CSCI567 Theory Assignment 3

Yifan Wang

3038184983

wang608@usc.edu

Problem 1 Lagrangian and Duality (35 points)

We are given N samples: $\{(\mathbf{x}_1, y_1)\dots(\mathbf{x}_N, y_N)\}, \mathbf{x}_i \in \mathcal{X}, y_i \in \{-1, 1\}\ \forall i \in \{1\dots N\}$. We say input $\mathbf{x}_i$ belongs to class $\mathcal{C}_1$ if its label y_i is 1 and it belongs to class $\mathcal{C}_{-1}$ if its label is -1. Mathematically, $\mathcal{C}_1 = \{(\mathbf{x}_i, y_i) : y_i = 1\}$ and $\mathcal{C}_{-1} = \{(\mathbf{x}_i, y_i) : y_i = -1\}$. Now, consider a two class classification problem formulation as follows: We want to find a separating hyper-plane $\mathbf{w}$ such that if input $\mathbf{x}_i$ belongs to $\mathcal{C}_1$ then $\mathbf{w}^T\mathbf{x}_i \geq 0$ and if it belongs to $\mathcal{C}_{-1}$ then $\mathbf{w}^T\mathbf{x}_i \leq 0$. Therefore, we can find the optimal weights $\mathbf{w}^*$ by maximizing the objective

$$f(\mathbf{w}) = \sum_{i=1}^{N} y_i \mathbf{w}^T \mathbf{x}_i$$

Note that $f(\mathbf{w})$ can be arbitrarily maximized by increasing the magnitude of $\mathbf{w}$ once we have found a vector $\mathbf{w}$ such that $f(\mathbf{w}) > 0$. Therefore, we add an additional constraint that $\mathbf{w}^T\mathbf{w} = \|\mathbf{w}\|^2 \leq 1$.

1.1 Write the down the constraint minimization problem. Solve it and find the explicit form the optimal weights $\mathbf{w}^*$. **(10 points)**

$$max \sum y_i W^T x_i = min - \sum y_i W^T x_i$$

subject to

$$||w||^2 - 1 \leq 0$$

Lagrange

$$L = -\sum y_i W^T x_i + \lambda(W^T W - 1)$$

solve

$$\frac{d\,L}{d\,W} = -\sum y_i x_i + 2\lambda W = 0$$

$$\lambda(W^T W - 1) = 0$$

$$\lambda^* = \frac{1}{2}||\sum y_i x_i||$$

$$W^* = \frac{\sum y_i x_i}{||\sum y_i x_i||}$$

1.2 Suppose we use a transformation function $\phi : \mathcal{X} \to^K$ to transform inputs and the corresponding kernel function is $k(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x})^T\phi(\mathbf{x}')$. Analogus to problem 1.1, write down the constraint minimization problem for this setup. **(4 points)**

$$max \sum y_i W^T \phi(x_i) = min - \sum y_i W^T \phi(x_i)$$

subject to

$$||w||^2 - 1 \leq 0$$

1.3 Write down the dual of the optimization problem in 1.2. **(15 points)**

Lagrange

$$L = -\sum y_i W^T \phi(x_i) + \lambda(W^T W - 1)$$

$$\frac{d\,L}{d\,W} = -\sum y_i \phi(x_i) + 2\lambda W = 0$$

$$W = \frac{\sum y_i \phi(x_i)}{2\lambda}$$

$$L = -\sum y_j [\frac{\sum y_i \phi(x_i)}{2\lambda}]^T \phi(x_j) + \lambda([\frac{\sum y_i \phi(x_i)}{2\lambda}]^T [\frac{\sum y_i \phi(x_i)}{2\lambda}] - 1)$$

$$L = -\frac{1}{2\lambda}\sum_i \sum_j y_i y_j \phi(x_i)^T \phi(x_j) + \lambda([\frac{1}{4\lambda^2}\sum_i \sum_j y_i y_j \phi(x_i)^T \phi(x_j) - 1)$$

$$L = -\frac{1}{4\lambda}\sum_i \sum_j y_i y_j \phi(x_i)^T \phi(x_j) - \lambda$$

$$max - \frac{1}{4\lambda}\sum_i \sum_j y_i y_j \phi(x_i)^T \phi(x_j) - \lambda$$

subject to

$$W = \frac{\sum y_i \phi(x_i)}{2\lambda}$$

$$||W||^2 - 1 \leq 0$$

$$\lambda \geq 0$$

1.4 Can the optimization problem in 1.2 be kernelized? Also, can you kernelize the prediction rule? Explain why or why not. **(6 points)**

By above equation, it can be kernalized by replacing $\phi(x_i)^T\phi(x_j)$ with $K(x_i, x_j)$

Problem 2 Support Vector Machines (40 points)

Consider the dataset consisting of points (x, y), where x is a real value, and $y \in \{-1, 1\}$ is the class label. Let's start with three points $(x_1, y_1) = (-1, -1)$, $(x_2, y_2) = (1, -1)$, $(x_3, y_3) = (0, 1)$.

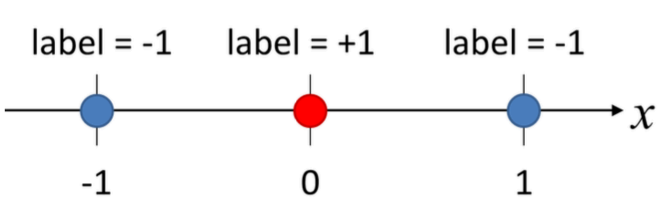


Figure 1: Three data points considered in this problem

2.1 Can three points shown in Figure 1, in their current one-dimensional feature space, be perfectly separated with a linear separator? Explain why or why not. **(2 points)**

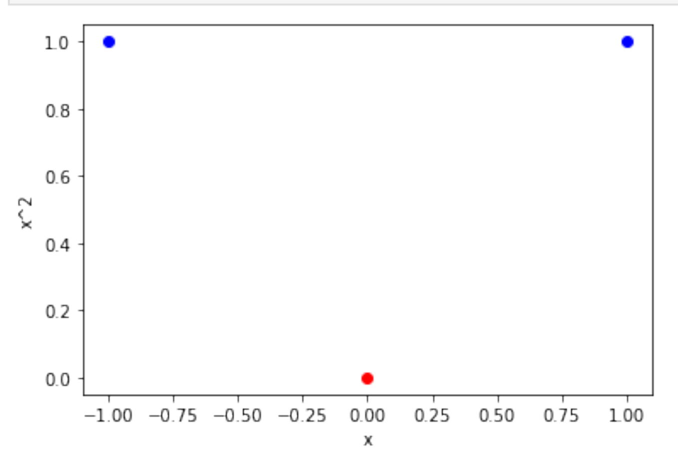
No

Since in 1D space, to be linear separable, the linear separator is a point but obviously, there is no such point existing.

2.2 Now we define a simple feature mapping $\phi(x) = [x, x^2]^T$ to transform the three points from one- to two-dimensional feature space. Plot the transformed points in the new two-dimensional feature space. Is there a linear decision boundary that can separate the points in this new feature space? Explain why or why not. **(2 points)**

In [1]:

```
import numpy as np
import matplotlib.pyplot as plt
x = np.array([-1, 0, 1]).reshape(-1,1)
y = np.array([-1, 1, -1])
phi_x = np.concatenate([x, np.square(x)], axis=1)
for i in range(x.shape[0]):
    if y[i] == 1:
        plt.scatter(phi_x[i, 0], phi_x[i, 1], c='r')
    else:
        plt.scatter(phi_x[i, 0], phi_x[i, 1], c='b')
plt.xlabel('x')
plt.ylabel('x^2')
plt.show()
```



It is linear separable in the new space. From the figure, we can always find a line that prefectly separate the data points.

2.3 Given the feature mapping $\phi(x) = [x, x^2]^T$, write down the kernel function $k(x, x')$. Moreover, write down the 3×3 kernel (or Gram) matrix $\mathbf{K}$ based on $k(x_i, x_j)$ of the three data points. Verify that $\mathbf{K}$ is a positive semi-definite (PSD) matrix. **(6 points)**

$$k(x, x') = xx' + x^2x'^2$$

$$\begin{bmatrix} k(x1, x1) & k(x1, x2) & k(x1, x3) \\ k(x2, x1) & k(x2, x2) & k(x2, x3) \\ k(x3, x1) & k(x3, x2) & k(x3, x3) \end{bmatrix}$$

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

the e-vaulue is 0 and 1, so it is PSD

2.4 Write down the dual formulation of this problem by plugging in the specific data points. **(15 points)**

$$max_\alpha \sum \alpha_i - \frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j y_i y_j \phi(x_i)\phi(x_j)$$

subject to

$$\alpha_i \geq 0$$

$$\sum_i \alpha_i y_i = 0$$

2.5 Solve the above dual form analytically and obtain primal solutions $\mathbf{w}^*$ and b^*. **(15 points)**

$$w^* = [0, -2]^T$$

$$b^* = 1$$

Problem 3 PCA (25 points)

Consider the following design matrix, representing four sample points $X_i \in \mathbb{R}$

$$\mathbf{X} = \begin{bmatrix} 4 & 1 \\ 2 & 3 \\ 5 & 4 \\ 1 & 0 \end{bmatrix}$$

We want to represent the data in only one dimension, so we turn to principal components analysis (PCA).

3.1 Compute the unit-length principal component directions of X, and state which one of the component directions would the PCA algorithm choose if you request just one principal component to be returned. Please provide an exact answer, without approximation. (You will need to use the square root symbol.) Show your work. **(15 points)**

$$X_{cov} = \frac{1}{4-1}(X - mean(X))^T(X - mean(X)) =$$

$$\begin{bmatrix} \frac{10}{3} & 2 \\ 2 & \frac{10}{3} \end{bmatrix}$$

Perform SVD get the

$$eigen - value = \{\frac{16}{3}, \frac{4}{3}\}$$

corresponding eigen-vector

$$[\frac{1}{\sqrt(2)}, -\frac{1}{\sqrt(2)}]^T,\ [\frac{1}{\sqrt(2)}, \frac{1}{\sqrt(2)}]^T$$

Transformed

$$[0, 0, 2\sqrt{2}, -2\sqrt{2}]^T$$

3.2 The plot below (Fig. 3.2) depicts the sample points from X. We want a one-dimensional representation of the data, so draw the principal component direction (as a line) and the projections of all four sample points onto the principal direction.

Label each projected point with its principal coordinate value (where the origin's principal coordinate is zero). Give the principal coordinate values exactly. **(10 points)**

The red line is the new 1D axis, the 4 red points (two at 0) are the projected coordinates