CSE 40647/60647: Data Science (Fall 2017)

Homework 4

Handed Out: October 12, 2017

Due: November 9, 2017

General Instructions

- This assignment is due at 11:59 PM on the due date.
- We will be using Sakai (https://sakailogin.nd.edu/portal/site/FA17-CSE-40647-CX-01) for collecting this assignment. Contact TA if you face technical difficulties in submitting the assignment. We shall NOT accept any late submission!
- The homework MUST be submitted in pdf format. You can handwrite trees/figures and scan them into PDF. Name your pdf file as YourNetid-HW4.pdf.
- Please use Piazza if you have questions about the homework. Also feel free to send TA emails and come to office hours.

Classification: Decision Tree, Naïve Bayes, and Classification Evaluation

Goal: Given Notre Dame's football game data for the last two seasons (2015 and 2016), can we construct four Classification models to test on past games at this season (the first six games in 2017) and predict the four upcoming games in 2017. The four classification models are ID3, C4.5, CART, and Naïve Bayes.

Data: Each data object 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 *build* classification models. (Background color: YELLOW)

Testing set: 6 games. Please use game ID 25-30 to *evaluate* the performance of classification models. (Background color: BLUE)
Suppose "Win" is the positive label and "Lose" is the negative label. Keep this in mind when you use Precision and Recall to evaluate.

Predicting set: 4 games. Please use game ID 31-33 and 35 to *predict* the future. (Background color: GREEN)

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	?
32	10/28/17	North Carolina State	Home	Out	1-NBC	?
33	11/4/17	Wake Forest	Home	Out	1-NBC	?
34	11/11/17	Miami Florida	Away	In	?	?
35	11/18/17	Navy	Home	Out	1-NBC	?
36	11/25/17	Stanford	Away	In	?	?

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

(1) Construct a decision tree based on the training set (24 games).
(2) Use the decision tree to predict labels of instances in the testing set (6 games) based on their attributes
(3) Calculate Accuracy, Precision, Recall, and F-1 score on the testing set, given the ground-truth labels of the 6 games.
(4) Predict labels of the future 4 games using the decision tree.

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

The same as above.
(1) Construct a decision tree based on the training set (24 games).
(2) Use the decision tree to predict labels of instances in the testing set (6 games) based on their attributes
(3) Calculate Accuracy, Precision, Recall, and F-1 score on the testing set, given the ground-truth labels of the 6 games.
(4) Predict labels of the future 4 games using the decision tree.

[20'] Question 3: CART model, a decision tree model using "Δgini"

The same as above.
(1) Construct a decision tree based on the training set (24 games).
(2) Use the decision tree to predict labels of instances in the testing set (6 games) based on their attributes
(3) Calculate Accuracy, Precision, Recall, and F-1 score on the testing set, given the ground-truth labels of the 6 games.
(4) Predict labels of the future 4 games using the decision tree.

[30'] Question 4: Naïve Bayes model

(1) For each instance in the testing set (6 games), use Naïve Bayes to predict the label based on the training set (24 games).
(2) Calculate Accuracy, Precision, Recall, and F-1 score on the testing set, given the ground-truth labels of the 6 games.
(3) For each instance in the predicting set (4 games), use Naïve Bayes to predict the label based on the training set (24 games).

[5'] Given your conclusion on which of the four models is the best.

[5'] (Project follow-up) Suppose you are working on paper classification: The data object is a paper. Attributes are the words in the paper (binary, if the word is in the paper). The label is "yes"/"no" on if the paper is in the SIGKDD conference. Which models do you plan to implement/use or have you implemented/used?