CS 4780/5780 Homework 7

Due: Tuesday 11/20/18 11:55pm on Gradescope

Problem 1: Kernelized Perceptron

In this problem, we are going to kernelize the perceptron algorithm. Recall the perceptron algorithm

Algorithm 1: Perceptron Algorithm

Initialize $\mathbf{w} = \vec{0}$;
while *TRUE* **do**
  m = 0 ;
  for $(\mathbf{x}_i, y_i) \in D$ **do**
    if $y_i(\mathbf{w}^T \mathbf{x}_i) \leq 0$ **then**
      $\mathbf{w} \leftarrow \mathbf{w} + y_i \mathbf{x}_i$;
      $m \leftarrow m + 1$;
    end
    if $m = 0$ **then**
      break
    end
  end
end

Now, in a classification task, there are α_i misclassifications for each training point $\mathbf{x}_i$. We start with $\mathbf{w} = 0$ and at each iteration we either add or subtract a vector $\mathbf{x}_i$. It follows that the final weight vector must take the form

$$\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i.$$

This observation allows us to modify the perceptron algorithm such that we only need to keep track of the number of misclassifications for each training point, instead of updating $\mathbf{w}$.

(a) Fill in the skeleton code so that the perceptron algorithm only needs to keep track of the number of misclassifications for each training point, instead of updating $\mathbf{w}$.

Algorithm 2: Modified Perceptron Algorithm

```
Initialize ______________________________ ;
while TRUE do
    m = 0 ;
    for (x_i, y_i) ∈ D do
        if ______________________________ then
            ______________________________;
            m ← m + 1;
        end
        if m = 0 then
            break
        end
    end
end
```

(b) How would you modify algorithm 2 to kernelize the perceptron algorithm?

Problem 2: Constructing Kernels

In class, we have shown how to construct new kernels from existing valid kernels, following a few "kernel-preserving" rules. In this problem, we will prove that these rules indeed produce valid kernels.
In order to justify that a function $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a valid kernel function, we can show that either of the following properties hold:

1. The matrix
$$K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$$
is symmetric and positive semidefinite for any set of vectors $\mathbf{x}_1, ..., \mathbf{x}_n \in \mathbb{R}^d$
2. $k(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j)$ for some transformation ϕ

Suppose k_1, k_2 are valid kernel functions. Show that the following kernels are valid:

(a) $k(\mathbf{x}_1, \mathbf{x}_2) = c k_1(\mathbf{x}_1, \mathbf{x}_2)$ for any $c \geq 0$.

(b) $k(\mathbf{x}_1, \mathbf{x}_2) = k_1(\mathbf{x}_1, \mathbf{x}_2) + k_2(\mathbf{x}_1, \mathbf{x}_2)$

(c) $k(\mathbf{x}_1, \mathbf{x}_2) = k_1(\mathbf{x}_1, \mathbf{x}_2) k_2(\mathbf{x}_1, \mathbf{x}_2)$

Problem 3: Gaussian Process Regression

Consider you have the following dataset:

No.	x	y
1	-1	0
2	1	2

and you are going to use a Gaussian Process

$$f(x) \sim GP(m, k)$$

to model the underlying relationship between x and y. After discussing with the professor, you decide to use a function for the mean which is identically zero and a quadratic kernel for your GP, namely,

$$m(x) = 0$$

$$k(x, x') = (x \cdot x' + 1)^2$$

(a) Write down the kernel (i.e. covariance) of the GP prior for the given dataset.

(b) How does the distribution of the GP posterior change if you assume Gaussian noise with variance $\sigma^2 = 0.1$?

(c) You are given the following test points:

No.	x	y
1	0	0.5
2	2	4.5

Assume a noise free setup. What is the mean and covariance of your GP posterior?