

# **DIPLOMATURA EN MACHINE LEARNING CON PYTHON**

## **Bayes Ingenuo**

## Bayes Ingenuo

- Instalación y uso del paquete
- Ejemplo conceptual en Excel
- Ejemplo en Python
- Principales parámetros de ajuste y control
- Problema concreto

## Clasificador basado en bayes ingenuo

### Un poco de contenido teórico:

#### Descripción del problema:

Necesitamos clasificar en unas pocas categorías a una serie de casos a partir de un conjunto de atributos.

Conocemos los atributos y la clase a la que pertenecen los  $N$  casos de una muestra. Queremos extender ese conocimiento al resto del universo para el cual NO tenemos la clase a la cual pertenece cada caso. Por lo tanto se trata de una técnica de aprendizaje supervisado.

Vamos a poder asumir sin pérdida de generalidad que nos interesa clasificar entre dos opciones pues si quisiéramos más nos bastaría con seguir subdividiendo las clases ya obtenidas.

Antes de plantear la matemática, siempre clara pero, a veces, hostil, vamos a tratar de entender la situación “con los dedos”

Asumamos que tenemos dos colaboradores que se turnan para hacer el trabajo de soporte, Juan y Pedro.

Juan viene 3 días por semana y los otros dos le toca venir a Pedro. Ellos ajustan los turnos de manera que siempre venga alguien. En principio no tenemos idea de quién va a venir.

Juan tiene la costumbre de tardar en atender. Sabemos que si el teléfono Juan tarda cuatro rings en atender el teléfono un 80% de las veces. Pedro, por otra parte, se desespera por atender rápido y, cuando atiende el sólo un 10% de las veces el teléfono suena 4 rings.

Vamos a distinguir entonces dos situaciones:

- Las probabilidades pre-existentes, que nos hablan de lo conocido, lo que está en el pasado.
- Las probabilidades de los eventos desconocidos, que están en el futuro, que son las que nos interesa predecir.

Nuestro planteo tiene que servirnos para calcular lo desconocido (y su incertidumbre) a partir de lo conocido.

Esto significa que estos atributos agregan poca o ninguna información a la tarea predictiva que queremos realizar pero consumen tanto tiempo de cálculo como los demás.

Un paso necesario es entonces ver qué pares de atributos están correlacionados entre sí para quedarnos con un conjunto de atributos no correlacionados, y así optimizar el modelo.

Esto es especialmente importante si el volumen de datos con el que vamos a trabajar es grande.

### **Selección de muestras:**

Una vez que conseguimos un conjunto independiente de atributos, puede convenirnos dividir el conjunto de datos de trabajo. Si resulta muy grande y la consideración del total pudiera ser muy lenta para el equipamiento que podemos utilizar siempre es posible tomar una muestra aleatoria de los datos y trabajar sobre la misma.

A veces conviene ensayar los procedimientos a utilizar con una muestra aleatoria en la que podamos trabajar más rápido hasta hacernos una idea cabal de lo que podemos obtener y luego utilizar muestras mayores.

Como veremos más adelante, siempre conviene dejar una parte de los datos sin tocar para validar sobre ellos el poder predictivo que hemos adquirido.

Usualmente, se trabaja con dos subconjuntos de los datos originales: el set de entrenamiento (*training*) y el set de validación (*testing*). A veces conviene considerar un subconjunto más, el de tuneo (*tuning*).

El subconjunto de entrenamiento representa, en general, cerca de  $2/3$  del set original, y es el conjunto de datos que se usa para entrenar al modelo (el modelo 'aprende' a partir de este set, relacionando la variable objetivo con las variables predictoras).

Con este set se construye el modelo, que luego es aplicado sobre el conjunto de validación para predecir los valores de la variable objetivo en este último set.

Estas predicciones se comparan con el valor conocido (del set de validación), lo que permite validar el modelo. El set de tuning se usa para tener una corroboración independiente y poder probar distintos valores de los parámetros cercanos a los parámetros óptimos encontrados en la fase de entrenamiento. De esta manera es posible optimizar aún más el modelo, o sea, 'hacer el tuneo fino'.

**Descripción del algoritmo básico de Bayes Ingenuo:**

Lo llamamos Bayes ingenuo no porque Bayes haya sido ningún ingenuo, fue un investigador muy sagaz de la estadística, sino porque incorpora una hipótesis que no es sencillo demostrar y que aceptamos entonces en forma bastante ingenua.

Como siempre vamos a considerar lo que tenemos y lo que queremos conseguir. Tenemos un universo de casos. Cada caso está descrito por un conjunto de atributos.

Además de los atributos, que nos son conocidos para todos los casos del universo tenemos un resultado, de carácter binario que sólo nos es conocido sobre una muestra aleatoria de ese universo:

Muestra Aleatoria		Resto Universo
Entrenamiento	Prueba	
Atributos conocidos	Atributos conocidos	Atributos Conocidos
Resultado conocido	Resultado conocido	Resultado Desconocido

Supongamos el siguiente ejemplo:

Queremos predecir si nuestros compañeros de secundaria aceptarán jugar un partido de fútbol o no. Como ya somos bastante veteranos no nos animamos a jugar en cualquier condición climática.

Vamos a tratar de inferir una regla que nos permita calcular la probabilidad de que juguemos a partir de la historia de las ocasiones anteriores:

Cielo	Temperatura	Humedad	Viento	¿Se jugó?
Soleado	Alta	Alta	No	No
Soleado	Alta	Alta	Si	No
Cubierto	Alta	Alta	No	Si
Lluvioso	Media	Alta	No	Si
Lluvioso	Fría	Media	No	Si
Lluvioso	Fría	Media	Si	No
Cubierto	Fría	Media	Si	Si
Soleado	Media	Alta	No	No
Soleado	Fría	Media	No	Si
Lluvioso	Media	Media	No	Si
Soleado	Media	Media	Si	Si
Cubierto	Media	Alta	Si	Si
Cubierto	Alta	Media	No	Si

Lluvioso	Media	Alta	Si	No
----------	-------	------	----	----

Vemos que, en total tenemos 14 casos. De los cuales se jugaron en 9 y no se jugaron en 5.

$$P(\text{jugar}) = 9/14$$

$$P(\text{no jugar}) = 5/14$$

También podemos calcular las probabilidades condicionales, por ejemplo

$$P(\text{Cielo} = \text{Soleado} / \text{jugar} = \text{si}) = 2/9$$

$$P(\text{Cielo} = \text{Soleado} / \text{jugar} = \text{no}) = 3/5$$

¿Por qué?

Porque contamos en la tabla que de los 9 casos que se jugó hubo dos casos soleados y el resto no lo fue.

Y si contamos de los cinco casos en los que no se jugó vemos que sólo en tres casos estaba soleado.

Análogamente calculamos:

$$P(\text{Temperatura} = \text{Fría} / \text{jugar} = \text{si}) = 3/9$$

$$P(\text{Temperatura} = \text{Fría} / \text{jugar} = \text{no}) = 1/5$$

$$P(\text{Humedad} = \text{alta} / \text{Jugar} = \text{si}) = 3/9$$

$$P(\text{Humedad} = \text{alta} / \text{Jugar} = \text{no}) = 4/5$$

$$P(\text{Viento} = \text{si} / \text{Jugar} = \text{si}) = 3/9$$

$$P(\text{Viento} = \text{si} / \text{Jugar} = \text{no}) = 3/5$$

Ahora vamos a asumir INGENUAMENTE que la temperatura, la humedad y el viento son independientes y lo usamos para calcular la probabilidad de que se dé el caso de que se cumplan las cuatro condiciones juntas tanto para jugar como para no jugar. Al asumir que la temperatura, la condición del cielo, la humedad y el viento obtenemos esa probabilidad conjunta simplemente multiplicando las probabilidades de cada situación:

	Jugar		No Jugar	
Total	9		5	
Temperatura Fría	3	3/9	1	1/5
Humedad Alta	3	3/9	4	4/5
Viento Si	3	3/9	3	3/5
Cielo Soleado	2	2/9	3	3/5
Temperatura fría y humedad alta y		$3/9 * 3/9 * 3/9 * 2/9 = 0.0082$		$1/5 * 4/5 * 3/5 * 3/5 = 0.0576$

viento y soleado				
------------------	--	--	--	--

Luego multiplicamos esa probabilidad que calculamos por la cantidad de casos en los que se jugó y no se jugó respectivamente:

	Jugar		No Jugar	
Total	9		5	
Temperatura Fría	3	3/9	1	1/5
Humedad Alta	3	3/9	4	4/5
Viento Si	3	3/9	3	3/5
Cielo Soleado	2	2/9	3	3/5
Temperatura fría y humedad alta y viento y soleado	$9 * .0082 = .074$	$3/9 * 3/9 * 3/9 * 2/9 = 0.0082$	$5 * .0576 = .288$	$1/5 * 4/5 * 3/5 * 3/5 = 0.0576$

Y ahora ya estamos listos para el truco final:

Si se cumplen las cuatro circunstancias climáticas jugaremos 0.074 veces y no jugaremos 0.288

O sea que la probabilidad de jugar será  $0.074/(0.074+.288) = .205$  y la de no jugar .795

Si repetimos este trabajo (con paciencia de computadora) sobre todas las combinaciones distintas de las cuatro circunstancias climáticas podemos calcular en cada caso la probabilidad de jugar y de no jugar.

Para una aproximación más intuitiva se sugiere ver el video:

<https://www.youtube.com/watch?v=ayQglkLE36I>

Pero, para hacer eso, mejor usar un algoritmo que lo tenga resuelto. Así que ahora que ya entendimos lo que está pasando vamos por el algoritmo implementado en Python.

### Construyendo el código Python – Instalación y uso de paquetes

Para la aplicación en Python se crean el clasificador, el set de entrenamiento. Luego el clasificador usando el set de entrenamiento realiza la predicción.

El set de entrenamiento (X) contiene altura, peso y tamaño de calzado. Y contiene etiquetas asociadas a masculino o femenino.

- Sklearn
- GaussianNB

Creamos el siguiente script en Sublime

```
#Paquete a utilizar
from sklearn.naive_bayes import GaussianNB

# create naive bayes classifier
gaunb = GaussianNB()

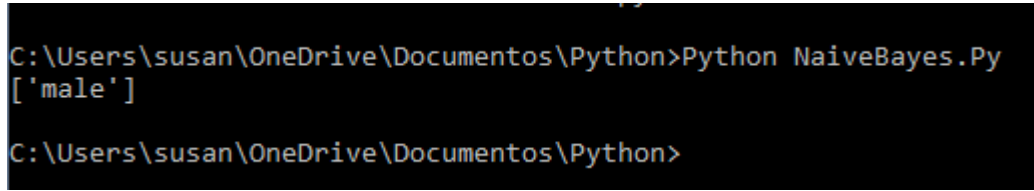
# create dataset
X = [[121, 80, 44], [180, 70, 43], [166, 60, 38], [153, 54, 37], [166, 65, 40], [190, 90, 47],
      [175, 64, 39],
      [174, 71, 40], [159, 52, 37], [171, 76, 42], [183, 85, 43]]

Y = ['male', 'male', 'female', 'female', 'male', 'male', 'female', 'female', 'female', 'male',
      'male']

# train classifier with dataset
gaunb = gaunb.fit(X, Y)

# predict using classifier
prediction = gaunb.predict([[190, 70, 43]])
print(prediction)
```

Invocamos al Script de la línea de comandos y obtenemos el resultado



```
C:\Users\susan\OneDrive\Documentos\Python>Python NaiveBayes.Py
['male']

C:\Users\susan\OneDrive\Documentos\Python>
```