CSE 40647/60647: Data Science **Spring 2018**

Homework 2

Handed Out: February 6, 2018 *Due: February 20, 2018 11:59 pm*

- This assignment is due at **11:59 PM** on the due date. Contact TA if you have technical difficulties in submitting it on **Sakai**. We shall NOT accept any late submission!
- Homework must be submitted in ZIP format (including .pdf and .py). Name your ZIP file as **YourNetid-HWx.zip**. Handwritten answers must be scanned into PDF.
  —YourNetid-HWx.zip
  ——YourNetid-HWx.pdf
  ——YourNetid-HWx-Qy.py
  ——... (and any supplementary materials)
- Please use **Piazza** if you have any question about the homework.

Classification: Decision Tree (DT), Naïve Bayes, and Classification Evaluation

Goal: Given Notre Dame's football game data for the last two seasons (2015 and 2016), can we construct three classification models to predict game results on games in 2017? Can we evaluate the model performance? The **three classification models** are **ID3, C4.5, and Naïve Bayes**.

Data: Each data object (or called instance) is a game. We have three attributes: (1) "Is Home/Away?", a 2-value attribute ("Home", "Away"), (2) "Is Opponent in AP Top 25 at Preseason?", a 2-value attribute ("In", "Out"), (3) "Media", a 5-value attribute ("1-NBC", "2-ESPN", "3-FOX", "4-ABC", "5-CBS"). The label "Win/Lose" is binary ("Win", "Lose").

Training set: 24 games. Please use game ID 1-24 to ***construct*** classification models. (Background color: YELLOW)
Testing set: 12 games. Please use your classification models to ***predict*** labels of game ID 25-36 and ***evaluate*** the performance of the classification models. (Background color: BLUE)

Suppose "Win" is the **positive** label and "Lose" is the **negative** label. Keep it in mind when you use Precision and Recall to evaluate the models.

For DT models, we stop splitting instances into child nodes when one of the criteria is satisfied:
(1) All features have been used;
(2) Information Gain or Gain Ratio will be **zero** with any feature that has not yet been used.
For prediction:
(1) If the node is not pure, we use the majority of this node for prediction: For example, if we have 5 positives and 1 negatives, we predict the testing case at this node to be a positive.
(2) If the node has a balance (half/half labels), e.g., 2 positives and 2 negatives, we use the majority of the root node (the entire dataset) for prediction.

ID	Date	Opponent	Is Home or Away?	Is Opponent in AP Top 25 at Preseason?	Media	Label: Win/Lose
1	9/5/15	Texas	Home	Out	1-NBC	Win
2	9/12/15	Virginia	Away	Out	4-ABC	Win
3	9/19/15	Georgia Tech	Home	In	1-NBC	Win
4	9/26/15	UMass	Home	Out	1-NBC	Win
5	10/3/15	Clemson	Away	In	4-ABC	Lose
6	10/10/15	Navy	Home	Out	1-NBC	Win
7	10/17/15	USC	Home	In	1-NBC	Win
8	10/31/15	Temple	Away	Out	4-ABC	Win
9	11/7/15	PITT	Away	Out	4-ABC	Win
10	11/14/15	Wake Forest	Home	Out	1-NBC	Win
11	11/21/15	Boston College	Away	Out	1-NBC	Win
12	11/28/15	Stanford	Away	In	3-FOX	Lose
13	9/4/16	Texas	Away	Out	4-ABC	Lose
14	9/10/16	Nevada	Home	Out	1-NBC	Win
15	9/17/16	Michigan State	Home	Out	1-NBC	Lose
16	9/24/16	Duke	Home	Out	1-NBC	Lose
17	10/1/16	Syracuse	Home	Out	2-ESPN	Win
18	10/8/16	North Carolina State	Away	Out	4-ABC	Lose
19	10/15/16	Stanford	Home	In	1-NBC	Lose
20	10/29/16	Miami Florida	Home	Out	1-NBC	Win
21	11/5/16	Navy	Home	Out	5-CBS	Lose
22	11/12/16	Army	Home	Out	1-NBC	Win
23	11/19/16	Virginia Tech	Home	In	1-NBC	Lose
24	11/26/16	USC	Away	In	4-ABC	Lose
25	9/2/17	Temple	Home	Out	1-NBC	Win
26	9/9/17	Georgia	Home	In	1-NBC	Lose
27	9/16/17	Boston College	Away	Out	2-ESPN	Win
28	9/23/17	Michigan State	Away	Out	3-FOX	Win
29	9/30/17	Miami Ohio	Home	Out	1-NBC	Win
30	10/7/17	North Carolina	Away	Out	4-ABC	Win
31	10/21/17	USC	Home	In	1-NBC	Win
32	10/28/17	North Carolina State	Home	Out	1-NBC	Win
33	11/4/17	Wake Forest	Home	Out	1-NBC	Win
34	11/11/17	Miami Florida	Away	In	4-ABC	Lose
35	11/18/17	Navy	Home	Out	1-NBC	Win
36	11/25/17	Stanford	Away	In	4-ABC	Lose

[25'] Question 1: ID3 model, a decision tree model using "Information Gain"

(1) Programming: Use **ID3** to construct a decision tree based on the training set (24 games). Use the tree to predict labels of instances in the testing set (12 games) based on their attributes. Calculate Accuracy, Precision, Recall, and F1 score on the testing result.
(2) Please submit your code as **YourNetid-HW2-Q1.py**. Attach a figure of your decision tree (either hand- or electronically drawn) and write down prediction label of the 12 testing games as well as evaluation result in the PDF.

[25'] Question 2: C4.5 model, a decision tree model using "Gain Ratio"

(1) Programming: Use **C4.5** to construct a decision tree based on the training set (24 games). Use the tree to predict labels of instances in the testing set (12 games) based on their attributes. Calculate Accuracy, Precision, Recall, and F1 score on the testing result.
(2) Please submit your code as **YourNetid-HW2-Q2.py**. Attach a figure of your decision tree (either hand- or electronically drawn) and write down prediction label of the 12 testing games as well as evaluation result in the PDF.

[35'] Question 3: Naïve Bayes model

(1) Programming: Use **Naïve Bayes** to predict labels of instances in the testing set (12 games) based on the training set (24 games). Calculate Accuracy, Precision, Recall, and F1 score on the testing result.
(2) Please submit your code as **YourNetid-HW2-Q3.py**. ~~Attach a figure of your decision tree (either hand- or electronically drawn) and~~ write down prediction label of the 12 testing games as well as evaluation result in the PDF.

[15'] For this Notre Dame game prediction task, which model is the best, which model performs the worst? Can you explain why? Write down in the PDF.