CSE 151: Introduction to Machine Learning Spring 2015

Problem Set 5

Instructor: Kamalika Chaudhuri **Due on:** May 29, 2015

Instructions

- This is a 40 point homework. For Problem 1, Parts 1-2 are worth 2 points each, and Parts 3-4 are worth 3 points each. For Problem 2, Parts 1-4 are worth 2 points each, and Parts 5-8 are worth 3 points each.
- Problem 3 is a programming assignment. For this problem, you are free to use any programming language you wish.
- To submit your code, please send email your code to cse151homeworks@gmail.com. **Please submit your solution with the homework, and email only the code to this address.**

Problem 1: 10 points

In the following problems, suppose that K, K_1 and K_2 are kernels with feature maps ϕ, ϕ^1 and ϕ^2. For the following functions $K'(x, z)$, state if they are kernels or not. If they are kernels, write down the corresponding feature map, in terms of ϕ, ϕ^1, ϕ^2 and c, c_1, c_2. If they are not kernels, prove that they are not.

1. $K'(x, z) = cK(x, z)$, for $c > 0$.
2. $K'(x, z) = cK(x, z)$, where $c < 0$, and there exists some x for which $K(x, x) > 0$.
3. $K'(x, z) = c_1K_1(x, z) + c_2K_2(x, z)$ for $c_1, c_2 > 0$.
4. $K'(x, z) = K_1(x, z)K_2(x, z)$.

Problem 2: 20 points

For the following functions $K(x, z)$, state if it is a kernel or not. If the function is a kernel, then write down its feature map. If it is not a kernel, prove that it is not one. For your proof, you can use the answers to Problem 1.

1. $x = [x_1, x_2]$, $z = [z_1, z_2]$, x_1, x_2, z_1, z_2 are real numbers. $K(x, z) = x_1z_2$.
2. Let $x = [x_1, \ldots, x_d]$, $z = [z_1, \ldots, z_d]$, x_is and z_is are real numbers. $K(x, z) = 1 - \langle x, z\rangle$.
3. $x = [x_1, \ldots, x_d]$, $z = [z_1, \ldots, z_d]$, x_is and z_is are real numbers. $K(x, z) = \|x - z\|^2$.
4. $x = [x_1, \ldots, x_d]$, $z = [z_1, \ldots, z_d]$, and f is a function. $K(x, z) = f(x_1, x_2)f(z_1, z_2)$.
5. $x = [x_1, \ldots, x_d]$, $z = [z_1, \ldots, z_d]$, x_is and z_is are real numbers. $K(x, z) = \frac{1-\langle x,z\rangle^2}{1-\langle x,z\rangle}$.
6. $x = [x_1, \ldots, x_d]$, $z = [z_1, \ldots, z_d]$, x_is and z_is are integers between 0 and 100. $K(x, z) = \sum_{i=1}^{d} \min(x_i, z_i)$.
7. $x = [x_1, \ldots, x_d]$, $z = [z_1, \ldots, z_d]$, x_is and z_is are real numbers.
$$K(x, z) = (1 + x_1z_1)(1 + x_2z_2) \ldots (1 + x_dz_d)$$
8. $x = [x_1, \ldots, x_d]$, $z = [z_1, \ldots, z_d]$, x_is and z_is are integers between 0 and 100. $K(x, z) = \sum_{i=1}^{d} \max(x_i, z_i)$.

Problem 3: Programming Assignment: 10 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. Download the files `hw5train.txt` and `hw5test.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 20. 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. For your kernel, use the string kernel function. 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. For this problem, use $p = 3$ and $p = 4$. Write down the training and test errors of kernel perceptron for $p = 3$ and $p = 4$ on this dataset.