模式识别与统计学习 决策树作业

题目：请使用使用sklearn里的DecisionTreeClassifier进行编程实现对下表中数据进行分类。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| day | outlook | temp | humidity | windy | play |
| 1 | sunny | hot | high | FALSE | no |
| 2 | sunny | hot | high | TRUE | no |
| 3 | overcast | hot | high | FALSE | yes |
| 4 | rainy | mild | high | FALSE | yes |
| 5 | rainy | cool | normal | FALSE | yes |
| 6 | rainy | cool | normal | TRUE | no |
| 7 | overcast | cool | normal | TRUE | yes |
| 8 | sunny | mild | high | FALSE | no |
| 9 | sunny | cool | normal | FALSE | yes |
| 10 | rainy | mild | normal | FALSE | yes |
| 11 | sunny | mild | normal | TRUE | yes |
| 12 | overcast | mild | high | TRUE | yes |
| 13 | overcast | hot | normal | FALSE | yes |
| 14 | rainy | mild | high | TRUE | no |

解答：

代码：

|  |
| --- |
| from sklearn import tree  from matplotlib import pyplot as plt  import numpy as np  # Category 0 1 2  # Outlook[0] Sunny Overcast Rain  # Temperature[1] Hot Mild Cold  # Humidity[2] High Normal  # Wind[3] Weak Strong  # Play Tennis No Yes  data = np.array([  [0, 0, 0, 0],  [0, 0, 0, 1],  [1, 0, 0, 0],  [2, 1, 0, 0],  [2, 2, 1, 0],  [2, 2, 1, 1],  [1, 2, 1, 1],  [0, 1, 0, 0],  [0, 2, 1, 0],  [2, 1, 1, 0],  [0, 1, 1, 1],  [1, 1, 0, 1],  [1, 0, 1, 0],  [2, 1, 0, 1]  ])  target = np.array([[0], [0], [1], [1], [1], [0], [1], [0], [1], [1], [1], [1], [1], [0]])  clf = tree.DecisionTreeClassifier(criterion="entropy")  clf.fit(data, target)  tree.plot\_tree(clf)  plt.show() |

输出：

数学推导：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| day | outlook | temp | humidity | windy | play |
| 1 | sunny | hot | high | FALSE | no |
| 2 | sunny | hot | high | TRUE | no |
| 3 | overcast | hot | high | FALSE | yes |
| 4 | rainy | mild | high | FALSE | yes |
| 5 | rainy | cool | normal | FALSE | yes |
| 6 | rainy | cool | normal | TRUE | no |
| 7 | overcast | cool | normal | TRUE | yes |
| 8 | sunny | mild | high | FALSE | no |
| 9 | sunny | cool | normal | FALSE | yes |
| 10 | rainy | mild | normal | FALSE | yes |
| 11 | sunny | mild | normal | TRUE | yes |
| 12 | overcast | mild | high | TRUE | yes |
| 13 | overcast | hot | normal | FALSE | yes |
| 14 | rainy | mild | high | TRUE | no |

1. 计算决策属性的熵
2. 根节点选取
3. Outlook
4. Temperature
5. Humidity
6. Wind

所以根节点选择Humidity.

3.1 Humidity = high

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| day | outlook | temp | humidity | windy | play |
| 1 | sunny | hot | high | FALSE | no |
| 2 | sunny | hot | high | TRUE | no |
| 3 | overcast | hot | high | FALSE | yes |
| 4 | rainy | mild | high | FALSE | yes |
| 8 | sunny | mild | high | FALSE | no |
| 12 | overcast | mild | high | TRUE | yes |
| 14 | rainy | mild | high | TRUE | no |

所以，下一节点选择Outlook.

3.1.1 Outlook != Sunny

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| day | outlook | temp | humidity | windy | play |
| 3 | overcast | hot | high | FALSE | yes |
| 4 | rainy | mild | high | FALSE | yes |
| 12 | overcast | mild | high | TRUE | yes |
| 14 | rainy | mild | high | TRUE | no |

3.1.2 Wind = Ture

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| day | outlook | temp | humidity | windy | play |
| 12 | overcast | mild | high | TRUE | yes |
| 14 | rainy | mild | high | TRUE | no |

3.2 Humidity = normal

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| day | outlook | temp | humidity | windy | play |
| 5 | rainy | cool | normal | FALSE | yes |
| 6 | rainy | cool | normal | TRUE | no |
| 7 | overcast | cool | normal | TRUE | yes |
| 9 | sunny | cool | normal | FALSE | yes |
| 10 | rainy | mild | normal | FALSE | yes |
| 11 | sunny | mild | normal | TRUE | yes |
| 13 | overcast | hot | normal | FALSE | yes |

3.2.1 Wind = true

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| day | outlook | temp | humidity | windy | play |
| 6 | rainy | cool | normal | TRUE | no |
| 7 | overcast | cool | normal | TRUE | yes |
| 11 | sunny | mild | normal | TRUE | yes |