CSCI 554 Pattern Recognition
Fall 2019
Proj#1

Due: 9/30/2019 @ 11:59:59pm
Submission: via Blackboard

30 points mathematical expressions
30 points design (functions)
20 points (evaluation)
20 points (discussion)

Total 100 points

Outcomes

Upon completion of this project students will reinforce theoretical concepts covered in class concerning conditional distributions through reduction to practice. Additionally this project will further strengthen student experience in working through the mathematics, designing a mathematical model, and crafting a prototype implementation.

The Assignment

In class, we discussed conditional probabilities and Bayes Rule. For the assignment, students will employ a publicly available data-set in the implementation of a simple inference system that makes use of conditional probabilities and Bayes Rule.

The assignment will consist of the following tasks

1. Locate and download the "Car Evaluation Data Set" from the UCI Machine Learning Repository:
   https://archive.ics.uci.edu/ml/datasets/Car+Evaluation
2. Study the information at the data-set's information section of the web page with particular attention to the "Data Set Information" and "Attribute Information" sections.
3. Make sure you understand the data-set before you continue.
   The data-set consists of a random vector of values that appears as a matrix. Each row of the matrix corresponds to a different random vector while each column of the matrix corresponds to a different random variable whose value appears jointly with values of the other random variables. Each random variable has a range that is described in the "Attribute Information" section of the data set description.
4. Once you understand the Car Evaluation Data Set, navigate to the "Data

Folder," right click, and save the file "car.data" to your file system.

5. Rename your "cars.data" file to "cars.csv" where the extension ".csv" stands for "comma separated value" (CSV).
6. Make note of the structure of the CSV file's contents by opening in with a text editor. You will note that the data appears such that a random vector of values occupies each line of the CSV file. Each value in the random vectors takes form as a comma separated list. Each element in the comma separated list is the value for each random variable in the random vector. There are a total of 7 random variables in the random vector. Each random variable is associated with a value and therefore the random vectors each have a dimensionality of 7.
7. Note that the range for each random variable is different.

You will begin by implementing a set of counting routines that will count the number of vectors that have a particular value for each variable. For example, the first random variable in the random vectors in cars.csv is "`buyingPrice`" this variable takes on values from the set `{vhigh, high, med, low}` representing the measured outcomes for the buying price of the vehicle described by the random vector in question.

For the buying price, you are to write a routine…

```
countBuyingPrice(value, dataArray)
```

…this function takes two parameters, namely the value for r.v. buyingPrice and the array of random vectors. The function will return the number of vectors from the data array for which `buyingPrice=value`. This is essentially how the "favorable outcome is counted."

You are to implement such counting routines for each of the 7 random variables in the random vector. Once you have done this, you will implement conditioning needed to build a classifier. The counting routines will be useful in implementing a Bayesian classifier by interpreting a conditional probability using the idea of filtering we discussed in class. The conditioning you will perform will employ only a single random variable, the class label, as the favorable variable and 3 of the random variables as the condition variables.

That is…

```
P(acceptable|buyPrice,maint,safety)
```

Please see the data description to learn what these variables mean. It is your responsibility to read.

Your conditional probability routine for the classifier will use your counting routines to compute the conditional probabilities using the "filtering" approach we

discussed in lecture. For example...

```
P(acceptable| vhigh,high,low)
```

will filter out those vectors for which "buyPrice=vhigh" and "maint=high" and "safety=low." Given the resulting total set, counts are then computed for the possible favorable outcomes within the resulting set for "acceptable={unacc, accep, good, vgood}. From this you will compute the conditional probability.

You are to do this with the class variable, acceptable, as the favorable variable and 3 of variables <buyPrice, maint, safety> as the condition variables. Because the favorable variable, acceptable, consists of 4 states acceptable={unacc, accep, good, vgood}, The conditional distribution is to be evaluated for each of the states of the favorable variable. You are to report, each of these conditional probabilities, with the intent of implementing the MAP decision rule.

Test your code

1. Randomize the rows of your CSV file once read into memory
2. Set aside 10% of the rows of the randomized CSV file. The remaining 90% will be used to compute the probabilities P(acceptable | .)
3. Randomly select a vector from the 10% set aside and use values for <buyPrice, maint, safety> as condition variables.
4. Evaluate the posterior class probability for each value of the favorable variable (note this is also the class label):

   ```
   P(acceptable|buyPrice,maint,safety)
   ```

5. Display the results
6. Select the class associated with the MAP probability. Print this class.
7. Determine if the predicted MAP decision coincides with the ground truth value for acceptable. Note the ground truth value is the value for the variable, acceptable,you encountered in step#3.
8. Repeat steps 1…4 for each of the set aside 10% of the rows and display your results.
9. Discuss how your MAP decision compared to the ground truth class label.
10. You have just implemented the posterior class distribution directly using conditioning. Assuming the condition variables <buyPrice, maint, safety> are independent, using Bayes rule, product rule, and sum rule, derive an equivalent mathematical expression for the posterior class distribution.
11. Given the expression for step #10 describe how you would implement it.

Submitting Your work

1. Include all of your MATLAB files, your data files, and your answer to the question (MS-Word or PDF) in a single Zip file.