COSC 311 - Project 2

Dr. Joe Anderson

Due: 17 December

1 Timeline

We will adhere to the following sub-due dates to keep groups and projects on track and organized. For each one, upload any supporting documents/code to the assignment submission page on MyClasses.

1. **24 November**: A short description of your planned project data, topics, and approach. Summarize the algorithms you plan to implement, and any additional supporting or motivating scratch work to explain your choices. What algorithms will you implement them? Why would the be successful on your data? How will you try visualizing the data in illustrative/creative ways? What conclusions do you think one can draw? Why or why not?
2. **8 December**: A followup summary to your plan, along with preliminary visualizations and experiments. What is going right or wrong, what are you next steps or any adjustments?
3. **17 December**: Submit the final report, code, slides, visualizations. You will present to the class during the scheduled final exam.

2 Description

In this project, you will work to 1) implement and understand several machine-learning algorithms and 2) present the data that you analyzed in Project 1, and show the effectiveness of the algorithms on this data. You can focus on one of the three datasets that you analyzed for Project 1. You may work in the same groups or prepare different individual presentations/analyses of your data.

1. Effectively summarize the data: what is it? what does it represent? How was it gathered?
2. What "classes" are present within the data? Just by looking at some primitive plots or graphical breakdowns, are there features that "give away" which class a sample belongs to? For instance, in the adults dataset, the classes are whether the person makes more or less than $50k; a relevant question to ask is that, since this doesn't correlate with any single attribute, perhaps there is a specific pairing of attributes for which a certain combination of values means that somebody is at a particular income level.
3. What attributes look like they are parameters of an underlying population that could be learned by an unsupervised machine learning algorithm? For example, in the adults dataset, does it seem like any of the numerical parameters follow a specific distribution within different sub-populations? One example might be that if you know a person's occupation, education, and income, can you effectively model their hours worked per week?
4. Implement your own version (not using pre-built libraries) of two supervised machine learning algorithms. Some suggestions include but are not limited to:

(a) Decision trees

(b) Support vector machine

(c) Nearest neighbors

(d) AdaBoost or other boosting (using your other algorithm as the weak learner)

(e) Standard neural network

5. Show a visualization of *each* feature in the dataset.

   (a) For **at least two** of these, use a non-standard graphic (outside the standard bar, line, scatter plot) to represent and summarize it. Be creative! Think about visualizations the rely on maps or external structures to present spacial relationships, or perhaps an animated visual of some feature/structure changing over time (`matplotlib.animation` can help here).

   (b) The visualization should emphasize both the values of the feature and their relationship to other aspects such as membership in a particular population within the data.

6. Show the behavior of your machine learning algorithms on the data. Specifically, consider the following:

   (a) Show the "learning curve" of your algorithm: how the error changes with more training (either more training data or more epochs, depending on the algorithm). Include the performance on both testing and training datasets during this process, when applicable.

   (b) Show the generalization error by using different sets of data for testing, training, and validation.

   (c) How does the algorithm perform when you restrict the dimension of your data? This may be done by removing columns, or by using other dimension reduction algorithms (Principle Component Analysis, Multi-dimensional scaling, etc.). If you can get good behavior using only 2 or 3 features, try to visualize the algorithms results on the full domain, to show the "decision surface".

   (d) Use more "classical" statistic techniques to motivate the success (or failure) of your machine learning algorithm: are the features uncorrolated with the label? Do hypothesis tests reject assumptions that would make prediction possible?

7. Finally, you will present this data and your results to your classmates on the final exam date, December 17. The presentation should be 10-15 minutes long, and will take the place of the final exam grade. Your presentation should focus on:

   (a) Creative ways to visualize the data, and the results/reasoning behind your machine learning algorithms,

   (b) Clearly explaining why your results make sense in a boarder context, and whether they inform new information that is not directly captured by the dataset,

   (c) Providing "next steps" to somebody with more domain knowledge, that might lead to better or more meaningful outcomes.

3 Submission

Use a Jupyter notebook for each one of your datasets, answering the questions with Markdown or Text cells.

Upload all responses, source files, and documentation to the course MyClasses assignment in a `.zip` or `.gz` archive.