课后作业答案：NaiveBayes

作者：欧新宇（Xinyu OU）

【作业提交】

将分类结果保存到文本文档进行提交(写上每一题的题号和题目，然后再贴答案)，同时提交源代码。

1. 测试结果命名为: ex05-结果-你的学号-你的姓名.txt
2. 源代码命名: ex05-01-你的学号-你的姓名.py, ex05-02-你的学号-你的姓名.py, ex05-03-你的学号-你的姓名.py

结果文件，要求每小题标注题号，两题之间要求空一行

要求在 “糖尿病预测” 数据集上使用高斯 (Gaussian) 朴素贝叶斯完成以下任务，要求如下：

1. 要求训练集和测试集的分割比例为80%:20%, 给出高斯 (Gaussian) 朴素贝叶斯在训练集和测试集上的分类精度（ex05-01, ex05-结果）
2. 对于第79个测试样本，输出对于该样本的类别预测值，以及每个类别的预测概率（ex05-02, ex05-结果）
3. 给定新样本，给出该样本的类别，以及每个类别的预测概率。（ex05-03, ex05-结果）

样本中各个参数的值为：

- Pregnancies: 【学号//6】
- Glucose：【学号*3】
- BloodPressure：【学号*2】
- SkinThickness：【学号】
- Insulin：【学号*4】
- BMI：30+【学号/7】
- DiabetesPedigreeFunction：【学号/6】
- Age：【学号】

【学号】= 你的学号的后两位

提示：由于糖尿病数据集时通过pandas进行输入的，所以在进行数据操作和处理的时候，需要转换为numpy数据组，实现方法参考如下：

```
X_test = np.array(X_test)[data_id]
```

参考代码:

1.要求训练集和测试集的分割比例为80%:20%, 给出高斯 (Gaussian) 朴素贝叶斯在训练集和测试集上的分类精度（ex05-01, ex05-结果）

```
# global.conf
file_path = 'D:\\CloudStation\\MyWebsites\\Teaching\\MachineLearning\\datasets\\diabetes.csv'

test_size = 0.3
```

```
# user.conf
noStudent = 18
```

```
# 加载 pandas库，并使用read_csv()函数读取糖尿病预测数据集diabetes
import pandas as pd
data = pd.read_csv(file_path)

# 将数据中的特征和标签进行分离，其中第0位位索引号，第1-8位位特征，第9位为标签
X = data.iloc[:, 0:8]
y = data.iloc[:, 8]

# 以 70%:30%的比例对训练集和测试集进行拆分
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size =
test_size)

# 引入KNN分类模型，并配置KNN分类器，设置近邻数 = 2
from sklearn.naive_bayes import GaussianNB
gnb = GaussianNB()
gnb.fit(X_train, y_train)

train_score = gnb.score(X_train, y_train)
test_score = gnb.score(X_test, y_test)

print("训练集评分:{0:.2f}；测试集评分:{1:.2f}".format(train_score, test_score))
```

```
训练集评分:0.79；测试集评分:0.73
```

2.对于第79个测试样本，输出对于该样本的类别预测值，以及每个类别的预测概率（ex05-02, ex05-结果）

```
import numpy as np

data_id = 39
data_x = [np.array(X_test)[data_id]]
data_y = np.array(y_test)[data_id]

print("样本的正确分类为: {}".format(data_y))

print("GaussianNB模型预测的分类是: {}".format(gnb.predict(data_x)[0]))
print("+ 属于分类0的概率值是：{:.5f}".format(gnb.predict_proba(data_x)[0][0]))
print("+ 属于分类1的概率值是：{:.5f}".format(gnb.predict_proba(data_x)[0][1]))
```

```
样本的正确分类为: 0
GaussianNB模型预测的分类是: 0
+ 属于分类0的概率值是：0.85878
+ 属于分类1的概率值是：0.14122
```

```
gnb.predict_proba(data_x)[0]
```

```
array([0.85877521, 0.14122479])
```

3.给定新样本，给出该样本的类别，以及每个类别的预测概率。（ex05-03, ex05-结果）

```
import numpy as np
noStudent = 18
X_new = np.array([[noStudent//6, noStudent*3,
                   noStudent*2, noStudent, noStudent*4,
                   noStudent/7, noStudent/6, noStudent]])

prediction = gnb.predict(X_new)
prediction_prob = gnb.predict_proba(X_new)
print("新样本的预测分类为：{}".format(prediction))
print("+ 属于分类0的概率值是：{:.5f}".format(prediction_prob[0][0]))
print("+ 属于分类1的概率值是：{:.5f}".format(prediction_prob[0][1]))
```

```
新样本的预测分类为：[1]
+ 属于分类0的概率值是：0.07877
+ 属于分类1的概率值是：0.92123
```