# Machine Learning, 2024 Fall
# Homework 4

## Notice

Due 23:59 (CST), Dec 26, 2024

Plagiarizer will get 0 points.

LaTeXis highly recommended. Otherwise you should write as legibly as possible.

## 1 Support Vector Machine [30pts]

1. Recall the hard-margin SVM objective:

$$\text{minimize } \frac{1}{2}\|\mathbf{w}\|^2$$
$$\text{subject to } y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 \quad \forall i$$

The constraints specify that the (functional) margin of each example is at least 1. If we change the constraint to require the margin to be at least $c$ ($c > 0$), i.e., solving:

$$\text{minimize } \frac{1}{2}\|\mathbf{w}\|^2$$
$$\text{subject to } y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq c \quad \forall i$$

(1) Would it change the separating hyperplane? Why or why not?

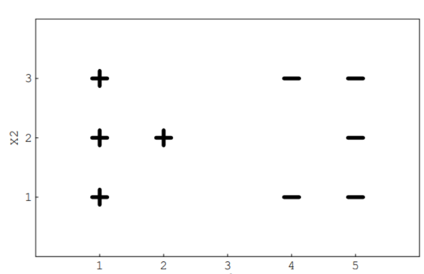
**Solution:** No. It would only scale w and b.

(2) Let $\mathbf{w}^*$ be the solution of the original hard-margin SVM, and $\mathbf{w}_0$ be the solution of the modified problem with margin at least $c$. Write an expression for $\mathbf{w}_0$ in terms of $\mathbf{w}^*$:

**Solution:**

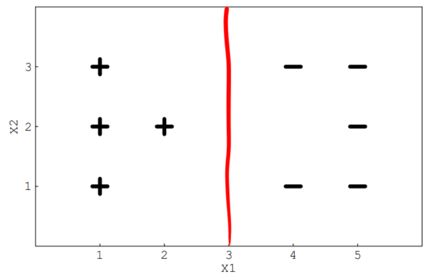
$$\mathbf{w}_0 = c \cdot \mathbf{w}^*$$

2. Suppose we are using a linear SVM (i.e., no kernel), with some large $C$ value, and are given the following data set.



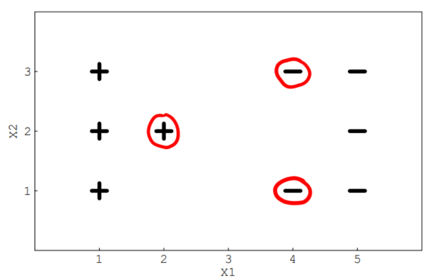
(1) Draw the decision boundary of linear SVM. Give a brief explanation.

**Solution:**




(2) In the above image, circle the points such that removing that example from the training set and retraining SVM, we would get a different decision boundary than training on the full sample. You need to offer a brief explanation.

**Solution:**



## 2 Kernel Function [20pts]

(a) Let $k_1(u, v)$ be a valid kernel. Consider the new kernel function $k(u, v) = \exp(k_1(u, v))$, where $\exp(x)$ is the standard exponential function. Prove that $k(u, v)$ is also a valid kernel, i.e., show that it is positive semi-definite.

**Solution:** Let $k_1(u, v)$ be a valid kernel. We show that $k(u, v) = \exp(k_1(u, v))$ is also a valid kernel.

(1) Symmetry: Since $k_1(u, v)$ is symmetric, we have:

$$k(u, v) = \exp(k_1(u, v)) = \exp(k_1(v, u)) = k(v, u).$$

Thus, $k(u, v)$ is symmetric.

(2) Positive Semi-Definiteness: For any $\mathbf{z} \in \mathbb{R}^n$, we need to show:

$$\mathbf{z}^T K \mathbf{z} = \sum_{i,j} z_i z_j \exp(k_1(u_i, u_j)) \geq 0.$$

Since $k_1(u, v)$ is positive semi-definite, $\mathbf{z}^T K_1 \mathbf{z} \geq 0$. The exponential function preserves positive semi-definiteness, so $K$ is also positive semi-definite:

$$\mathbf{z}^T K \mathbf{z} \geq 0.$$

Since $k(u, v)$ is symmetric and positive semi-definite, it is a valid kernel.

(b) Let $K_1(x, z)$ and $K_2(x, z)$ be valid kernels. Proving that for non-negative constants $c_1$ and $c_2$, $K_0(x, z) = c_1 K_1(x, z) + c_2 K_2(x, z)$ is a valid kernel function.

**Solution:** Let $K_1(x, z) = \langle \varphi_1(x), \varphi_1(z) \rangle$ and $K_2(x, z) = \langle \varphi_2(x), \varphi_2(z) \rangle$.

Then,

$$K_0(x, z) = c_1 K_1(x, z) + c_2 K_2(x, z) = c_1 \langle \varphi_1(x), \varphi_1(z) \rangle + c_2 \langle \varphi_2(x), \varphi_2(z) \rangle.$$

Rewriting:

$$K_0(x, z) = \langle \sqrt{c_1} \varphi_1(x), \sqrt{c_1} \varphi_1(z) \rangle + \langle \sqrt{c_2} \varphi_2(x), \sqrt{c_2} \varphi_2(z) \rangle.$$

Define the new feature map $\varphi_0(x)$ as:

$$\varphi_0(x) = \left( \sqrt{c_1} \varphi_1(x), \sqrt{c_2} \varphi_2(x) \right).$$

Thus,

$$K_0(x, z) = \langle \varphi_0(x), \varphi_0(z) \rangle.$$

Since $K_0(x, z)$ is the inner product of $\varphi_0(x)$ and $\varphi_0(z)$, $K_0(x, z)$ is a valid kernel.

## 3 Support Vector Machine with Kernel [20pts]

Suppose we use a Support Vector Machine (SVM) with a custom kernel defined as:

$$K(x, x') = \begin{cases} 1 & \text{if } x = x', \\ -1 & \text{if } x \neq x'. \end{cases}$$

This corresponds to mapping each $x$ to a vector $\psi(x)$ in some high-dimensional space (that need not be specified) so that

$$K(x, x') = \psi(x)^T \psi(x').$$

As in the original setup, we are given $m$ training samples $(x_1, y_1), \ldots, (x_m, y_m)$, where $y_i \in \{-1, +1\}$, and all the data points $x_i$ are distinct (i.e., $x_i \neq x_j$ for $i \neq j$).

Based on the standard SVM optimization problem, derive the expression of $\alpha_i$ when using the kernel defined above. Recall that the weight vector w used in SVMs has the form

$$\mathbf{w} = \sum_i \alpha_i y_i \psi(x_i)$$

**Solution:**

$$\mathcal{L}(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j K(x_i, x_j)$$

$$K(x_i, x_j) = \begin{cases} 1 & \text{if } x_i = x_j, \\ -1 & \text{if } x_i \neq x_j. \end{cases}$$

$$\mathcal{L}(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \alpha_i^2 + \frac{1}{2} \sum_{i \neq j} \alpha_i \alpha_j y_i y_j$$

$$\frac{\partial \mathcal{L}(\alpha)}{\partial \alpha_i} = 1 - \alpha_i + \sum_{j=1}^{m} \alpha_j y_i y_j$$

$$1 - \alpha_i + \sum_{j=1}^{m} \alpha_j y_i y_j = 0$$

$$\alpha_i = 1 + \sum_{j=1}^{m} \alpha_j y_i y_j$$

## 4 K-Means [30pts]

Recalling the K-means, we iteratively find the cluster centers $\mu_t^k$ and upadte the class $C_t^k$ for all data. Given a clusters number $K$, our goal is to minimize SSE

$$SSE = \sum_{k=1}^{K} \sum_{i \in C_t^k} \left\| x_i - \mu_t^k \right\|_2^2$$

(1) **Please prove that the K-means algorithm converges.**

(2) **Implement the K-means algorithm on the dataset we provide in Zip.** Your answer should include: embedded code, comment on your code and visual screenshot of your clustering results of K=2,5,10. Hint: Implement the K-means algorithm by hand (Dont use the sklearn implementation)

**Solution:** For K-means, there are two steps (I) finding the closest cluster (II) setting each cluster to the mean of all assigned data. We need to prove both steps decrease SSE

(I) SSE calculate the sum of distances between each data point and the assigned cluster. For arbitrary data point, let the origin assigned cluster be $\mu_t$ and the updated cluster $\mu_{t+1}$. The updated distance $\|x_i - \mu_{t+1}\| \leq \|x_i - \mu_t\|$, therefore the overall $SSE_{t+1} \leq SSE_t$.

(II) Let us consider one cluster $k$ and the corresponding part of SSE $\sum_{i \in C_t} \|x_i - \mu_t\|_2^2$. Let $\bar{x}_i$ denotes the new cluster center for this cluster. The proof is as follows:

$$\begin{aligned}
\sum_{i \in C_t} \|x_i - \mu_t\|_2^2 &= \sum_{i \in C_t} \|x_i - \bar{x}_i + \bar{x}_i - \mu_t\|_2^2 \\
&= \sum_{i \in C_t} \|x_i - \bar{x}_i\|_2^2 + \sum_{i \in C_t} \|\bar{x}_i - \mu_t\|_2^2 + 2 \sum_{i \in C_t} \left\| x_i \bar{x}_i - x_i \mu_t - \bar{x}_i^2 + \bar{x}_i \mu_t \right\| \\
&= \sum_{i \in C_t} \|x_i - \bar{x}_i\|_2^2 + \sum_{i \in C_t} \|\bar{x}_i - \mu_t\|_2^2 + 2\,|C_t| * \left\| \bar{x}_i \bar{x}_i - \bar{x}_i \mu_t - \bar{x}_i^2 + \bar{x}_i \mu_t \right\| \\
&= \sum_{i \in C_t} \|x_i - \bar{x}_i\|_2^2 + \sum_{i \in C_t} \|\bar{x}_i - \mu_t\|_2^2 = \sum_{i \in C_t} \|x_i - \bar{x}_i\|_2^2 + 2 \mid C_t\| * \|\bar{x}_i - \mu_t\|_2^2 \\
&\geq \sum_{i \in C_t} \|x_i - \bar{x}_i\|_2^2
\end{aligned}$$

All clusters follow the above proof, then we prove that step (II) decrease SSE.