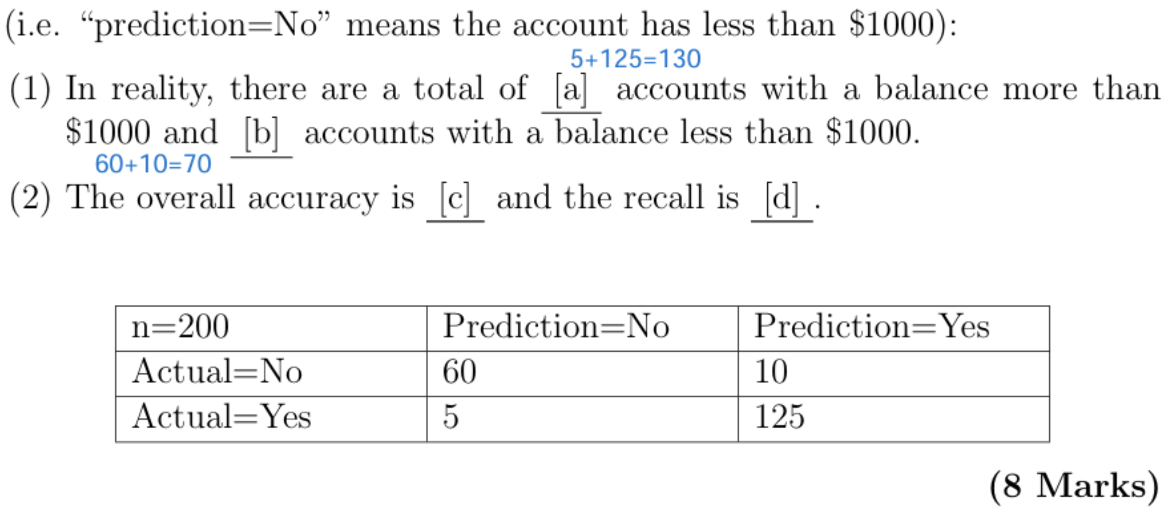
INT104W12_往年试卷解析v0.1

大题部分

混淆矩阵



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

5+125=130

(1) In reality, there are a total of [a] accounts with a balance more than $1000 and [b] accounts with a balance less than $1000.

60+10=70

(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。

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

- Recall：真例中的正例/所有真例。

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

- Accuracy：模型正确分类的样本/总例。

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

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                        100             100
number = np.random.randint(1, __[e]__ , size=( __[f]__ ))
```

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。

$$
\begin{aligned}
P(Yes|Rainy, Humidity < 65\%) &= \frac{P(Rainy, Humidity < 65\%|Yes)P(Yes)}{P(Rainy, Humidity < 65\%)} && (1)\\
&= \frac{P(Rainy|Yes)P(Humidity < 65\%|Yes)P(Yes)}{P(Rainy, Humidity < 65\%)} && (2)\\
&= \frac{3/5 * 4/5 * 5/10}{3/10} && (3)\\
&= 80\% && (4)
\end{aligned}
$$

kNN用K近邻算法归类偏好

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?

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