámetro

Dado que Naive Bayes es tan fácil de optimizar y tan rápido desde el punto de vista computacional, es un buen "**baseline**" para un problema de clasificación.

Si performa bien, podemos quedarnos con este modelo y si necesitamos mejorar la precisión, tenemos una línea de base sobre la cual mejorar.

Naive Bayes tiende a funcionar bien en las siguientes situaciones:

- ***Cuando se cumplen los supuestos*** (cosa que rara vez pasa en casos reales)
- ***Cuando las clases están muy bien separadas*** y no es necesaria tanta complejidad en el modelo
- ***Cuando tenemos datos con muy alta dimensionalidad*** (por

Práctica Guiada NB

Lab NB