机器学习课程实验七

2022 年 11 月 3 日　苏博南　202000460020

考虑这样一个数据集，输入为：

1. Pregnancies: 怀孕次数
2. Glucose: 口服葡萄糖耐量试验中 2 小时的血糖浓度
3. BloodPressure: 舒张压（毫米汞柱）
4. SkinThickness: 三头肌皮褶厚度（毫米）
5. Insulin:2 小时血清胰岛素（mu U/ml）
6. BMI: 体质指数（体重 kg/(身高 m)^2）
7. DiabetesPedigreeFunction: 糖尿病谱系功能
8. Age: 年龄

输出为一个二分类 01，即有无患病。首先引入必要的库：

```
# import dependencies
import numpy as np
import pandas as pd

# other dependencies that you might not need
# just for publishing image in notebook
from IPython.display import Image
from IPython.core.display import HTML
%matplotlib  inline
```

[1] ✓ 2.2s

然后读入数据集：

```
# column has all the name of column name
# our data is stored in dataframe: data

column = ["Pregnancies","Glucose","BloodPressure","SkinThickness","Insulin","BMI","DiabetesPedigreeFunction","Age","Outcome"]
data = pd.read_csv('pima-indians-diabetes.data.csv',names=column)
```

[2] ✓ 0.9s

```
data.head()
```

[3] ✓ 0.1s

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

根据概率学公式：

$$P(c \mid x) = \frac{P(x \mid c)P(c)}{P(x)}$$

其中，

1. $P(c \mid x)$ 被称为给定特征 x 后，类别 c 的**后验概率**。
2. $P(c)$ 是类别 c 的**先验概率**。
3. $P(x \mid c)$ 被称为给定类别 c 后特征 x 的**似然度**。
4. $P(x)$ 被称为特征 x 的**边缘概率**，可以理解为特征的先验主观概率。

根据贝叶斯公式，我们分别计算给定特征数据后，outcome0 和 outcome1 的后验概率：

$$P(\frac{Outcome\ 0}{Data}) = \frac{P(\frac{Pregnancies}{Outcome\ 0}) \times P(\frac{Glucose}{Outcome\ 0}) \times P(\frac{BloodPressure}{Outcome\ 0}) \times P(\frac{SkinThickness}{Outcome\ 0}) \times P(\frac{Insulin}{Outcome\ 0}) \times P(\frac{BMI}{Outcome\ 0}) \times P(\frac{DiabetesPedigreeFunction}{Outcome\ 0}) P(\frac{Age}{Outcome\ 0}) \times P(Outcome\ 0)}{marginal\ property}$$

$$P(\frac{Outcome\ 1}{Data}) = \frac{P(\frac{Pregnancies}{Outcome\ 1}) \times P(\frac{Glucose}{Outcome\ 1}) \times P(\frac{BloodPressure}{Outcome\ 1}) \times P(\frac{SkinThickness}{Outcome\ 1}) \times P(\frac{Insulin}{Outcome\ 1}) \times P(\frac{BMI}{Outcome\ 1}) \times P(\frac{DiabetesPedigreeFunction}{Outcome\ 1}) P(\frac{Age}{Outcome\ 1}) \times P(Outcome\ 1)}{marginal\ property}$$

整个朴素贝叶斯分类器就需要五个步骤：

1. 计算先验概率。
2. 计算似然度。
3. 计算边缘概率。
4. 利用概率公式计算数据点的后验概率。
5. 进行分析。

1 先验概率的计算

在分类中，我们就把两个类别的先验概率看作是两个类别的样本数量占总样本数的比例：

```
# Number of patients of outcome 1
n_outcome1 = data['Outcome'][data['Outcome'] == 1].count()

# Number of patients of outcome 0
n_outcome0 = data['Outcome'][data['Outcome'] == 0].count()

# Total people
total_ppl = data['Outcome'].count()
```

[6] ✓ 0.9s

```
# Number of people of outcome1 divided by the total people
P_outcome1 = n_outcome1/total_ppl

# Number of people of outcome0 divided by the total people
P_outcome0 = n_outcome0/total_ppl
```

[7] ✓ 0.9s

2 计算似然度

- 我们假定给定类别后，每个特征数据都是服从标准正态分布的。
- 根据样本数据，我们可以给出特征分布的参数的无偏估计。以怀孕次数为例：

$$p\left(\frac{Pregnancies}{Outcome\ 1}\right) = \frac{1}{\sqrt{2\pi \text{ variance of Outcome1 pregnancies in the data}}} \times e^{-\frac{(\ observation's\ pregnancies\ -\ average\ pregnancy\ of\ \ outcome0\ in\ the\ data)^2}{2\ variance\ of\ \ outcome0\ pregnancy\ in\ the\ data}}$$

根据概率学知识，样本均值就是概率分布期望的似然估计，样本方差就是概率分布方差的似然估计。可以计算得出样本的均值和方差为：

```
# Now first calculate the means of the data according to outcome

# Group the data by gender and calculate the means of each feature
data_means = data.groupby('Outcome').mean()

# View the values
data_means
```

[9] ✓ 0.1s

Outcome	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age
0	3.298000	109.980000	68.184000	19.664000	68.792000	30.304200	0.429734	31.190000
1	4.865672	141.257463	70.824627	22.164179	100.335821	35.142537	0.550500	37.067164

```
# Second calculate the variance of the data according to outcome

# Group the data by gender and calculate the variance of each feature
data_variance = data.groupby('Outcome').var()

# View the values
data_variance
```

[10] ✓ 0.1s

Outcome	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age
0	9.103403	683.362325	326.274693	221.710525	9774.345427	59.133870	0.089452	136.134168
1	13.996870	1020.139457	461.897968	312.572195	19234.673319	52.750693	0.138648	120.302588

同理可以计算得到其他特征的概率分布函数。

3 边缘概率

现实中边缘概率是极难计算的，我们无法得知每个特征在全体中的占比，就像我们很难知道” 红心的西瓜占全体西瓜的多少” 一样。但在朴素贝叶斯分类器中，我们直接认为边缘概率就是具有特征 x 占全部样本数量的比例。

4 贝叶斯分类器的应用

我们用已知的贝叶斯分类器数据去预测新的数据点，计算得到给定 data 后 outcome0 和 outcome1 的数据比较哪个大确认结果是 0 还是 1。最后在测试集上预测准确率有 73%：

```
# now the model will train in training dataset and then we will use test dataset to predict its accuracy

from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.33, random_state=42)
```

[23] ✓ 0.8s

```
# now preparing our model as per Gaussian Naive Bayesian

from sklearn.naive_bayes import GaussianNB

model = GaussianNB().fit(X_train, y_train) #fitting our model
```

[24] ✓ 0.9s

```
predicted_y = model.predict(X_test) #now predicting our model to our test dataset
```

[25] ✓ 0.9s

```
from sklearn.metrics import accuracy_score

# now calculating that how much accurate our model is with comparing our predicted values and y_test values
accuracy_score = accuracy_score(y_test, predicted_y)
print (accuracy_score)
```

[26] ✓ 0.1s

```
0.7362204724409449
```