**CSE 151: Introduction to Machine Learning** Spring 2018

# Programming Assignment 4

*Instructor: Kamalika Chaudhuri* **Due on:** Jun 1, 2018

## 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, containing the answers to the questions first, followed by the code. Please submit both files through Gradescope.

## Problem 1: Programming Assignment: 20 points

In this problem, we will look at classifying protein sequences according to whether they belong to a particular protein family or not. For this task, we will use the string kernel that we discussed in class, as well as a modified version of this kernel. Download the files `pa4train.txt` and `pa4test.txt` from the class website. These files contain your training and test data sets respectively.

The data files are in ASCII text format, and each line of the file contains a string, which represents a protein sequence, followed by a label, which is 1 or $-1$, to indicate whether the protein sequence belongs to a protein family or not. Each letter in the protein sequence represents an amino acid, and thus the alphabet size is $|\Sigma| = 21$ (20 amino acids + a symbol to represent missing data). Different protein sequences in the file have different length; this is not surprising because even the same protein will have different lengths in different species, for example, in mouse and human. Assume that the data is linearly separable by a hyperplane through the origin. Run a single pass of kernel perceptron algorithm on the training dataset to find a classifier that separates the two classes.

1. First, we will use the string kernel function for our kernel. Recall from class that given two strings $s$ and $t$, the string kernel $K_p(s,t)$ is the number of substrings of length $p$ that are common to both $s$ and $t$, where a string that occurs $a$ times in $s$ and $b$ times in $t$ is counted $ab$ times.

   For this problem, use $p = 3$, $p = 4$ and $p = 5$. Write down the training and test errors of kernel perceptron for $p = 3, 4, 5$ on this dataset.

   [Hint: If your code is correct, the training error for $p = 2$ will be about 0.0711.]

2. Next, repeat Part (1) with a slight modification of the string kernel, $M_p(s,t)$. Given two strings $s$ and $t$, the modified string kernel $M_p(s,t)$ is the number of substrings of length $p$ that are common to both $s$ and $t$, where a string that occurs $a$ times in $s$ and $b$ times in $t$ *is counted only once.* What are the training and test errors for this kernel for $p = 3, 4, 5$?

3. Finally, we will try to interpret the classifier that we built. For this, consider the kernel perceptron classifier $w$ from part (1) for $p = 5$. This classifier can be written in the form: $w = \sum_i^n \alpha_i \phi(x_i)$, where $x_i$-s are the training data points, and $\phi$ is the feature map corresponding to the string kernel. Recall from lecture that $\phi$ has $21^5$ coordinates, where each coordinate corresponds to a substring of size 5 on the alphabet $\Sigma$.

   Find the two coordinates in $w$ with the highest positive values. You should be able to do this without explicitly computing *all* the coordinates of $w$. What are the substrings corresponding to these coordinates? These coordinates correspond to those substrings whose presence most strongly indicates that the protein belongs in the family.