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. Please email your code to cse151homeworks@gmail.com

Problem 1: 8 points

In the following problems, suppose that $K$, $K_1$ and $K_2$ are kernels with feature maps $\phi$, $\phi^1$ and $\phi^2$. 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 $\phi, \phi^1, \phi^2$ 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. $K'(x, z) = c_1 K_1(x, z) + c_2 K_2(x, z)$ for $c_1, c_2 > 0$.
4. $K'(x, z) = K_1(x, z) K_2(x, z)$.

Solution

1. Suppose $K(x, z) = \langle \phi(x), \phi(z) \rangle$ for some feature map $\phi$, and let $\phi'(x) = \sqrt{c}\phi(x)$. Then, for all $x$ and $z$,
   \[ K'(x, z) = cK(x, z) = c\langle \phi(x), \phi(z) \rangle = \langle \sqrt{c}\phi(x), \sqrt{c}\phi(z) \rangle \]
   Therefore $K'(x, z)$ is a kernel corresponding to the feature map $\phi'$.

2. Suppose $x_0$ is the $x$ for which $K(x, x) > 0$. Consider the $1 \times 1$ kernel matrix $K' = K'(x_0, x_0)$ for the kernel $K'$ and the data point $x_0$. Then, $K' = cK(x_0, x_0)$. If $z = 1$, then $z^T K' z = cK(x_0, x_0) < 0$, which violates the kernel Positive Semi Definiteness (PSD) property. Thus $K'$ is not a kernel.

3. Suppose $K_1(x, z) = \langle \phi^1(x), \phi^1(z) \rangle$ and $K_2(x, z) = \langle \phi^2(x), \phi^2(z) \rangle$. Then, for all $x$ and $z$,
   \[
   K'(x, z) = c_1 \langle \phi^1(x), \phi^1(z) \rangle + c_2 \langle \phi^2(x), \phi^2(z) \rangle = \langle \sqrt{c_1}\phi^1(x), \sqrt{c_1}\phi^1(z) \rangle + \langle \sqrt{c_2}\phi^2(x), \sqrt{c_2}\phi^2(z) \rangle
   = \langle \phi'(x), \phi'(z) \rangle
   \]
   where $\phi'(x)$ is a concatenation of the feature maps $\sqrt{c_1}\phi^1(x)$ and $\sqrt{c_2}\phi^2(x)$. In other words, if the feature maps $\phi^1$ and $\phi^2$ have $m_1$ and $m_2$ coordinates respectively, then $\phi'$ has $m_1 + m_2$ coordinates; for any $x$, the first $m_1$ coordinates of $\phi'(x)$ are $\sqrt{c_1}\phi^1_1(x), \sqrt{c_1}\phi^1_2(x), \ldots, \sqrt{c_1}\phi^1_{m_1}(x)$ and the remaining $m_2$ coordinates of $\phi'(x)$ are $\sqrt{c_2}\phi^2_1(x), \sqrt{c_2}\phi^2_2(x), \ldots, \sqrt{c_2}\phi^2_{m_2}(x)$. Therefore $K'(x, z)$ is a kernel corresponding to the feature map $\phi'$. 
Problem 2: 14 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.

1. $x = [x_1, x_2], z = [z_1, z_2], x_1, x_2, z_1, z_2$ are real numbers. $K(x, z) = x_1 z_2.$

2. Let $x = [x_1, \ldots, x_d], z = [z_1, \ldots, z_d], x_i$s and $z_i$s are real numbers. $K(x, z) = 1 - \langle x, z \rangle.$

3. $x = [x_1, \ldots, x_d], z = [z_1, \ldots, z_d],$ and $f$ is a function. $K(x, z) = f(x_1, x_2)f(z_1, z_2).$

4. $x = [x_1, \ldots, x_d], z = [z_1, \ldots, z_d], x_i$s and $z_i$s are integers between 0 and 100. $K(x, z) = \sum_{i=1}^d \min(x_i, z_i).$

5. $x = [x_1, \ldots, x_d], z = [z_1, \ldots, z_d], x_i$s and $z_i$s are real numbers.

$$K(x, z) = (1 + x_1 z_1)(1 + x_2 z_2) \ldots (1 + x_d z_d)$$

6. $x = [x_1, \ldots, x_d], z = [z_1, \ldots, z_d], x_i$s and $z_i$s are integers between 0 and 100. $K(x, z) = \sum_{i=1}^d \max(x_i, z_i).$

7. $x$ are $z$ are documents with words from some dictionary $D$. $K(x, z)$ is the number of words that occur in both $x$ and $z,$ where each unique common word is counted once.

Solution

1. $K(x, z)$ is not a kernel.

   For $x = [1, -1],$ we have $K(x, x) = 1 \times -1 = -1.$ The corresponding kernel matrix $K = -1.$ For $v = 1,$ $v^\top K v = -1 < 0,$ which violates the PSD property. Thus $K$ is not a kernel.

2. $K(x, z)$ is not a kernel.

   For $x = [2, 2, \ldots],$ we have $K(x, x) = 1 - \langle x, x \rangle = 1 - 4d.$ The corresponding kernel matrix $K = 1 - 4d.$ For $v = 1,$ $v^\top K v = 1 - 4d < 0,$ which violates the kernel PSD property for $d > 0.$ Thus $K$ is not a kernel.
3. \( K(x, z) \) is a kernel corresponding to the feature map \( \phi(x) = f(x_1, x_2) \).

4. \( K(x, z) \) is a kernel.
   Let \( K_i(x, z) = \min(x_i, z_i) \). From Problem 1, we know that the sum of two kernels \( K_1 \) and \( K_2 \) is also a kernel whose corresponding feature map is the concatenation of the feature maps corresponding to \( K_1 \) and \( K_2 \). Thus if we can find the feature maps for all \( K_i(x, z) \), then we can get the feature map for \( K(x, z) \) by concatenating these maps. Consider following feature map:
   \[
   \phi_i(x) = [f_1(x_i), f_2(x_i), \ldots, f_{100}(x_i)]^\top
   \]
   where \( f_k(t) = I(t \geq k) = \begin{cases} 1 & t \geq k \\ 0 & t < k \end{cases} \). Without loss of generality, suppose that \( x_i \leq z_i \). Then \( \phi_i(x) = [1, \ldots, 1, 0, \ldots, 0]^\top \) where only the first \( x_i \) entries are 1. Analogously, \( \phi_i(z) = [1, \ldots, 1, 0, \ldots, 0]^\top \) where only the first \( z_i \) entries are 1. Then
   \[
   \langle \phi_i(x), \phi_i(z) \rangle = \sum_{i=1}^{x_i} 1 \cdot 1 + \sum_{i=x_i+1}^{z_i} 0 \cdot 1 + \sum_{i=z_i+1}^{100} 0 \cdot 0 = x_i = \min(x_i, z_i)
   \]
   Therefore \( K_i(x, z) \) is a kernel corresponding to the feature map \( \phi_i(x) = [f_1(x_i), f_2(x_i), \ldots, f_{100}(x_i)]^\top \), and \( K(x, z) \) is a kernel corresponding to the feature map \( \phi(x) \) which is a concatenation of the feature maps \( \phi_1(x), \phi_2(x), \ldots, \phi_{100}(x) \).

5. \( K(x, z) \) is a kernel.
   Let \( K_i(x, z) = 1 + x_i z_i \), then \( K(x, z) = \prod_{i=0}^{d} K_i(x) \). From Problem 1, we know that the product of two kernels is also a kernel. Since \( K_i(x, z) \) is a kernel corresponding to the feature map \( \phi_i(x) = [1, x_i]^\top \), \( K(x, z) \) is also a kernel. More specifically, \( K(x, z) \) is a kernel corresponding to the feature map \( \phi(x) \), where for any \( x \), \( \phi(x) \) has \( 2d \) coordinates, one corresponding to each subset \( S \) of \( \{1, 2, \ldots, d\} \). \( \phi_S(x) \), the coordinate of \( \phi(x) \) corresponding to the set \( S \) is \( \prod_{i \in S} x_i \). This kernel is called the All Subsets kernel.

6. \( K(x, z) \) is not a kernel.
   One way to prove this is by showing a violation of the PSD property. Let \( x = [0, \ldots, 0] \), \( z = [1, 0, \ldots, 0] \) and \( v = [1, -1]^\top \). Then the kernel matrix
   \[
   K = \begin{bmatrix}
   K(x, x) & K(x, z) \\
   K(z, x) & K(z, z)
   \end{bmatrix} = \begin{bmatrix}
   0 & 1 \\
   1 & 1
   \end{bmatrix}
   \]
   Thus, \( v^\top K v = -1 < 0 \), which violates positivity.
   Another nice way is through a violation of the Cauchy-Schwartz inequality. Consider \( x = [0, \ldots, 0] \) and \( z = [1, 0, \ldots, 0] \). Then \( K(x, x) = 0, K(x, z) = K(z, x) = 1 \), which violates Cauchy-Schwartz inequality – that is \( K(x, z)^2 \geq K(x, x) \cdot K(z, z) \).

7. \( K \) is a kernel. The feature map \( \phi \) has a coordinate for each word \( u \) in the dictionary \( D \). Given a document \( x \), the coordinate of \( \phi(x) \) corresponding to word \( u \), \( \phi_u(x) \), is 1 if \( x \) contains the word \( u \) and 0 otherwise. Notice that this kernel is very similar to the string kernel we discussed in class.

**Problem 3: Programming Assignment: 18 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 `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 $|\Sigma| = 21$ (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.

**Solution**

1. After a single pass of the kernel perceptron algorithm, the training error and test error are listed as follows:

<table>
<thead>
<tr>
<th>$p$</th>
<th>training error</th>
<th>test error</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>1.35%</td>
<td>4.09%</td>
</tr>
<tr>
<td>4</td>
<td>0.716%</td>
<td>2.90%</td>
</tr>
<tr>
<td>5</td>
<td>0.634%</td>
<td>4.62%</td>
</tr>
</tbody>
</table>

2. After a single pass of the kernel perceptron algorithm, the training error and test error are listed as follows:

<table>
<thead>
<tr>
<th>$p$</th>
<th>training error</th>
<th>test error</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>1.29%</td>
<td>5.27%</td>
</tr>
<tr>
<td>4</td>
<td>0.881%</td>
<td>2.90%</td>
</tr>
<tr>
<td>5</td>
<td>0.606%</td>
<td>4.35%</td>
</tr>
</tbody>
</table>

3. There are 5 strings tied for first place. They are ‘DTAGQ’, ‘KVGPD’, ‘LFLNK’, ‘WDTAG’ and ‘GKSSL’.