**CSE 40647/60647 Data Science (Spring 2018)**

**Lecture 6: Classification: Concepts and Decision Tree**

**Goals:**

* Describe the difference between classification and clustering
* Describe two steps of the classification process
  + Describe what is entropy; describe and compare the following “feature selection measures” or called “splitting criteria”: information gain, gain ratio, and gini index.
* Given training instances and their attributes, construct by hand and implement using Python Decision Tree models:
  + ID3: information gain
  + C4.5: gain ratio
  + CART: gini index

**Exercise 1:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  | Is Home/Away? | Is Opponent in AP Top 25 at Preseason? | Media | Label: Win/Lose |
| 1 | 9/2/17 | Temple | Home | Out | 1-NBC | Win |
| 2 | 9/9/17 | Georgia | Home | In | 1-NBC | Lose |
| 3 | 9/16/17 | Boston College | Away | Out | 2-ESPN | Win |
| 4 | 9/23/17 | Michigan State | Away | Out | 3-FOX | Win |
| 5 | 9/30/17 | Miami Ohio | Home | Out | 1-NBC | Win |
| 6 | 10/7/17 | North Carolina | Away | Out | 4-ABC | Win |
| 7 | 10/21/17 | USC | Home | In | 1-NBC | ? |
| 8 | 10/28/17 | North Carolina State | Home | Out | 1-NBC | ? |
| 9 | 11/4/17 | Wake Forest | Home | Out | 1-NBC | ? |
| 10 | 11/11/17 | Miami Florida | Away | In | 4-ABC | ? |
| 11 | 11/18/17 | Navy | Home | Out | 1-NBC | ? |
| 12 | 11/25/17 | Stanford | Away | In | 4-ABC | ? |

We have

1. 6 training instances and 6 testing instances
2. 3 attributes: (a) 2-value attribute (Home/Away), (b) 2-value attribute (In/Out), (c) 4-value attribute (NBC/ESPN/FOX/ABC)

|  |
| --- |
| Solution: |

**Exercise 2:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ID | Date | Outlook | Temperature | Humidity | Windy | Label: Play? |
| 1 | 9/1/17 | Sunny | Hot | High | "False" | No |
| 2 | 9/8/17 | Sunny | Hot | High | "True" | No |
| 3 | 9/15/17 | Overcast | Hot | High | "False" | Yes |
| 4 | 9/22/17 | Rainy | Mild | High | "False" | Yes |
| 5 | 9/29/17 | Rainy | Cool | Normal | "False" | Yes |
| 6 | 10/1/17 | Rainy | Cool | Normal | "True" | No |
| 7 | 10/8/17 | Overcast | Cool | Normal | "True" | Yes |
| 8 | 10/15/17 | Sunny | Mild | High | "False" | No |
| 9 | 10/22/17 | Sunny | Cool | Normal | "False" | Yes |
| 10 | 10/29/17 | Rainy | Mild | Normal | "False" | Yes |
| 11 | 11/1/17 | Sunny | Mild | Normal | "True" | Yes |
| 12 | 11/8/17 | Overcast | Mild | High | "True" | Yes |
| 13 | 11/15/17 | Overcast | Hot | Normal | "False" | Yes |
| 14 | 11/22/17 | Rainy | Mild | High | "True" | No |
| 15 | 11/29/17 | Rainy | Hot | High | "False" | ? |

We have

1. 14 training instances and 1 testing instance
2. 4 attributes: (a) 3-value attribute (Sunny/Overcast/Rainy), (b) 3-value attribute (Hot/Mild/Cool), (c) 2-value attribute (High/Normal), (d) 2-value attribute (True/False)

|  |
| --- |
| Solution: |

**Name: NetID:**

|  |
| --- |
| **Please write down whatever question you have about this course:** |