CSE 151: Introduction to Machine Learning Summer 2019, Session 2

Programming Assignment 3

Instructor: Joseph Geumlek **Due on:** Aug 29, 2019

Instructions

- This is a 20 point homework. The assignment should be done individually.
- You are free to use any programming language that you wish.
- The programming assignment should be submitted as a single pdf file through Gradescope. Please enter your answers first, followed by the code.

Problem 1: Programming Assignment: 20 points

In this problem, we look at the task of classifying by topic posts made in two different internet newsgroups – comp.windows.x and rec.sport.baseball, that correspond to labels 1 and 2 respectively.

For your convenience, we have already pre-processed the posts and converted them to feature vectors, where each feature or coordinate corresponds to the count of a single word. Download the files `pa3train.txt` and `pa3test.txt` from the class website. These files contain your training and test data sets respectively. Each line of the training or test set is a feature vector of length 819, followed by a label 1 or 2.

A dictionary is also provided in the file `pa3dictionary.txt`; the first line in the dictionary is the word that corresponds to the first coordinate, the second line to the second coordinate, and so on.

1. First, we will learn a linear classifier that can predict if a post belongs to class 1 or class 2. For this purpose, your training data is in `pa3train.txt`, and your test data is in `pa3test.txt`.

   Assume that the data is linearly separable by a hyperplane through the origin. Run two, three, four, and five passes of a perceptron on the training dataset to find classifiers that separate the two classes. What are the training errors and the test errors of perceptron after two, three, four and five passes? [Hint: If your code is correct, the training error after a single pass of a perceptron would be about 0.04.]

2. **Interpreting a model:** Consider the weights w that you built by running five passes on the data. We will now try to interpret this classifier. Find the three coordinates in w with the highest and the three coordinates in w with the lowest values. What are the words (from `pa3dictionary.txt`) that correspond to these coordinates? The three highest coordinates are those words whose presence indicates the positive class most strongly, and the three lowest coordinates are those words whose presence indicates the negative class most strongly. Does this interpretation match up with what you might expect for these two newsgroups?

3. For the third part of the question, we will kernelize our perceptron, using two different kernels, an exponential kernel and a polynomial kernel:

$$K_{exp}(x, z) = e^{-||x-z||/20}$$

$$K_{poly}(x, z) = (\langle x, z \rangle + 10)^2$$

   For each kernel, K_{exp} and K_{poly}, perform one, two, three, four, and five passes of a kernelized perceptron. Report the training error and test error after each pass for both kernels. [Hint: if your code is correct, and you provide it with the kernel $K(x, z) = \langle x, z \rangle$, you should get the same results as your non-kernelized perceptron.]