DSC 140A - Homework 06

Due: Wednesday, February 22

Problem 1.

This week's homework problem is an opportunity to review the machine learning methods we have seen so far.

Data. At the link below is a data set of over 24,000 data points in $\mathbb{R}^6$ that is suitable for binary classification:

`https://f000.backblazeb2.com/file/jeldridge-data/004-competition/train-data.csv`

The file has seven columns: the first six contain the features, while the last column contains the label (either zero or one). Note that there are no column headers – the first row of the file contains data.

At the next link is a *test* data set of over 8,000 data points in $\mathbb{R}^6$:

`https://f000.backblazeb2.com/file/jeldridge-data/004-competition/test-data.csv`

Note that this file has only 6 columns – the label is missing!

Task. Implement a machine learning algorithm – any algorithm – that has been discussed in DSC 140A so far. Train the algorithm on the training set, and make predictions (1 or 0) on the test set.

You can use whatever programming language you like, but you should not use any machine learning packages, such as `sklearn`. Instead, you should implement the method "from scratch". You may use numerical packages, such as `numpy`.

Submitting your results. There are two Gradescope submissions for this assignment. First, there is "Homework 06 - Predictions". Upload your predictions to this submission' autograder as a file named `predictions.csv` containing one prediction (1 or 0) per line. The autograder will check the format of your submission for correctness, so be sure to wait to see its output.

Second, there is "Homework 06 - Code". Upload your code to this submission as a PDF; we will look at your code, but not run it.

Grading. Recall that the test set does not have labels – we have kept them a secret. At grading time, we will test your predictions against the true labels. To get full credit, at least 60% of your predictions should be correct.

Note: the training and test set are unbalanced (there are more instances of one class than another), but we will test your predictions on a *balanced* subset of the test set; to get full credit, your predictions must achieve 60% accuracy on this balanced subset.

Competition. For fun, we'll give out extra credit to people who achieve the top prediction performance on the unseen data. Overall, this assignment is worth 5 points. However, if your prediction performance is in one of the categories below, you'll earn the stated amount of extra credit:

- **Accuracy above 65%**: 1 point
- **Accuracy above 70%**: 2 points
- **Top 10% of submitted accuracies:** 3 points
- **Top accuracy:** 6 points

Hint: Some methods are easier to implement than others. You should be able to achieve full credit by implementing simpler methods.