Assignment 3

Learning Objectives:

1. Gain insight into trade-offs between Naïve Bayes and SVM
2. Analyze the effect of skewness of distribution of class attribute values on classification performance
3. Learn about smoothing
4. Get some programming experience (if you take the programming option)

Description:

In this assignment you will work with two data sets. One will be the Play Tennis data set we have worked with several times, and the other will be a data set constructed from chat data originally collected in Chinese, with each contribution assigned one of 17 topic codes. For the purpose of this assignment, it is not important to understand what the features represent or what the specific topics are, so the attribute names have been replaces with word1-word1000. The class attribute is called Topic, and its values are of the form c<number>-<number>.

The Topic dataset is provided in ".arff" format. The data from the Play Tennis data set is included here for your reference:

outlook	temperature	humidity	windy	play
sunny	hot	high	FALSE	no
sunny	hot	high	TRUE	no
overcast	hot	high	FALSE	yes
rainy	mild	high	FALSE	yes
rainy	cool	normal	FALSE	yes
rainy	cool	normal	TRUE	no
overcast	cool	normal	TRUE	yes
sunny	mild	high	FALSE	no
sunny	cool	normal	FALSE	yes
rainy	mild	normal	FALSE	yes
sunny	mild	normal	TRUE	yes
overcast	mild	high	TRUE	yes
overcast	hot	normal	FALSE	yes
rainy	mild	high	TRUE	no

Part 1 Step-by-Step Guide (Nonprogramming Option):

1. List all of the counts you would store for a simple Naïve Bayes model trained from the Play Tennis data set. Note that for this assignment we are adding an additional possible value for the Humidity feature, namely low, which never occurs in the training data. Compute these counts by hand and put them in a table in your answer document.

2. Now construct a second table from the same data using a form of smoothing where you simply add 1 to all counts.

3. Compute your prediction for the following test instances using Naïve Bayes applied to the two tables above so that you have a prediction with the original model and with the smoothed model, and show all of your work. Comment on the impact of zero counts on predictions. Comment on the impact of smoothing on prediction.

Outlook	Temperature	Humidity	Windy	Play
overcast	hot	normal	TRUE	
rainy	hot	high	FALSE	
overcast	cool	normal	TRUE	
rainy	mild	low	FALSE	

4. You will turn in a write up with your answers to questions 1-3.

Part 1 Step-by-Step Guide (Programming Option):

1. You should already have Python installed on your computer from Assignment 2. Download and unpack thinkbayescode.zip from blackboard. Most of this code came from the freely available code that comes with the Think Bayes book. I just added amlnaivebayes.py.
2. Read Supplementary Stat Reading 2, which you will find in the Assignment 3 folder near the thinkbayescode.zip file. This is an excerpt from Chapter 2 of the Think Bayes book, which begins to describe how Naïve Bayes works and how it is implemented in a simple way for the Cookie problem.
3. Launch Python with the working directory set to your thinkbayescode folder. Load the thinkbayes.py file. You can do this with execfile('thinkbayes.py')
4. Now, read through amlnaivebayes.py and compare it with cookie2.py, which is described in Supplementary Stat Reading 2, and which can be found in thinkbayescode/thinkbayes_code.
5. What you are going to do is compute a probability for each class value for each instance in the following table, both with a model that does not use smoothing and one that does. You'll do this by following the instructions in the comments in the amlnaivebayes.py file for modifying 3 functions, then running the file using execfile('amlnaivebayes.py') and then copying the output into a text file.

Outlook	Temperature	Humidity	Windy	Play
overcast	hot	normal	TRUE	
rainy	hot	high	FALSE	
overcast	cool	normal	TRUE	
rainy	mild	low	FALSE	

6. You will turn in your modified version of amlnaivebayes.py and your text file with the results. Also, answer the following question: What do you notice is the impact of smoothing on prediction.

Part 2 Step-by-Step Guide (Optional):

Load the Topic dataset into weka and run a cross-validation experiment using Naïve Bayes (from the bayes folder) and then one using SVM (called SMO under the functions folder). Which one performed better? Why do you think this was the case? Turn in your answers to these questions with the assignment.

Deliverables:

1. Submit your answers for Part 1 according to the specifications of the nonprogramming or programming option, depending on which you selected.
2. Optional: Submit your answers for Part 2

Miscellaneous Notes:

1. If you have not increased your heap size yet in your computer, please increase it now!
2. The experiments involving support vector machines take more than 10 minutes depending on your computer.