WPI Worcester Polytechnic Institute

Computer Science

CS539 Machine Learning
Homework 4 - Spring 2017

PROF. CAROLINA RUIZ

Due Date: Thursday, April 20, 2017

HW Instructions

- Carefully study Chapters 14, 15, 17, and 18 of the textbook, your class notes, and any other related materials posted on the course webpage.
- Solve each of the problems and exercises assigned in this Homework.
- **The programming assignments must be completed in Matlab** (or R if you obtained permission from the professor in advance). **You must write your own code.** *No other programming languages are allowed on this project*. Make sure to consult online documentation for Matlab. Also, my miscellaneous notes on Matlab may be useful for this project.
- **You don't need to submit a written report or homework solutions. Instead, an in-class Test will be given the day that the homework is due. This Test will evaluate your mastering of the material covered by the homework.**
- *This is meant to be an individual homework*. That is, you are expected to work on this homework on your own to make sure you know the material and know how to solve the problems, since you'll be tested individually in the Test. *Nevertheless you can discuss your questions about the homework on the Canvas' discussion forums, and consult with the professor and the TA during office hours, and with classmates if you have any trouble solving the homework problems*.

Section A: Bayesian Networks (50 points)

Dataset: For this part of the project, you will use the Adult Dataset (use the adult.data file) available at the UCI Machine Learning Repository.

- Carefully read the description provided for this dataset and familiarize yourself with the dataset as much as possible.
- For classification, use the attribute ">50K, <=50K" as the target.
- Using *stratified sampling*, split the dataset into 2 parts: *75% for training and 25% for testing*.

I. **Naive Bayes Models:**

For this part, it would be useful to look at my Matlab Naive Bayes example: diabetes_no_attribute_names.dat and naive_bayes_example_diabetis.m.

1. (10 points) Using Matlab functions, create a Naive Bayes model over the training dataset. Look at the conditional probability tables and select one that looks interesting. Include it in your report and explain why you think it is interesting.
2. (5 points) Classify the data instances in the test dataset using this Naive Bayes model. Include in your report the accuracy, precision, and recall values obtained.

II. **Bayesian Network:**

1. (10 points) Investigate what functions exist in Matlab to construct (non-Naive) Bayesian Networks. Describe those functions in your report.
2. (20 points) Using Matlab functions, create a (non-Naive) Bayesian network over the training dataset. For this, I suggest you modify function parameters until you obtain a "reasonable" graph of nodes and connections among them. Plot the graphical model obtained. Describe any interesting facts about this graphical model. Look at the conditional probability tables and select one that looks interesting. Include it in your report and explain why you think it is interesting.
3. (5 points) Classify the data instances in the test dataset using this Bayesian Network. Include in your report the accuracy, precision, and recall values obtained.

III. **Homework Problems:**

These homework problems are for you to study this topic. You do NOT need to submit your solutions.

- Chapter 14 Exercises 2, 3, 5 and 8 of the textbook (pp. 413-415).
- Given a dataset, you should be able to construct a Naive Bayes Model and calculate the conditional probability table for each node in the graph, as in the handout used during the lecture.
- Given a dataset and a graph of conditional dependencies among attributes (Bayesian network), you should be able to calculate the conditional probability table for each node in the graph, as in the handout used during the lecture.
- Given a Bayesian model (either Naive or not) and a test data instance, you should be able to determine the classification of the test instance according to the Bayesian model.
- You need to know how the K2 algorithm constructs a Bayesian model from a dataset.

Section B: Observable Markov Models and Hidden Markov Models (60 points)

For this part, it would be useful to look at my Matlab examples: hmmgenerate_fair_loaded_coins_HMMs_tutorial_example.m and mmgenerate_pepsi_coke_HMMs_tutorial_example.m

I. (5 points) Use Matlab to solve Exercise 1 of Chapter 15 (pp. 440-441).

II. (10 points) Use Matlab to solve Exercise 2 of Chapter 15 (p. 441).

III. Consider the Coke/Pepsi hidden Markov Model (HMM) used in Prof. Ruiz's example of Viterbi's, Forward, and Backward algorithms. Using the Matlab implementations of the Viterbi's, Forward, and Backward algorithms as appropriate, answer the following questions (include in your answers what algorithms and what Matlab commands you used and how you used them to solve each problem):

1. (10 points) Consider the following sequence of observables:

   PPPPCCPPPCCCPCCCCCPPPCPCP

   What is the probability that this sequence was generated by our HMM? Explain.
2. (10 points) What is the most likely sequence of hidden states that generated this sequence? Explain.
3. (10 points) Assume that the sequence is numbered stating at 1 (i.e., the first element of the sequence, "P", is at position 1). What is the most likely hidden state that generated the "C" in position 10 of the sequence? Explain.

4. (15 points) Use the HMM to generate a sequence of observables of length 2000. Then, use this generated sequence to learn the transition probabilities and the emision probabilities of a new hidden Markov model with 3 hidden states (let's forget about the "Start" state to simplify things). Compare the transition and emission probabilities of this new hidden Markov model with those of our original HMM.

IV. **Homework Problems:**
These homework problems are for you to study this topic. You do NOT need to submit your solutions.
Chapter 15 Exercises 3, 4, 5, 8, and 9 of the textbook (pp. 440-442).

Section C: Combining Multiple Models

In your course project, you are experimenting with *combining multiple models*, also called *meta-learning*. Investigate this topic using the following resources:

- Chapter 17 of the textbook and corresponding slides.
- my notes and resources on combining multiple models
- any other reliable online resources and research papers you find about this topic.

Your investigation must include in particular *Boosting, Bagging, and Stacking*, but you are encouraged to investigate other techniques you are interested in in addition to these three techniques.

I. Once that you have learned about these techniques, run experiments to see how they work. Follow these guidelines:
   - Use Matlab to run your experiments.
   - Design and run experiments. These experiments should compare the metalearning techniques you are investigating. Ideally also the experiments should compare the combination of models against individual models.
   - Analyze your results in depth.

II. **Homework Problems:**
These homework problems are for you to study this topic. You do NOT need to submit your solutions.
Chapter 17 Exercises 2, 6, and 9 of the textbook (pp. 511-513).

Section D: Reinforcement Learning

I. **Homework Problems:**
These homework problems are for you to study this topic. You do NOT need to submit your solutions.
Chapter 18 Exercises 1, 2, 3, and 4 of the textbook (pp. 542-544).
