CS 4780/5780 Homework 7 Solution

Problem 1: Kernelized Perceptron

(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 $\vec{w}$

Algorithm 1: Modified Perceptron Algorithm

Initialize $\vec{\alpha} = \vec{0}$;
while $TRUE$ **do**
    m = 0 ;
    for $(x_i, y_i) \in D$ **do**
        if $y_i \sum_{j=1}^{n} \alpha_j y_j \vec{x_j}^T \vec{x_i} \leq 0$ **then**
            $\alpha_i \leftarrow \alpha_i + 1$;
            $m \leftarrow m + 1$;
        end
    end
    if $m = 0$ **then**
        break
    end
end

(b) Now, how would you modify algorithm 2 to kernelize the perceptron algorithm?

Change $y_i \sum_{j=1}^{n} \alpha_j y_j \vec{x_j}^T \vec{x_i}$ to $y_i \sum_{j=1}^{n} \alpha_j y_j K(x_j, x_i)$, where K is the kernel function.

Problem 2: Constructing Kernels

Solution: Suppose ϕ_1 and ϕ_2 are the transformations associated with k_1 and k_2 repsectively.

(a) Notice that $k(\vec{x}_i, \vec{x}_j) = ck_1(\vec{x}_i, \vec{x}_j) = c\phi_1(\vec{x}_i)^T \phi_1(\vec{x}_j) = (\sqrt{c}\phi_1(\vec{x}_i))^T(\sqrt{c}\phi_1(\vec{x}_j))$. We can take $\phi_4(\vec{x}_i) = \sqrt{c}\phi_1(\vec{x}_i)$ as a transformation for $ck_1(\vec{x}_i, \vec{x}_j)$

(b) Observe that $k(\vec{x}_i, \vec{x}_j) = k_1(\vec{x}_i, \vec{x}_j) + k_2(\vec{x}_i, \vec{x}_j) = \phi_1(\vec{x}_i)^T \phi_1(\vec{x}_j) + \phi_2(\vec{x}_i)^T \phi_2(\vec{x}_j) = \begin{bmatrix} \phi_1(\vec{x}_i) \\ \phi_2(\vec{x}_i) \end{bmatrix}^T \begin{bmatrix} \phi_1(\vec{x}_i) \\ \phi_2(\vec{x}_i) \end{bmatrix}$.

We can take $\phi_5(\vec{x}_i) = \begin{bmatrix} \phi_1(\vec{x}_i) \\ \phi_2(\vec{x}_i) \end{bmatrix}$ as a transformation for $k_1(\vec{x}_1, \vec{x}_2) + k_2(\vec{x}_1, \vec{x}_2)$

(c) Notice that

$$\begin{aligned}
k(\vec{x}_i, \vec{x}_j) &= k_1(\vec{x}_i, \vec{x}_j)k_2(\vec{x}_i, \vec{x}_j) \\
&= \phi_1(\vec{x}_i)^T\phi_1(\vec{x}_j)\phi_2(\vec{x}_i)^T\phi_2(\vec{x}_j) \\
&= \sum_{a=1}^{n_1}[\phi_1(\vec{x}_i)]_a[\phi_1(\vec{x}_j)]_a \sum_{b=1}^{n_2}[\phi_2(\vec{x}_i)]_b[\phi_2(\vec{x}_j)]_b \\
&= \sum_{a=1}^{n_1}\sum_{b=1}^{n_2}[\phi_1(\vec{x}_i)]_a[\phi_2(\vec{x}_i)]_b[\phi_1(\vec{x}_j)]_a[\phi_2(\vec{x}_j)]_b
\end{aligned}$$

Suppose $\phi_6(\vec{x}_i) = [[\phi_1(\vec{x}_i)]_1[\phi_2(\vec{x}_i)]_1, ..., [\phi_1(\vec{x}_i)]_1[\phi_2(\vec{x}_i)]_{n_2}, [\phi_1(\vec{x}_i)]_2[\phi_1(\vec{x}_1)]_1, ..., [\phi_1(\vec{x}_1)]_{n_1}[\phi_1(\vec{x}_1)]_{n_2}]^T$. Then,

$$\phi_6(\vec{x}_i)^T\phi_6(\vec{x}_j) = \sum_{a=1}^{n_1}\sum_{b=1}^{n_2}[\phi_1(\vec{x}_i)]_a[\phi_2(\vec{x}_j)]_b[\phi_1(\vec{x}_i)]_a[\phi_2(\vec{x}_j)]_b = k_1(\vec{x}_i, \vec{x}_j)k_2(\vec{x}_i, \vec{x}_j)$$

Problem 3: Gaussian Process Regression

(a) The mean and covariance of the GP prior are defined by the mean and covariance functions given, and are independent of Y. Thus, the prior will have mean $\begin{bmatrix}0 & 0\end{bmatrix}^T$ and covariance

$$\begin{bmatrix}4 & 0\\ 0 & 4\end{bmatrix}$$

(b) Based on the definition, the GP posterior is defined as a normal distribution with mean and covariance as follows:

$$K(X_*, X)(K(X,X) + \sigma^2 I)^{-1}Y = K(X_*, X)(K(X,X) + 0.1I)^{-1}Y$$

$$K(X_*, X_*) - K(X_*, X)(K(X,X)+\sigma^2 I)^{-1}K(X, X_*) = K(X_*, X_*) - K(X_*, X)(K(X,X)+0.1I)^{-1}K(X, X_*)$$

(c) With a noise free setup, the GP posterior is defined as a normal distribution with mean and covariance as follows.

$$K(X_*, X)K(X,X)^{-1}Y$$

$$K(X_*, X_*) - K(X_*, X)K(X,X)^{-1}K(X, X_*)$$

where the function $K(X, X')$ defines a matrix K such that $K_{ij} = k(X_i, X'_j)$. Accordingly, the relevant matrices are:

$$Y = \begin{bmatrix}0 & 2\end{bmatrix}^T$$

$$K(X,X) = \begin{bmatrix}4 & 0\\ 0 & 4\end{bmatrix}$$

$$K(X_*, X) = \begin{bmatrix}1 & 1\\ 1 & 9\end{bmatrix} = K(X, X_*)^T$$

$$K(X_*, X_*) = \begin{bmatrix}1 & 1\\ 1 & 25\end{bmatrix}$$

Accordingly, we can compute the posterior mean and covariance, which are (respectively)

$$\begin{bmatrix}0.5 & 4.5\end{bmatrix}^T$$

$$\begin{bmatrix}0.5 & -1.5\\ -1.5 & 4.5\end{bmatrix}$$

(So this does not match to the target values at each of the test points with variance $\begin{bmatrix}0.5 & -1.5\\ -1.5 & 4.5\end{bmatrix}$).