模式识别与统计学习作业

学号：19001353
姓名：丁一鸣

Last TEX in April 5, 2022

1 基本概念

1.1 混淆矩阵与评价指标

Listing 1.1: Insert code directly in your document

```python
import numpy as np
import matplotlib.pyplot as plt
from sklearn import metrics
import csv

# read the data
with open('score.csv', newline='') as f:
    reader = csv.reader(f)
    s = list(reader)
tmp = s[0]
scores = np.array([float(item) for item in tmp])
with open('label.csv', newline='') as f:
    reader = csv.reader(f)
    l = list(reader)
tmp = l[0]
label = np.array([float(item) for item in tmp])

# build confusion matrix with a threshold of 0.05
threshold = 0.05
label_pred=scores.copy()
label_pred[scores>threshold]=1
label_pred[scores<threshold]=0
cm = metrics.confusion_matrix(label, label_pred)
disp = metrics.ConfusionMatrixDisplay(confusion_matrix=cm)
disp.plot()
plt.show()

# get TP, FP, TN, FN from confusion matrix
TN = cm[0][0]
FN = cm[1][0]
TP = cm[1][1]
FP = cm[0][1]

# compute Precision, Recall, F1-score and Accuracy
P=TP/(TP+FP)
R=TP/(TP+FN)
F1=(2*P*R)/(P+R)
acc=(TP+TN)/len(label)
print("Precision: ",P,"\nRecall:",R,"\nF1-score: ",F1,"\nAccuracy: ",
    acc)
```

```python

# compute FPR, TPR, AUC and draw ROC curve
fpr, tpr, thresholds = metrics.roc_curve(label, scores, pos_label=1)
print(thresholds)
roc_auc = metrics.auc(fpr, tpr)
plt.plot(
    fpr,
    tpr,
    color="darkorange",
    label="ROC curve (area = %0.2f)" % roc_auc,
)
plt.plot([0, 1], [0, 1], color="navy", linestyle="--")
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel("False Positive Rate")
plt.ylabel("True Positive Rate")
plt.title("Receiver operating characteristic example")
plt.legend(loc="lower right")
plt.show()
```

2 朴素贝叶斯

2.1 条件独立性证明

定义: A 和 B 在给定事件 C 的条件下相互独立，如果

$$P(A, B \mid C) = P(A \mid C)P(B \mid C)$$

证明: 事件 A，B 和 C 在给定事件 C 的条件下相互独立，当且仅当 $P(C) > 0$，且

$$P(A \mid B, C) = P(A \mid C)$$

证明: “⇒”：因为 A，B 和 C 在给定事件 C 的条件下相互独立，根据定义有

$$P(A, B \mid C) = P(A \mid C)P(B \mid C)$$

又有 $P(C) > 0$，则

$$\begin{aligned}
P(A \mid B, C) &= P(A \mid B, C)\frac{P(B \mid C)}{P(B \mid C)} \\
&= \frac{P(A \mid B, C)P(B \mid C)}{P(B \mid C)} \\
&= \frac{\frac{P(A,B,C)}{P(C)}}{P(B \mid C)} \\
&= \frac{P(A, B \mid C)}{P(B \mid C)} \\
&= P(A \mid C)
\end{aligned}$$

“⇐”：显然，当

$$P(A \mid B, C) = P(A \mid C)$$

两边同乘 $P(B \mid C)$，则有定义式。

2.2 骰子问题

假设有两对不同的骰子，一对是标准的骰子（每个面的点数为 1 到 6 中的一个），另一对为“增广" 的骰子，每个面的点数都增加了两个（介于 3 到 8 个点）。游戏者甲从一个装有 60% 标准对和 40% 增广对的袋子里随机选择一对进行投掷，游戏者乙在没有骰子信息的情况下，通过获知点数的和进行决策。

1. 应如何决策，使平均错误概率最小化？最小平均错误概率是多少？
2. 如果乙猜对是标准骰子对，可获得 10 元钱，猜对是增广骰子对获得 30 元钱，猜错损失 10 元钱，应如何决策，平均风险如何？

2.3 性别分类问题

训练样本：

Person	height (feet)	weight (lbs)	foot size(inches)
male	6	180	12
male	5.92 (5'11")	190	11
male	5.58 (5'7")	170	12
male	5.92 (5'11")	165	10
female	5	100	6
female	5.5 (5'6")	150	8
female	5.42 (5'5")	130	7
female	5.75 (5'9")	150	9

测试样本：

Person	height (feet)	weight (lbs)	foot size(inches)
sample	6	130	8

Figure 2.1: 性别分类数据

3 最小二乘线性回归

3.1 糖尿病数据的回归与预测

4

最近邻分类器

4.1 FashionMNIST 数据的分类

5 决策树

5.1 打网球数据的分类

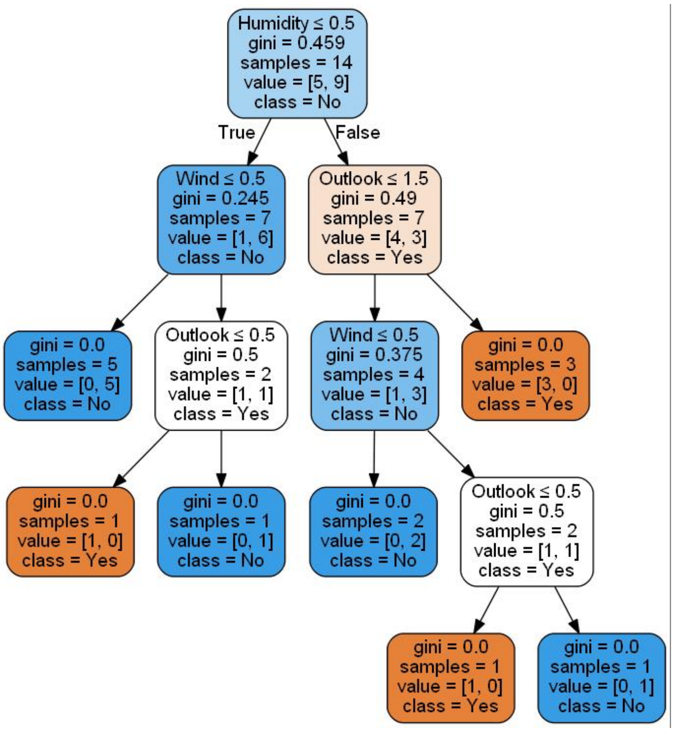

Figure 5.1: Decision tree of tennis data

Listing 5.1: Insert code directly in your document

```
from sklearn import tree
import pydotplus
import csv
f = csv.reader(open('1111.csv','r'))
# Outlook (0:Rain, 1:Overcast, 2:Suuny)
# Temprature (0:Cool, 1:Mild, 2:Hot)
```

```
# Humidity (0:Normal, 1:High)
# Wind (0:Weak, 1:Strong)
X = [[2,2,1,0],
     [2,2,1,1],
     [1,2,1,0],
     [0,1,1,0],
     [0,0,0,0],
     [0,0,0,1],
     [1,0,0,1],
     [2,1,1,0],
     [2,0,0,0],
     [0,1,0,0],
     [2,1,0,0],
     [1,1,1,1],
     [1,2,0,0],
     [0,1,1,1]]
y = [0,0,1,1,1,0,1,0,1,1,1,1,1,0]


# 建立并训练决策树
clf = tree.DecisionTreeClassifier()
clf = clf.fit(X, y)

# 预测结果
dot_data = tree.export_graphviz(clf, out_file=None,
                        feature_names=['Outlook', 'Temprature','Humidity'
                            ,'Wind'],
                        class_names=['Yes', 'No'],
                        filled=True, rounded=True,
                        special_characters=True)
graph = pydotplus.graph_from_dot_data(dot_data)
graph.write_jpg("tree.jpg")      # 生成jpg文件
```

5.2 不纯度指数的计算

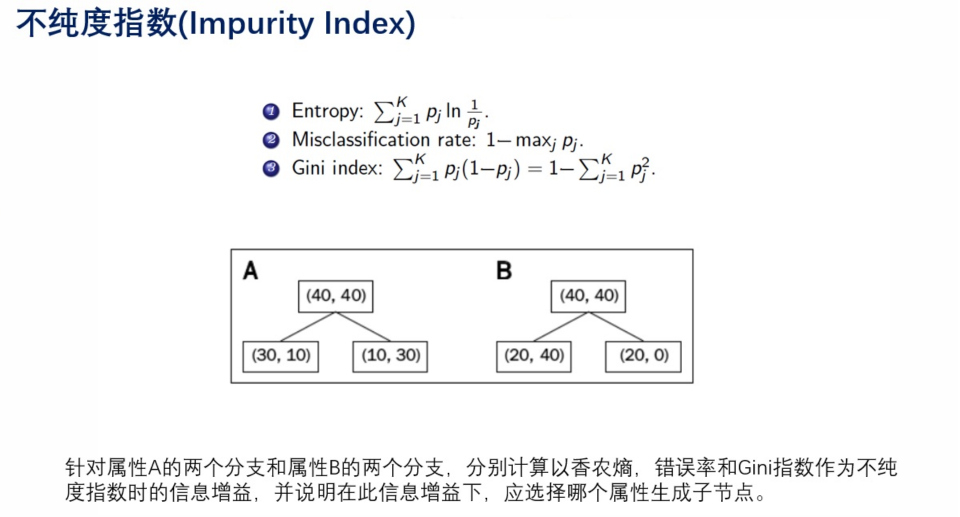

Figure 5.2: 不纯度计算题目

解：

1. Entropy

2.