第四课　分类:贝叶斯分类

主讲人：丁兆云

dingzhaoyun1983@163.com　17607310865　https：//github.com/zyding1983/datamining

数据挖掘

Data Mining

01 回顾

02 实践

分类概念

- 什么是分类?
  - 找出描述和区分数据类或概念的**模型**，以便能够使用模型**预测类标号未知的对象的类标号**
- 一般过程
  - 学习阶段
    - 建立描述预先定义的数据类或概念集的**分类器**
    - **训练集**提供了每个训练元组的类标号，分类的学习过程也称为监督学习（supervised learning）
  - 分类阶段
    - 使用定义好的分类器进行分类的过程

分类中的训练集与测试集

训练集

id	年龄	收入	爱好	信用	购买
1	青	高	否	中	否
2	青	高	否	优	否
3	中	高	否	中	是
4	老	中	否	中	是
5	老	低	是	中	是
6	老	低	是	优	否
7	中	低	是	优	是
8	青	中	否	中	否
9	青	低	是	中	是
10	老	中	是	中	是
11	青	中	是	优	是
12	中	中	否	优	是
13	中	高	是	中	是
14	老	中	否	优	否

测试集

一个收入中等、信用度良好的青年爱好游戏顾客。

是否会购买电脑呢?

分类中的训练集与测试集

训练集

id	年龄	收入	爱好	信用	购买
1	青	高	否	中	否
2	青	高	否	优	否
3	中	高	否	中	是
4	老	中	否	中	是
5	老	低	是	中	是
6	老	低	是	优	否
7	中	低	是	优	是
8	青	中	否	中	否
9	青	低	是	中	是
10	老	中	是	中	是
11	青	中	是	优	是
12	中	中	否	优	是
13	中	高	是	中	是
14	老	中	否	优	否

测试集

一个收入中等、信用度良好的青年爱好游戏顾客。

是否会购买电脑呢?

讨论:

条件概率是什么?

先验概率是什么?

后验概率是什么?

分类中的训练集与测试集

训练集

id	年龄	收入	爱好	信用	购买
1	青	高	否	中	否
2	青	高	否	优	否
3	中	高	否	中	是
4	老	中	否	中	是
5	老	低	是	中	是
6	老	低	是	优	否
7	中	低	是	优	是
8	青	中	否	中	否
9	青	低	是	中	是
10	老	中	是	中	是
11	青	中	是	优	是
12	中	中	否	优	是
13	中	高	是	中	是
14	老	中	否	优	否

测试集

一个收入中等、信用度良好的青年爱好游戏顾客。

是否会购买电脑呢?

h的似然概率　　h的先验概率

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

h的后验概率　　D的先验概率

D：待测试数据
h：假设类别

01

回顾

02

实践

贝叶斯分类编程实践


```
from sklearn import tree
import pandas as pd
from sklearn.naive_bayes import GaussianNB
data_url = "diabetes.csv"
df = pd.read_csv(data_url)
x = df.ix[:, 0:8]
#print(x)
y= df.ix[:, 8]
#print(y)
#X = [[0, 0], [1, 1]]
#Y = [0, 1]
clf = GaussianNB()
clf = clf.fit(x, y)
data_urltest = "diabetestest.csv"
dftest = pd.read_csv(data_urltest)
print(clf.predict(dftest))
#print(clf.predict_proba([[2., 2.]]))
```

贝叶斯分类编程实践

https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.GaussianNB.html#sklearn.naive_bayes.GaussianNB

https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.MultinomialNB.html#sklearn.naive_bayes.MultinomialNB

https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.ComplementNB.html#sklearn.naive_bayes.ComplementNB

https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.BernoulliNB.html#sklearn.naive_bayes.BernoulliNB

https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.BernoulliNB.html#sklearn.naive_bayes.BernoulliNB

同学们可以尝试利用python读入本地iris数据集，来完成贝叶斯分类，分析其分类效果

Any Questions?

谢　谢！