Naïve Bayes Learning and Learning the KNN Classifier

Task 1 Programming and Evaluation on A Small Dataset :

Given a university's football game data for the last two seasons, please construct Naïve Bayes classification models to predict game results on games, and evaluate the model performance.

- 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.
- 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.
- Predictive labels
  - 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.

Q1: Programming (you can implement from scratch, use open-sourced code, or use machine learning platforms)**:** Use Naïve Bayes and KNN 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. This posting discusses the four measurements: https://blog.exsilio.com/all/accuracy-precision-recall-f1-score-interpretation-of-performance-measures/
Q2: Write down the prediction labels of the 12 testing games in the PDF.

Training Data:

ID	Date	Opponent	Is_Home_or_Away	Is_Opponent_in_AP25_Preseason	Media	Label
1	9/5/15	Texas	Home	Out	1-NBC	Win
2	9/12/15	Virginia	Away	Out	4-ABC	Win
3	9/19/15	GeorgiaTech	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	WakeForest	Home	Out	1-NBC	Win
11	11/21/15	BostonCollege	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	MichiganState	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	NorthCarolinaState	Away	Out	4-ABC	Lose
19	10/15/16	Stanford	Home	In	1-NBC	Lose
20	10/29/16	MiamiFlorida	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	VirginiaTech	Home	In	1-NBC	Lose
24	11/26/16	USC	Away	In	4-ABC	Lose

Testing Data

ID	Date	Opponent	Is_Home_or_Away	Is_Opponent_in_AP25_Preseason	Media	Label
25	9/2/17	Temple	Home	Out	1-NBC	Win
26	9/9/17	Georgia	Home	In	1-NBC	Lose
27	9/16/17	BostonCollege	Away	Out	2-ESPN	Win
28	9/23/17	MichiganState	Away	Out	3-FOX	Win
29	9/30/17	MiamiOhio	Home	Out	1-NBC	Win
30	10/7/17	NorthCarolina	Away	Out	4-ABC	Win
31	10/21/17	USC	Home	In	1-NBC	Win
32	10/28/17	NorthCarolinaState	Home	Out	1-NBC	Win
33	11/4/17	WakeForest	Home	Out	1-NBC	Win
34	11/11/17	MiamiFlorida	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

Task 2 Programming and Evaluation on A Large Dataset (Titanic):

Q1: Test your naïve Bayesian classification on the Titanic dataset. Report the average Accuracy, Precision, Recall, and F1 score of your five-fold cross validation. The five-folds of the Titanic data are split randomly. What do you observe and learn by applying Bayesian learning to small datasets and larger datasets?

Q2: Implement KNN classification from scratch, and evaluate how K impacts the overall accuracy of kNN on the dataset. Plot the accuracies of kNN over k, and identify the best K. You can read sample code and try to implement by yourself. Below are some sample implementations from Github for your fast references:

https://github.com/sagarmk/Knn-from-scratch
https://github.com/senavs/knn-from-scratch

https://github.com/mavaladezt/kNN-from-Scratch
https://github.com/tugot17/KNN-Algorithm-From-Scratch
https://github.com/varmichelle/KNN

Q3: According to your algorithm analysis, which machine learning model performs better, Naïve Baysian or KNN on the Titanic dataset?

Please submit a PDF report. In your report, please answer each question with your explanations, plots, results in brief. DO NOT paste your code or snapshot into the PDF. At the end of your PDF, please include a website address (e.g., Github, Dropbox, OneDrive, GoogleDrive) that can allow the TA to read your code.