Subject: Probability and Statistics

Class XIV: Clasification

UdeA

Bioengineering

Francisco José Campuzano Cardona

Bioengineerer, MSc in Engineering

Clasificación

UdeA

Como dijimos antes, la clasificación es similar a la regresión, en tanto los predictores son variables numéricas, pero las etiquetas no, en este caso son categóricas.

Las categorías también se conocen como clases, y los problemas pueden ser de ***dos clases***, es decir binarios, o ***multiclase***, en donde hay más de dos posibles categorías dentro de las cuales clasificar las observaciones.

En el caso de dos clases, lo común es asignar a cada clase los valores de 0 y 1. En el caso de los problemas multiclase podría ser una clasificación ordinal, 1, 2, 3...

UdeA

Alternativa a los problemas multiclase

Otra forma de abordar los problemas multiclase, es convertirlos en varios problemas de dos clases. Por ejemplo: Suponga que en el término de un contrato de suscripción a algún servicio, el usuario podría decidir: no continuar el contrato, prorrogarlo mes a mes, o prorrogarlo por otro año. Entonces podría abordarse como un problema multiclase (tres clases) o dos problemas binarios.

En el segundo caso, se tendría un primer problema donde las opciones sería, ¿renueva el contrato, o no lo hace?. Y el segundo problema sería, dado que renovó el contrato, ¿lo hará por meses o lo hará a un año?
Esta estrategia es sobre todo útil, si una de las clases es más común.

UdeA

Algoritmo Naives Bayes

(*Naives* = Ingenuo)

Este es un algoritmo de clasificación bayesiano, en donde se hace uso de la probabilidad condicional.

El término Naives hace referencia a que hace algunas suposiciones “ingenuas” sobre los datos, que prácticamente nunca son reales, pero funcionan, la suposición principal es que los clasificadores, no están relacionados. Pero veamos como funcionaría un algoritmo no naives, para entender como funciona.

UdeA

Un algoritmo No Naives

1. Se tiene una base de datos con etiquetas.

2. Se tienen las características de otro *record* que se quiere clasificar. Entonces se busca en la base de datos, la combinación exacta de características para ver qué etiqueta tiene.

3. Se asigna la etiqueta al nuevo *record.*

Pero esto no es práctico, porque no siempre es posible encontrar exactamente la misma combinación de características del nuevo *record*

UdeA

Por ejemplo, se tiene la base de datos siguiente, y se desea asignar una etiqueta al *record*: (1,1,0) Y = ?

X1	X2	X3	Y
1	0	0	**5**
0	1	0	**5**
0	0	1	**3**
1	1	0	**3**
0	1	1	**5**
1	0	1	**3**
1	1	1	**5**

UdeA

Por ejemplo, si en un problema dado se quiere clasificar según ciertas características demográficas si una persona va a votar por alguien en una elección específica, por más grande que sea la base de datos es probable que no se encuentre un caso exactamente igual al siguiente: Hombre, Caucásico, de ingresos económicos altos, vive en el pacífico Colombiano, votó en las últimas elecciones, pero no en las penúltimas, tiene 3 hermanas, un hijo, está divorciado….

Entonces este enfoque no es muy práctico, por esto se recurre mejor a la probabilidad condicional, y se asume independencia entre las características.

UdeA

Entonces no se busca una clasificación exacta, sino probable, así para una clase L en un problema de clasificación, según el Teorema de Bayes se tendría:

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

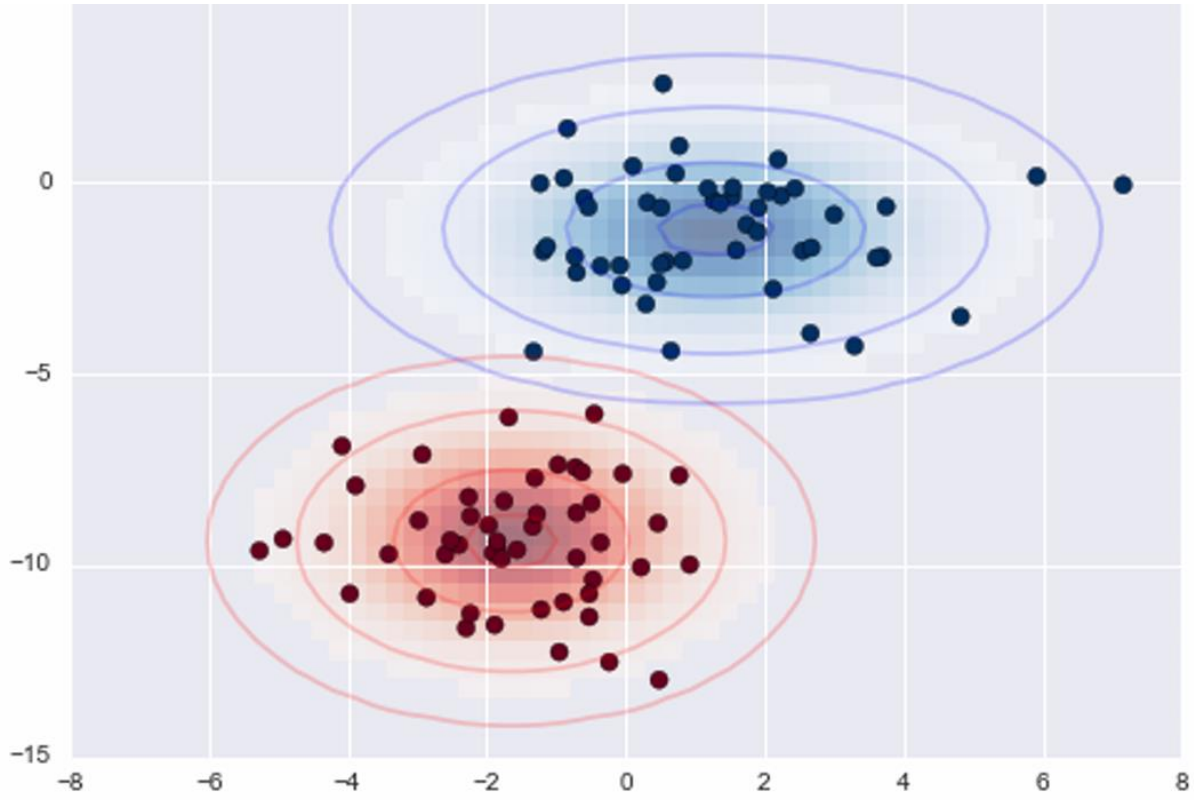
Sin embargo, el cómputo de esta ecuación, no es directo, se requiere un modelo que se conoce como ***Modelo generativo***, para poder determinar un proceso aleatorio que genere esos datos. Pero esto es muy complejo, entonces entran las suposiciones ingenuas.

UdeA

Gaussian Naive Bayes

En este caso se asume que los datos de cada etiqueta provienen de una distribución gausiana, en general multivariada, y que no hay covarianza entre sus *features.*

En este modelo, basta con tener la media y la desviación estándar de cada etiqueta, para poder reconstruir la distribución.



UdeA

Gaussian Naive Bayes

Luego de ajustado el modelo de clasificación, para un punto adicional, se determina la probabilidad de pertenecer a una clase, y se asigna la clase más probable.

Aquí los *features* son variables continuas !!

Veamos cómo se hace esto en sklearn ...

UdeA

Multinomial Naives Bayes

Aquí la suposición es que los *features* provienen de una distribución multinomial.

Una distribución multinomial, es una generalización de una distribución binomial, para más de 2 categorías.

Recordemos que en la distribución binomial, la variable aleatoria es la cantidad de veces que se obtiene fallo o éxito. En la distribución multinomial, se podría tener más de 2 opciones, y la variable sería la cantidad de veces que ocurre cada una de las opciones.

UdeA

Multinomial Naives Bayes

Entonces aquí los datos no se modelan según la mejor distribución gausiana, sino según la mejor distribución multinomial.

En este caso los *features* son variables discretas, específicamente contables, es decir, son el número de veces que ocurre algo.

Por ejemplo: se podría querer clasificar correos entre Spam y No Spam, según la frecuencia de aparición de ciertas palabras como: Urgente, Gratis, Oferta...

Subject: Probability and Statistics

UdeA

Bioengineering

¡Thanks!

Francisco José Campuzano Cardona

Bioengineering. MSc in Engineering