EE 518: Machine Learning Theory & Algorithms **Spring 2024**

Homework 4
Due: May 10, 2024, 11:59PM PT

Student Name: *Instructor Name: John Lipor*

Problem 1 Multiclass AdaBoost (5 pts, 3 pts, 3 pts)

The AdaBoost algorithm implemented last week is only capable of performing binary classification. In this problem, you will see how a binary classifier can be used to perform multiclass classification. For cleaner code, you may wish to package your AdaBoost classifier into its own class, as with the decision stump.

(a) Extend your AdaBoost implementation from Homework 3 to multiclass classification by following the One-versus-All approach described in Section 17.1 of UML. Test your code on the synthetic data generated in `prob1a.py` using $T = 500$ rounds of boosting. **Turn in** your code and a plot of the training and test error versus number of boosting rounds. A logarithmic y-axis may be helpful to see the resulting behavior.

(b) Generate the same plot as for part (a) using XGBoost with `max_depth` set to 1. What similarities or differences do you notice?

(c) Apply your multiclass AdaBoost to the MNIST dataset. **Turn in** your training and test error, as well as the number of boosting rounds you performed and why you chose this value.

Problem 2 DSS: Visualizing with Random Forests (5 pts each)

(*DSS rules apply.*) Aside from their use in classification and regression, decision trees can also be used to understand the similarity between datapoints, even when the features have different scales and types. This is important, since a good notion of similarity can be used to produce meaningful embeddings via algorithms such as UMAP. First, read the articles here and here.

(a) Apply the technique described in the articles to perform supervised dimensionality reduction/visualization with UMAP on the MNIST **test** dataset (you may choose to do the full 70,000-digit dataset if time allows). **Turn in** your plot of data embedded into two dimensions, as well as a plot of the UMAP embedding on the raw data (*without* random forest features, as in Homework 2, Problem 2).

(b) Apply the technique described in the articles to perform supervised dimensionality reduction/visualization with UMAP on the Titanic dataset from MP1. Note that you may wish to use the raw data after filling in missing entries instead of the feature-engineered data you used in your project. **Turn in** your plot of data embedded into two dimensions, as well as a plot of the UMAP embedding on the raw data.

(c) Describe the procedure you used to generate these embeddings *in your own words* (i.e., do not copy text from the articles). Your description should be such that you could have understood it prior to taking this course.

Problem 3 Information Gain (5 pts)

UML mentions that concave measures of information gain for decision trees are preferred. Show that if $C(a)$ is concave, then the `Gain` function is nonnegative. This indicates that splitting a node will not ever decrease the measures of gain defined in the text.

Problem 4 SLT (10 pts)

(*SLT rules apply.*) UML, Ch. 10, Exercise 4.2. Note that the boook has a typo, and the upper bound should be

$$\mathrm{VCdim}(B_d) \leq 16 + 2\log_2(d).$$

State how long you worked on the problem before looking at the solution.

Problem 5 SLT (10 pts)

(*SLT rules apply.*) UML, Ch. 18, Exercise 1. State how long you worked on the problem before looking at the solution.