INT104W12_往年试卷解析v0.1

大题部分

混淆矩阵

(i.e. "prediction=No" means the account has less than \$1000):

(1) In reality, there are a total of [a] (5+125=130) accounts with a balance more than \$1000 and [b] (60+10=70) accounts with a balance less than \$1000.

(2) The overall accuracy is [c] and the recall is [d].

n=200	Prediction=No	Prediction=Yes
Actual=No	60	10
Actual=Yes	5	125

(8 Marks)

横着看是事实上的真/假例（即真实的各个类别）；竖着看是预测得的正/负例（即预测得出的各类别）。

- Precision：真例中的正例/所有正例。

$$查准率Precision = \frac{\sum_1^n TP_n}{\sum_1^n (TP_n + FP_n)}$$

- Recall：真例中的正例/所有真例。

$$查全率Recall = \frac{\sum_1^n TP_n}{\sum_1^n (TP_n + FN_n)}$$

- Accuracy：模型正确分类的样本/总例。

$$准确率Accuracy = \frac{\sum_1^n (TP_n + TN_n)}{\sum_1^n (TP_n + FP_n + FN_n + TN_n)}$$

- accuracy=60+125/60+10+5+125=0.925
- recall=125/125+5=0.962

Python看api填空代码

numpy.random.randint(low, high=None, size=None, dtype=int)

- `size`: Output shape
- `dtype`: Desired dtype of the result

```
import numpy as np
number = np.random.randint(1, [e] , size=( [f] ))
```

100 100

size类型：size=(100,2)代表结果的大小是100行2列，size=100代表大小是一个有100个元素的数组。

整除print(10//3)=3

取余print(10%3)=1

朴素贝叶斯计算后验概率

Outlook	Humidity	Wind Speed	Preference
Rainy	80%	0.5m/s	Yes
Rainy	40%	0.2m/s	Yes
Rainy	50%	5.0m/s	No
Rainy	50%	0.2m/s	Yes
Rainy	75%	4.0m/s	No
Sunny	70%	5.0m/s	No
Sunny	75%	0.4m/s	No
Sunny	80%	0.1m/s	No
Sunny	50%	0.2m/s	Yes
Sunny	40%	4.0m/s	Yes

请计算在Outlook=Rainy，Humidity<65%的情况下玩家的表现。

$$P(Yes|Rainy, Humidity < 65\%) = \frac{P(Rainy, Humidity < 65\%|Yes)P(Yes)}{P(Rainy, Humidity < 65\%)} \tag{1}$$
$$= \frac{P(Rainy|Yes)P(Humidity < 65\%|Yes)P(Yes)}{P(Rainy, Humidity < 65\%)} \tag{2}$$
$$\propto P(Rainy|Yes)P(Humidity < 65\%|Yes)P(Yes) \tag{3}$$
$$= 3/5 * 4/5 * 5/10 = 0.24 \tag{4}$$
$$P(No|Rainy, Humidity < 65\%) = \frac{P(Rainy, Humidity < 65\%|No)P(No)}{P(Rainy, Humidity < 65\%)} \tag{5}$$
$$= \frac{P(Rainy|Yes)P(Humidity < 65\%|Yes)P(Yes)}{P(Rainy, Humidity < 65\%)} \tag{6}$$
$$\propto P(Rainy|No)P(Humidity < 65\%|No)P(No) \tag{7}$$
$$= 2/5 * 1/5 * 5/10 = 0.04 \tag{8}$$
$$\text{因为} P(Yes|Rainy, Humidity < 65\%) > P(No|Rainy, Humidity < 65\%) \text{故更有可能去} \tag{9}$$

kNN用K近邻算法归类偏好（本次期末考试不会出现）

knn属于监督学习，看该样本最近的k个样本的label，投票决定样本类别

湿度65%，风速3m/s，晴天。

是否需要计算晴天阴天等非数值距离？自己附一个值，或考试中会告诉你

城市（曼哈顿）距离：

然后你会发现在计算距离前需要zscore标准化，否则由于数据规模不同，湿度的影响太小。

计算交叉验证的准确度

k-means聚类迭代

层次聚类的几种距离计算方式

- 最小距离：两个簇的最近样本决定，又称为单链接算法（Single linkage）。
- 最大距离：两个簇的最远样本决定，又称为全链接算法（Complete linkage）。
- 平均距离：两个簇的所有样本对距离平均值决定，又称为均链接算法。
- 中心距离：两个簇的中心间的距离决定。
- 最小方差/离差平方和（ward）：两个簇的所有样本对的距离平方和的平均决定。

看清楚要什么样的linkage

计算信息熵信息增益（涉及log不太会考），基尼不纯度