Assignment #5: Experiments and Technical Writing
CSCI 374 Fall 2022 Oberlin College
Due: Wednesday November 23 at 11:59 PM

Background

Our final assignment this semester has three main goals:

1. Practice the process of scientific experimentation with machine learning
2. Gain experience utilizing pre-implemented solutions from available APIs
3. Work on our technical writing skills in the context of computer science

Students may work in groups of up to two members for this assignment.

Gitting Started

To begin this assignment, you please follow this link:

https://classroom.github.com/a/jeovOqPm

Data Set

For this assignment, we will learn from four pre-defined data sets:

1. **monks1.csv**: A data set describing two classes of robots using all nominal attributes and a binary label. This data set has a simple rule set for determining the label: `if head_shape = body_shape OR jacket_color = red, then yes, else no`. Each of the attributes in the monks1 data set are nominal. Monks1 was one of the first machine learning challenge problems (http://www.mli.gmu.edu/papers/91-95/91-28.pdf). This data set comes from the UCI Machine Learning Repository: http://archive.ics.uci.edu/ml/datasets/MONK%27s+Problems
2. **mnist_100.csv**: A data set of optical character recognition of numeric digits from images. Each instance represents a different grayscale 28x28 pixel image of a handwritten numeric digit (from 0 through 9). The attributes are the intensity values of the 784 pixels. Each attribute is ordinal (treat them as continuous for the purpose of this assignment) and a nominal label. This version of MNIST contains 100 instances of each handwritten numeric digit, randomly sampled from the original training data for MNIST. The overall MNIST data set is one of the main benchmarks in machine learning: http://yann.lecun.com/exdb/mnist/. It was converted to CSV file using the python code provided at: https://quickgrid.blogspot.com/2017/05/Converting-MNIST-Handwritten-Digits-Dataset-into-CSV-with-Sorting-and-Extracting-Labels-and-Features-into-Different-CSV-using-Python.html
3. **votes.csv**: A data set describing the voting histories of members of the U.S. Congress. The objective is to predict a member's political party based on their previous votes. All attributes are nominal, although some values are missing (indicated by a "?" character). This data comes from Weka 3.8: http://www.cs.waikato.ac.nz/ml/weka/

4. **hypothyroid.csv**: A data set describing patient health data using a mix of nominal and continuous attributes that can be used to diagnose the health of a patient's thyroid into four possible labels. This data set comes from Weka 3.8: http://www.cs.waikato.ac.nz/ml/weka/

Experiments

Your assignment is to conduct experiments comparing different supervised machine learning approaches on the four data sets described above. In particular, you should:

1) Write a program that takes in a single parameter (a random seed). You can call this program whatever you like.
2) Your program should read in each of the four data sets, then:
    a. preprocess the categorical attributes into one-hot encodings
    b. shuffle the data sets so that they are randomly ordered (based on your random seed)
    c. split each data set so that X% are used for training and 100-X% are used for testing (you can choose your own X)
3) Your program should train models for three different supervised learning approaches on each training set (for a total of 12 models = 3 approaches x 4 training sets)
    a. Two of your approaches should be Decision Trees and Neural Networks
    b. Your third approach should be either Random Forests or Naïve Bayes (you can use both if you wish)
    c. You are encouraged to use pre-existing implementations of these approaches for your experiments (e.g., scikit-learn in Python, Weka in Java), although you are welcome to use your own implementations, if you prefer
4) You program should evaluate your learned models on each test set. You should save the resulting confusion matrices to file, where the output file names follow the pattern results-<Approach>-<DataSet>_<Seed>.csv (e.g., results-NeuralNetworks-votes-12345.csv).
    a. The format of these files should be the same as in homeworks 1-4
5) From your confusion matrices, you should calculate the accuracy of each model and the recall of each label for each model.
    a. For each dataset, you should also calculate a confidence interval for the accuracy of each supervised learning approach and compare the confidence intervals to see if any overlap. The ultimate goal of our experiments is to see when the approaches statistically significantly outperform each other.
        i. Since we are comparing more than two types of approaches, we need to use a slightly different $Z_{1-\frac{\alpha}{2}}$ value in the confidence intervals: instead of 1.96, use 2.24 if you are comparing three different approaches (and 2.39 if

you are comparing all four approaches). This adjustment is called the Bonferroni correction – we need to have slightly wider intervals (corresponding to slightly smaller α values to account for how many comparisons we are making to reduce the risk of concluding there is a statistically significant difference when there is actually none).

Technical Report

You should document the results of your experiments in a technical report, which is good practice for the final group report due next month. Your grade for this assignment is based primarily on your technical report.

Technical reports commonly have X sections (the abstract is not numbered in your report):

0) **Abstract**: 100-200 words summarizing the entire report that quickly help the reader know what you will be discussing.
1) **Introduction**: a short overview of the entire report that should motivated the reader to be interested in what you will be discussing
    a. We typically start with one paragraph describing the "bigger picture" – what is the greater context of your experiment or project? Here, the "bigger picture" might be the ability of supervised machine learning to automate tasks for people
    b. The second paragraph is often the focus of the experiment or project – inside that bigger picture, what is the specific problem you aiming to solve? Here, that might be evaluating machine learning on four real-world data sets to gain an understanding of when one solution approach might be better than another.
    c. The third paragraph is often a short summary of your solution – how are going to solve the problem from the previous paragraph? Here, we are evaluating three (or four) popular supervised machine learning approaches.
    d. The remainder of the introduction is a summary of what experiments you have run, followed by a brief sneak peek at the results.
2) **Background and Related Work**: A description of any details the reader might need to understand your problem or solution, as well as a summary of what work others have done to either (i) solve the same problem with different solutions, or (ii) used the same solution to solve other problems.
    a. You do *not* need this section in your report for this assignment, but you will need one in your final group reports.
3) **Problem**: A description of what problem you are trying to solve or what you aim to study with your experiments. This should be more elaborate than the second paragraph of the introduction and should give the reader enough detail to know how a problem they are interested in might related to the one that you are studying.
    a. For this assignment, your problem is that we have four different real-world data sets and we want to learn how to automatically make predictions for new instances of that real-world data.

i. Your section would then include a description of each data set and what type of real-world problem they represent.

ii. You should also include a table that provides summary statistics for each data set, such as the number of attributes, the number of possible labels, and the proportion of each label in the data set.

4) **Solution**: A description of your approach to solving the problem from the previous section. Again, this will be more elaborate than the third paragraph of the introduction. You should give the reader enough details so that they (i) know how your solution works, and (ii) could replicate your solution if they had a similar problem.

    a. For this assignment, we have three solutions – the three supervised machine learning approaches you've used in your experiments (Decision Trees, Neural Networks, and either Random Forests or Naïve Bayes). For each approach, you should provide a paragraph summarizing how the method works.

5) **Experimental Setup**: A description of the design of your experiments. This should provide everything a reader needs to know what you've done, as well as all the parameters they would need to recreate your experiments on their own.

    a. For this assignment, your details will include:

        i. What performance measures you use to evaluate your solutions (a description of accuracy and recall)

        ii. Where your solution implementations come from (e.g., scikit-learn, Weka)

        iii. What your training/test percentages were

        iv. What your random seeds were

        v. Any hyperparameters used in your solutions (number of neurons, number of hidden layers, learning rate, number of trees, number of attributes in each tree, etc.)

6) **Results**: A discussion of the results you found in your experiments. We typically include one subsection for each data set that we look at, and a final subsection summarizing the results across all data sets.

    a. When you present your results, you are really making an argument for a conclusion you think should be drawn from your results. You do so by writing a POEM:

        i. Presentation: show the reader your results (usually in a table and/or chart)

        ii. Observation: point out to the reader the key results you want them to see in your presentation

        iii. Explanation: answer the question of "why did these observations occur?"

        iv. Meaning: explain to the reader the main take-away message they should have – what is the implication of your results

    b. For this assignment, the primary results are a comparison of the accuracies across the different data sets. Our research question is "did any of the machine learning

approaches statistically significantly outperform the others on each data set?" If one method does outperform the others, we want to try to explain why and see what kind of general conclusions we can draw from the results. If no single approach is best, then again an explanation for why is important.

   i. You should also present the recalls of each algorithm on each label for each data set in a table, but you do not have to compare these with statistical tests

c. For this assignment, you should have five subsections – one for each data set (each only 1 or 2 paragraphs long) and one summarizing the results across all data sets (again, 1 or 2 paragraphs).

   i. In the final subsection, you should create a table showing how many times each approach (i) statistically significantly outperformed another approach, (ii) did statistically significantly worse than another approach, and (iii) had statistically similar results (neither was better than another). From this table, you should try to conclude how we might choose one of the methods for a new data set.

7) **Conclusion and Future Work**: typically one paragraph summarizing the paper (similar to the abstract), briefly recapping the problem, solution, and the key findings of your results, and a second paragraph listing ideas you have for future work of what someone might do next that builds on what you've accomplished in this technical report. Do you have ideas for new experiments (new data sets or new solutions)?

Altogether, your report should probably be 4-6 pages long, but length with vary based on your discussions (I'm looking for content and not length).

README

Within a README file, you should include:

1) Who were the members of your group (so that I know who to give credit to!)
2) A short paragraph describing your experience during the assignment (what did you enjoy, what was difficult, etc.)
3) An estimation of how much time you spent on the assignment, and
4) An affirmation that you adhered to the honor code

Please remember to commit your technical report (as a Word document), solution code, results files, and README file to your repository on GitHub. You do not need to wait to commit everything until you are done with the assignment; it is good practice to do so after hitting important milestones or solving bugs during a coding session. ***Make sure to document your code***, explaining how you implemented the different components of the assignment.

Honor Code

Each student is allowed to work with a partner to complete this assignment. Groups are also allowed to collaborate with one another to discuss the abstract design and processes of their

implementations. For example, please feel free to discuss how to use the API you chose to work with, or how you wrote up a particular part of the report. However, sharing code or writing between groups is not permitted.

Grading Rubric

Your solution and README will be graded based on the following rubric:

Correctly implemented the experiments: /30 points
Correctly wrote the technical report: /65 points
Provided requested README information: /5 points