PROEA-821: Machine Learning Spring 2018

Homework 1

Handed out: March 5, 2018
Due date: March 19, 2018

General Instructions

- You are welcome to talk to other members of the class about the homework. I am more concerned that you understand the underlying concepts. However, you should write down your own solution. Please keep the class collaboration policy in mind.
- Feel free discuss the homework with the instructor or the TAs.
- Handwritten solutions will not be accepted.
- The homework is due by midnight of the due date. Please submit the homework on Canvas.
- You should submit two files: a separate pdf or word file for the homework report, and a zip file containing everything else, e.g. code, data, etc.

1 Experiments

In this question you will be implementing a decision tree learner. You will experiment with the decision tree hyperparameters using cross-validation.

There is a secret computer science conference that only a selected group of computer scientists will be invited. Fortunately, we have a secret agent that has access to the guest list. But, she only has part of it. The accessible name list is in the **Dataset/training.data** file. Your job is to use this training data to learn a decision tree classifier that can predict if the names in the **Dataset/test.data** file will be invited to the secret conference. We suggest some features for the dataset, but you need to extract the features yourself. You are also welcome to add your own features.

Cross-Validation

The depth of the tree is a hyper-parameter to the decision tree algorithm that helps reduce overfitting. You will see later in the semester that many machine learning algorithm (SVM, logistic-regression etc) have some hyper-parameters as their input. One way to determine a proper value for the hyper-parameter is to use a technique called cross-validation.

As usual we have a training set and a test set. Our goal is to discover good hyper-parameters using the training set. To do so, you can put aside some of the training data aside, and when training is finished, you can test the resulting classifier on the held out data.

This allows you to get an idea of how well the particular choice of hyper-parameters does. However, since you did not train on your whole dataset you may have introduced a statistical bias in the classifier. To correct for this, you will need to train many classifiers with different subsets of the training data removed and average out the accuracy across these trials.

For problems with small data sets, a popular method is the leave-one-out approach. For each example, a classifier is trained on the rest of the data and the chosen example is then evaluated. The performance of the classifier is the average accuracy on all the examples. The downside to this method is for a data set with n examples you must train n different classifiers. Of course, this is not practical for the data set you will use in this problem, so you will hold out subsets of the data many times instead.

Specifically, for this problem, you should implement k-fold cross validation. The general approach for k-fold cross validation is the following: Suppose you want to evaluate how good a particular hyper-parameter is. You split the training data into k parts. Now, you will train your model on $k-1$ parts with the chosen hyper-parameter and evaluate the trained model on the remaining part. You should repeat this k times, choosing a different part for evaluation each time. This will give you k values of accuracy. Their average cross-validation accuracy gives you an idea of how good this choice of the hyper-parameter is. To find the best value of the hyper-parameter, you will need to repeat this procedure for different choices of the hyper-parameter. Once you find the best value of the hyper-parameter, use the value to retrain you classifier using the entire training set.

1. [20 points] **Implementation**

   For this problem, your will be using the data in **Dataset** folder. This folder contains two files: **training.data** and **test.data**. You will train your algorithm on the training file. Remember that you should not look at or use your testing file until your training is complete.

   (a) [8 points] Implement the decision tree data structure and the ID3 algorithm for your decision tree (Remember that the decision tree need not be a binary tree!). Discuss what approaches or choices you had to make during this implementation.

   (b) [4 points] Suggest at least 4 other features you could have extracted from this dataset.

   (c) [2 points] Report the error of your decision tree on the **Dataset/training.data** file.

   (d) [5 points] Report the error of your decision tree on the **Dataset/test.data** file.

   (e) [1 points] Report the maximum depth of your decision tree.

2. [20 points] **Limiting Depth**

   In this section, you will be using 4-fold cross-validation in order to limit the depth of your decision tree, effectively pruning the tree to avoid overfitting. You will be using the 4 cross-validation files for this section, titled Dataset/CVSplits/training 0X.data where X is a number between 0 and 3 (inclusive)

(a) [10 points] Run 4-fold cross-validation using the specified files. Experiment with depths in the set $\{1, 2, 3, 4, 5, 10, 15, 20\}$, reporting the cross-validation accuracy and standard deviation for each depth. Explicity specify which depth should be chosen as the best, and explain why.

(b) [5 points] Using the depth with the greatest cross-validation accuracy from your experiments: train your decision tree on the **Dataset/training.data** file. Report the accuracy of your decision tree on the **Dataset/test.data** file.

(c) [5 points] Discuss the performance of the depth limited tree as compared to the full decision tree. Do you think limiting depth is a good idea? Why?

Experiment Submission Guidelines

1. The report should detail your experiments. For each step, explain in no more than a paragraph or so how your implementation works. You may provide the results for the final step as a table or a graph.

2. You should include a shell script, `run.sh`, that will execute your source code. Your code should produce similar output to what you include in your report. It is suggested (but not required) to include a ReadMe file describing how to run your code.

3. Please do not hand in binary files! We will *not* grade binary submissions.