Worcester Polytechnic Institute

Computer Science

CS539 Machine Learning
Homework 3 - Spring 2017

PROF. CAROLINA RUIZ

Due Date: Thursday, March 30th, 2017

HW Instructions

- Carefully study Chapters 9, 11, and 13 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 with advanced professor's permission). **You must write your own code.** *No other programming languages are allowed on this project*. Make sure to consult online documentation for Matlab (or for R). Also, my miscellaneous notes on Matlab (and R) 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: Trees (80 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.

I. **Classification Trees:**
For classification, use the attribute ">50K, <=50K" as the target.
   1. (20 points) Use Matlab functions to construct decision trees over the dataset using 4-fold cross-validation. Briefly describe in your report the functions that you use and their parameters. Run at least 5 different experiments varying parameter values. Repeat the same experiments but now using pruning. Show the results of all of your experiments neatly organized on a table showing parameter

values, classification accuracy, size of the tree (number of nodes and/or number of leaves), and runtime.

2. (10 points) Select the pruned tree with smallest size. Use Matlab plotting functions to depict the tree. Include the plot in your report (or at least the top levels if the tree is too large). Briefly comment on any interesting aspect of this tree.
3. (10 points) Research what the random forest technique does. Describe this technique briefly in your report, including what the inputs to this technique are, what it outputs, and how it constructs its output.
4. (10 points) Include also what Matlab function constructs random trees. Run at least 5 different experiments varying parameter values. Show the results of your experiments neatly organized on a table showing parameter values, classification accuracy, size of the random forest, and runtime.

II. **Regression Trees:**

For regression, use the attribute "education-num" as the target.

1. (20 points) Use Matlab functions to construct regression trees over the dataset using 4-fold cross-validation. Briefly describe in your report the functions that you use and their parameters. Run at least 5 different experiments varying parameter values. Repeat the same experiments but now using pruning. Show the results of all of your experiments neatly organized on a table showing parameter values, Sum of Square Errors (SSE), Root Mean Square Error (RMSE), Relative Square Error (RSE), Coeffient of Determination (R^2), size of the tree (number of nodes and/or number of leaves), and runtime.
2. (10 points) Select the pruned tree with smallest size. Use Matlab plotting functions to depict the tree. Include the plot in your report (or at least the top levels if the tree is too large). Briefly comment on any interesting aspect of this tree.

III. **Homework Problems:**

These homework problems are for you to study this topic. You do NOT need to submit your solutions. Chapter 9 Exercises 1, 2, 4, 6, 8, 9, 10 of the textbook (pp. 235-236).

Section B: Artificial Neural Networks and Deep Learning (50 points)

Dataset: For this part of the project, you will use the OptDigit Dataset 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.
- Use the following files:
  - optdigits.names
  - optdigits.tra: training dataset
  - optdigits.tes: test dataset

I. **Classification using Artificial Neural Networks (ANNs):**

Use Matlab functions to construct and train ANNs over optdigits.tra and then test them over optdigits.tes.

Topology of your Neural Net:

- Layers: I suggest that you use a 3-layer, feedforward architecture. More specifically, a net consisting of (1 input layer,) 2 hidden layers, and 1 output layer. Each node in a layer is connected to each and everyone of the nodes in the next layer, and no nodes on the same layer are connected.
- Number of nodes per layers: The input layer will have an entry for each input attribute. You will need to determine experimentally how many nodes to use in each hidden layer (I recommend to start with relatively small numbers of hidden nodes and increase the number as needed. Also, it makes sense for the first hidden layer to contain more nodes than the 2nd hidden layer). For the output

layer, you need to determine if you want to use 10 different output nodes or just one. Your training the network should be consistent with this decision.

- *However, you can experiment with other architures in addition to the one suggested here*.

Experiments:

1. (5 points) Briefly describe in your report the Matlab functions that you use and their parameters.
2. (5 points) Explain also how many nodes you use on the output layer, and how you use the output from the output node(s) to assign a classification label to a test instance.
3. (35 points) Run at least 10 different experiments varying parameter values. Show the results of all of your experiments neatly organized on a table showing parameter values, number of hidden nodes in each layer, classification accuracy, and runtime.
4. (5 points) Pick the experiment that you think produced the best result. Justify your choice in your report. Include the confusion matrix for this experiment. See what misclassifications are most common and elaborate on your observations.

II. **Deep Learning:**

1. Read the following article: Yann LeCun, Yoshua Bengio, Geoffrey Hinton. "Deep learning". Nature 521, 436-444 (28 May 2015) doi:10.1038/nature14539.
2. Watch one of the following videos about deep learning (if you have time try to watch both). You're not expected to understand all the details, but try to get from the videos some of the theoretical foundations of deep learning and some of its applications.
   1. "Deep Learning" by Ruslan Salakhutdinov from the collection of Deep Learning Summer School, Montreal 2015
   2. "Recent developments on Deep Learning" Geoffrey Hinton's GoogleTech Talk, March 2010.

   Links to both videos (and several others) are available at deeplearning.net tutorials

III. **Homework Problems:**

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

- Chapter 11 Exercises 1, 2, 3, 6, 12 of the textbook (pp. 311-313).
- Study *convolutional neural networks* (section 11.8.3), *autoencoders* (section 11.11), and *recurrent networks* (section 11.12.2). You should know the topology (= structure) of these nets and the procedure to train them.

Section C: Support Vector Machines (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.

I. **Classification using Support Vector Machines (SVMs):**

For classification, use the attribute ">50K, <=50K" as the target.

1. (9 points) Use Matlab functions to construct a support vector machine over the dataset using 4-fold cross-validation. Briefly describe in your report the functions that you use and their parameters.
2. (36 points) Run at least 12 different experiments varying parameter values for each of the following kernel functions (run at least 4 experiments for each one of the 3 kernel functions required):
   - polynomial (including linear, quadratic, ...)
   - radial-basis functions (Gaussian)
   - sigmoid (tanh)

   Show the results of all of your experiments neatly organized on a table showing kernel function used, parameter values, classification accuracy, and runtime.
3. (5 points) Pick the experiment that you think produced the best result. Justify your choice in your report. Use Matlab functionality to plot a 2 or 3 dimensional depiction of data instances in each of the two classes, support vectors, and the decision boundary.

II. **Homework Problems:**

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

Chapter 13 Exercises 1 and 2 of the textbook (pp. 382-383).