机器学习作业--朴素贝叶斯

分层采样

从一个可以分成不同子总体（或称为层）的总体中，按规定的比例从不同层中随机抽取样品的方法。这种方法的优点是，样本的代表性比较好，抽样误差比较小。

这里对于三种不同的酒分别进行采样，可以直接采样选取前面的一定比例，也可以随机抽取一定比例。

a类59个，b类71个，c类48个。

查看数据集的分布，大致符合钟形曲线，因此我们之后也是按照高斯分布的前提去进行贝叶斯分类的。

image-20211116170118726

下面的代码是随机抽取的代码，分别选了a类50个，b类61个，c类41个，是按照比例选取的，符合分层采样的要求。

```
import pandas as pd
data=pd.read_csv("F:\\wine.data",header=None)
a=np.linspace(0,59,60).astype(int)
temp=np.random.choice(a,50)
a1=data.iloc[temp]
b=np.linspace(60,130,71).astype(int)
temp=np.random.choice(b,61)
b1=data.iloc[temp]
a=np.linspace(131,177,48).astype(int)
temp=np.random.choice(c,41)
c1=data.iloc[temp]
```

朴素贝叶斯估计

贝叶斯公式如下：

$$P(c|x)=\frac{p(x1|c)p(x2|c)...p(x13|c)p(c)}{p(X)}$$

这里假定所有的数据都服从高斯分布，因此需要计算**训练集**中每种分布的标准差和平均值。

```
mean_a=np.mean(train.iloc[0:50])
std_a=np.std(train.iloc[0:50])
mean_b=np.mean(train.iloc[50:109])
std_b=np.std(train.iloc[50:109])
mean_c=np.mean(train.iloc[110:150])
std_c=np.std(train.iloc[110:150])
```

然后将**测试集**中的数据带入高斯分布的概率分布公式中计算贝叶斯概率，进行分类。

```
from scipy.stats import norm
def calc(x):
    pdf_a=1
    pdf_b=1
    pdf_c=1
    for i in range(13):
        pdf_a*=1/(np.sqrt(2*np.pi)*std_a[i+1])*np.exp(-np.square(x[i+1]-
mean_a[i+1])/(2*np.square(std_a[i+1])))
        pdf_b*=1/(np.sqrt(2*np.pi)*std_b[i+1])*np.exp(-np.square(x[i+1]-
mean_b[i+1])/(2*np.square(std_b[i+1])))
        pdf_c*=1/(np.sqrt(2*np.pi)*std_c[i+1])*np.exp(-np.square(x[i+1]-
mean_c[i+1])/(2*np.square(std_c[i+1])))
    out_a=pdf_a*44/134
    out_b=pdf_b*54/134
    out_c=pdf_c*36/134
    return np.argmax([out_a,out_b,out_c])
```

```
abt=['a','b','c']
for i in range(44):
    if(calc(test.iloc[i])+1-test[0].iloc[i]!=0):
        print("wrong prediction")
```

```
test
```

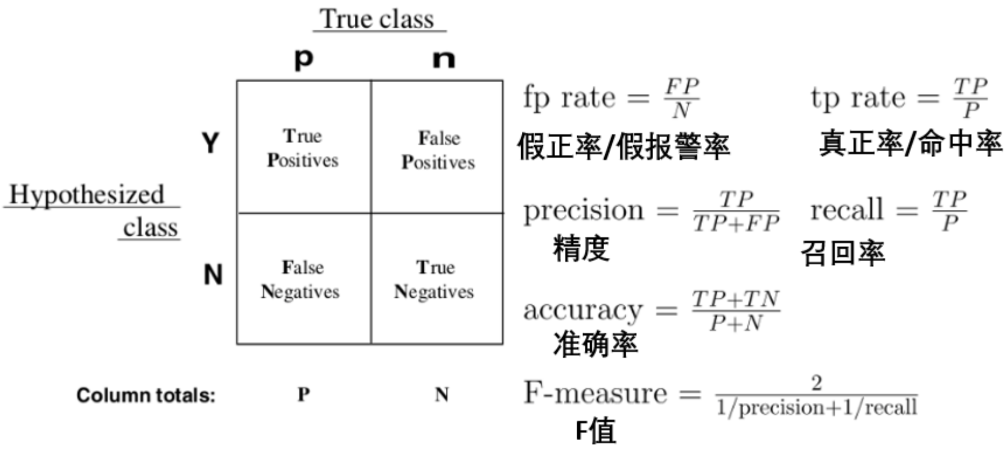
对模型在整个测试集（总数据集除去134个训练集样本之后，总共44个数据）上进行估计，测试结果如图所示，没有错误估计输出。

正确率百分之百。

混淆矩阵

因为这是一个三分类任务，因此混淆矩阵应该是3×3格式的。在许多次测试之后，终于有了一次正确率不是百分百的。

预测值\真实值	a	b	c
a	14	1	0
b	0	17	0
c	0	0	12



这里的混淆矩阵是3×3的，之后我们需要对每一种分类计算精度，召回率，准确率和F值。

精度，召回率，准确率和F值的计算方式如上图所示。

a的准确率为93.3%，精度为93.3%，召回率为100%，f值为0.967

b的准确率为100%，精度为100%，召回率为94.4%，f值为0.972

c的准确率为100%，精度为100%，召回率为100%，f值为1.