

Entregable Naive Bayes

Laura Rodríguez Navas
rodrigueznabas@posgrado.uimp.es

Febrero 2020

Contador	Tamaño	Órbita	Temperatura	Habitable
170	Grande	Cercana	M	Sí
20	Grande	Cercana	A	Sí
45	Pequeño	Lejana	M	Sí
139	Pequeño	Cercana	M	Sí
30	Grande	Lejana	B	No
130	Grande	Cercana	B	No
255	Pequeño	Lejana	B	No
11	Pequeño	Cercana	A	No

Tomamos la siguiente base de datos relativa a la clasificación de planetas extrasolares como habitables o no. Para ello disponemos de tres variables predictivas: Tamaño, Órbita y Temperatura. La variable clase (Habitable) es binaria sí, no. Nótese que la columna “Contador” no es una variable predictiva, si no que nos sirve para mostrar de forma “comprimida” la base de datos, indicándose cuantas instancias existen para cada una de las configuraciones mostradas. Así, podemos ver que nuestra base de datos tiene 800 instancias, y no 8 como correspondería al número de filas de la tabla.

1. Construye un clasificador Naive Bayes para la tabla suministrada. Debes detallar todas las tablas de probabilidad.

Para este entregable, primero he preparado un fichero de datos CSV (“dataset.csv”), a partir de la tabla proporcionada, usando el lenguaje de programación Python, donde el fichero CSV resultante contiene las 800 instancias.

A continuación se puede observar el código utilizado,

```

import csv
import pandas as pd

data = {
    "Contador": [170, 20, 45, 139, 30, 130, 255, 11],
    "Tamaño": ["Grande", "Grande", "Pequeño", "Pequeño", "Grande",
               "Grande", "Pequeño", "Pequeño"],
    "Órbita": ["Cercana", "Cercana", "Lejana", "Cercana", "Lejana",
               "Cercana", "Lejana", "Cercana"],
    "Temperatura": ["M", "A", "M", "M", "B", "B", "B", "A"],
    "Habitable": ["Si", "Si", "Si", "Si", "No", "No", "No", "No"]
}

df = pd.DataFrame(data, columns=list(data.keys()))
columnnames = df.columns.tolist()
contador = data["Contador"]

with open("dataset.csv", "w") as fd:
    csv.writer(fd).writerow(columnnames)  # add column names
    for n in range(len(contador)):
        for i in range(contador[n]):
            df.loc[[n]].to_csv(fd, index=False, header=False, mode="a")

```

Después he transformado el conjunto de datos para eliminar la variable Contador, ya que no es una variable predictiva.

A continuación se puede observar el código añadido al anterior,

```

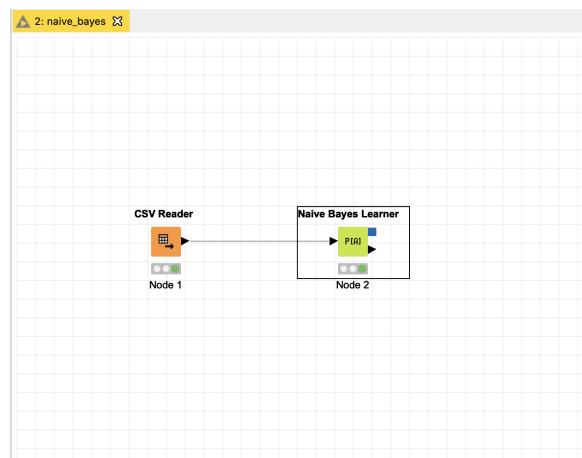
df = pd.read_csv('dataset.csv')
df = df.drop(["Contador"], axis=1)
df.to_csv('dataset.csv', index=False)

```

Finalmente, he utilizado el software KNIME para la construcción del clasificador Naive Bayes para los datos y también para detallar todas las tablas de probabilidad.

Concretamente, el clasificador crea un modelo bayesiano a partir de los datos. Calcula el número de filas por valor de atributo por clase para atributos nominales y la distribución gaussiana para atributos numéricos.

Si usamos Naive Bayes Learner,



Por ejemplo, podemos obtener obtener:

Naive Bayes Learner View - 2:2 - Naive Bayes Learner			
File			
Class counts for Habitable			
Class:	No	Si	
Count:	426	374	
Total count: 800			
Threshold to used for zero probabilities: 1.0E-4			
Missing values are ignored during learning and prediction phase.			
P(Tamaño class=?)			
Class/Tamaño	Grande	Pequeño	
No	160	266	
Si	190	184	
Rate:	44%	56%	
P(Temperatura class=?)			
Class/Temperatura	A	B	M
No	11	415	0
Si	20	0	354
Rate:	4%	52%	44%
P(Órbita class=?)			
Class/Órbita	Cercana	Lejana	
No	141	285	
Si	329	45	
Rate:	59%	41%	

Tablas de probabilidad

Suavizando mediante la corrección de Laplace,

$$P(\text{Habitable}=\text{Sí}) = (374+1)/(800+2) = 0.467$$

$$P(\text{Habitable}=\text{No}) = (426+1)/(800+2) = 0.532$$

P(Habitable)	Habitable = Sí	Habitable = No
	0.467	0.532

$$P(\text{Tamaño}=\text{Grande}|\text{Habitable}=\text{Sí}) = (190+1)/(800+2) = 0.238$$

$$P(\text{Tamaño}=\text{Grande}|\text{Habitable}=\text{No}) = (160+1)/(800+2) = 0.201$$

$$P(\text{Tamaño}=\text{Pequeña}|\text{Habitable}=\text{Sí}) = (184+1)/(800+2) = 0.231$$

$$P(\text{Tamaño}=\text{Pequeña}|\text{Habitable}=\text{No}) = (266+1)/(800+2) = 0.333$$

P(Tamaño Habitable)	Habitable = Sí	Habitable = No
Tamaño = Grande	0.238	0.201
Tamaño = Pequeña	0.231	0.333

$$P(\text{Órbita}=\text{Cercana}|\text{Habitable}=\text{Sí}) = (329+1)/(800+2) = 0.411$$

$$P(\text{Órbita}=\text{Cercana}|\text{Habitable}=\text{No}) = (141+1)/(800+2) = 0.177$$

$$P(\text{Órbita}=\text{Lejana}|\text{Habitable}=\text{Sí}) = (45+1)/(800+2) = 0.057$$

$$P(\text{Órbita}=\text{Lejana}|\text{Habitable}=\text{No}) = (285+1)/(800+2) = 0.357$$

P(Órbita Habitable)	Habitable = Sí	Habitable = No
Órbita = Cercana	0.411	0.177
Órbita = Lejana	0.057	0.357

$$P(\text{Temperatura}=\text{A}|\text{Habitable}=\text{Sí}) = (20+1)/(800+2) = 0.026$$

$$P(\text{Temperatura}=\text{A}|\text{Habitable}=\text{No}) = (11+1)/(800+2) = 0.015$$

$$P(\text{Temperatura}=\text{B}|\text{Habitable}=\text{Sí}) = (0+1)/(800+2) = 1.25 \times 10^{-3}$$

$$P(\text{Temperatura}=\text{B}|\text{Habitable}=\text{No}) = (415+1)/(800+2) = 0.519$$

$$P(\text{Temperatura}=\text{M}|\text{Habitable}=\text{Sí}) = (354+1)/(800+2) = 0.445$$

$$P(\text{Temperatura}=\text{M}|\text{Habitable}=\text{No}) = (0+1)/(800+2) = 1.25 \times 10^{-3}$$

P(Temperatura Habitable)	Habitable = Sí	Habitable = No
Temperatura = A	0.026	0.015
Temperatura = B	1.25×10^{-3}	0.519
Temperatura = M	0.445	1.25×10^{-3}

2. ¿Cuántos parámetros necesita tu clasificador? ¿Cuántos necesitaría para especificar la distribución de probabilidad conjunta (DPC)?

El clasificador necesita las tres variables predictivas: Tamaño, Órbita y Temperatura. Y la variable clase Habitable, que es binaria (Sí, No).

3. Clasifica los siguientes registros y muestra la distribución de probabilidad resultante:

1. (Tamaño=Grande, Órbita=Cercana, Temperatura=M)

Se estima para las dos etiquetas de la variable clase Habitable (Sí, No):

$$P(\text{Sí}) * P(\text{Tamaño}=\text{Grande}|\text{Sí}) * P(\text{Órbita}=\text{Cercana}|\text{Sí}) * P(\text{Temperatura}=\text{M}|\text{Sí}) = 0.467 * 0.238 * 0.411 * 0.445 = 0.020$$

$$P(\text{No}) * P(\text{Tamaño}=\text{Grande}|\text{No}) * P(\text{Órbita}=\text{Cercana}|\text{No}) * P(\text{Temperatura}=\text{M}|\text{No}) = 0.532 * 0.201 * 0.177 * 1.25 \times 10^{-3} = 2.36 \times 10^{-5}$$

$$\text{Así, } P(\text{Tamaño}=\text{Grande, Órbita}=\text{Cercana, Temperatura}=\text{M}) = 0.020 + 2.36 \times 10^{-5} = 0.0200236$$

2. (Tamaño=Grande, Órbita=Lejana, Temperatura=A)

Se estima para las dos etiquetas de la variable clase Habitable (Sí, No):

$$P(\text{Sí}) * P(\text{Tamaño=Grande} \mid \text{Sí}) * P(\text{Órbita=Lejana} \mid \text{Sí}) * P(\text{Temperatura=A} \mid \text{Sí}) = \\ 0.467 * 0.238 * 0.057 * 0.026 = 1.65 \times 10^{-4}$$

$$P(\text{No}) * P(\text{Tamaño=Grande} \mid \text{No}) * P(\text{Órbita=Lejana} \mid \text{No}) * P(\text{Temperatura=A} \mid \text{No}) = \\ 0.532 * 0.201 * 0.357 * 0.015 = 5.73 \times 10^{-4}$$

$$\text{Así, } P(\text{Tamaño=Grande, Órbita=Lejana, Temperatura=A}) = 1.65 \times 10^{-4} + 5.73 \times 10^{-4} = \\ 7.38 \times 10^{-4}$$

4. Si usáramos la DPC como clasificador y para todo registro obtuviéramos el mismo resultado que usando el clasificador Naive Bayes, ¿qué conclusión obtendríamos?