Assignment Six
ECE 4200

- Provide credit to **any sources** other than the course staff that helped you solve the problems. This includes **all students** you talked to regarding the problems.
- You can look up definitions/basics online (e.g., wikipedia, stack-exchange, etc).
- **The due date is 10/25/2020, 23.59.59 ET**.
- Submission rules are the same as previous assignments.

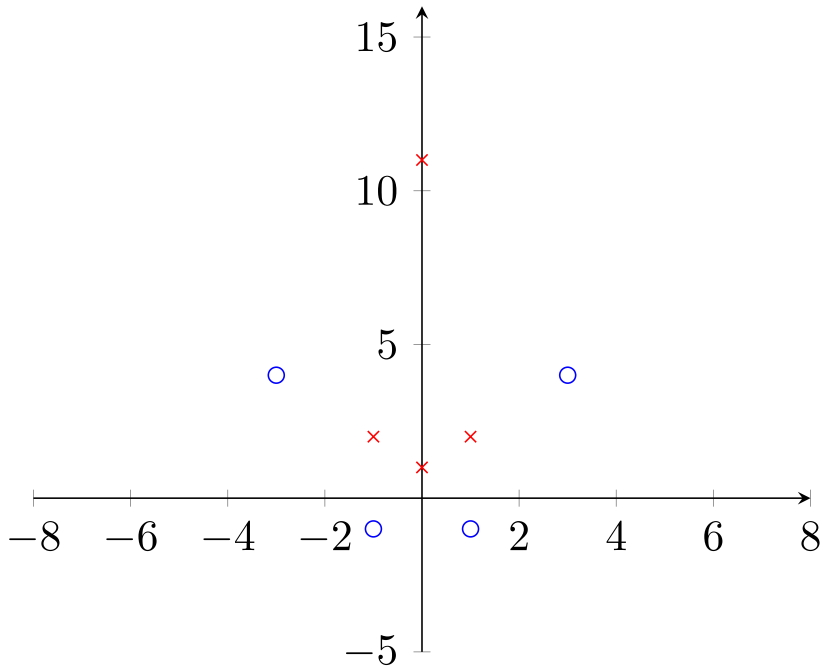
Problem 1. (15 points). SVM's obtain *non-linear* decision boundaries by mapping the feature vectors $\bar{X} \in \mathbb{R}^d$ to a possibly high dimensional space via a function $\phi : \mathbb{R}^d \to \mathbb{R}^m$, and then finding a linear decision boundary in the new space.

We also saw that to implement SVM, it suffices to know the kernel function $K(\bar{X}_i, \bar{X}_j) = \phi(\bar{X}_i) \cdot \phi(\bar{X}_j)$, without even explicitly specifying the function ϕ.

Recall **Mercer's theorem**. K is a kernel function if and only if for any n vectors, $\bar{X}_1, \ldots, \bar{X}_n \in \mathbb{R}^d$, and **any** real numbers $c_1, \ldots, c_n$, $\sum_{i=1}^n \sum_{j=1}^n c_i c_j K(\bar{X}_i, \bar{X}_j) \geq 0$.

1. Prove the following half of Mercer's theorem (which we showed in class). If K is a kernel then $\sum_{i=1}^n \sum_{j=1}^n c_i c_j K(\bar{X}_i, \bar{X}_j) \geq 0$.
2. Let $d = 1$, and $x, y \in \mathbb{R}$. Is the function $K(x, y) = x + y$ a kernel?
3. Let $d = 1$, and $x, y \in \mathbb{R}$. Is $K(x, y) = xy + 1$ a kernel?
4. Suppose $d = 2$, namely the original features are of the form $\bar{X}_i = [\bar{X}^1, \bar{X}^2]$. Show that $K(\bar{X}, \bar{Y}) = (1 + \bar{X} \cdot \bar{Y})^2$ is a kernel function. This is called as **quadratic kernel**.

   (**Hint**: Find a $\phi : \mathbb{R}^2 \to \mathbb{R}^m$ (for some m) such that $\phi(\bar{X}) \cdot \phi(\bar{Y}) = (1 + \bar{X} \cdot \bar{Y})^2$).
5. Consider the training examples $\langle[0, 1], 1\rangle, \langle[1, 2], 1\rangle, \langle[-1, 2], 1\rangle, \langle[0, 11], 1\rangle, \langle[3, 4], -1\rangle, \langle[-3, 4], -1\rangle, \langle[1 - 1], -1\rangle, \langle[-1, -1], -1\rangle$. We have plotted the data points below.
   - Is the data **linearly classifiable** in the original 2-d space? If yes, please come up with *any* linear decision boundary that separates the data. If no, please explain why.
   - Is the data linearly classifiable in the feature space corresponding to the quadratic kernel. If yes, please come up with *any* linear decision boundary that separates the data. If no, please explain why.



Problem 2. (10 points). Let $f, h_i, 1 \le i \le n$ be real-valued functions and let $\alpha \in \mathbb{R}^n$. Let $L(z, \boldsymbol{\alpha}) = f(z) + \sum_{i=1}^{n} \alpha_i h_i(z)$. In this problem, we will prove that the following two optimization problems are equivalent.

$$\begin{aligned} &\min_z f(z) \\ &\text{s.t. } h_i(z) \le 0,\ i = 1, \ldots, n. \end{aligned} \tag{1}$$

$$\min_z \left[\max_{\boldsymbol{\alpha} \ge \mathbf{0}} L(z, \boldsymbol{\alpha}) \right] \tag{2}$$

Let (z^*, α^*) be the solution of (2) and let z_p^* be the solution of (1). Prove that:

$$L(z^*, \alpha^*) = f(z_p^*)$$

Hint: Use the fact that for any z, $\boldsymbol{\alpha} \ge \mathbf{0}$, $L(z^*, \alpha^*) \ge L(z^*, \alpha)$ and $L(z^*, \alpha^*) \le L(z, \alpha_z)$, where $\alpha_z = \arg\max_{\alpha \ge \mathbf{0}} L(z, \alpha)$.

You may follow the following steps but it is not required as long as your proof is correct.

1. Prove that $L(z^*, \alpha^*) \le f(z_p^*)$
2. Prove that $L(z^*, \alpha^*) \ge f(z_p^*)$

Problem 3. (15 points). In this problem, we derive the dual formulation of the soft-margin SVM problem with $\bar{\xi} = \xi_1, \ldots, \xi_n$.

$$\begin{aligned} &\min_{\bar{w}, \bar{\xi}} \frac{1}{2} \|\bar{w}\|_2^2 + C \cdot \sum_{i=1}^{n} \xi_i \\ &\text{such that} \\ &\qquad 1 - \xi_i - y_i(\bar{X}_i \cdot \bar{w} - t) \le 0,\ i = 1, \ldots, n, \\ &\qquad -\xi_i \le 0,\ i = 1, \ldots, n. \end{aligned} \tag{3}$$

Now we can define $2n$ Lagrangian variables $\boldsymbol{\alpha} = \alpha_1, \ldots, \alpha_n$, and $\boldsymbol{\beta} = \beta_1, \ldots, \beta_n$ corresponding to these equations and obtain the following Lagrangian.

$$L(\bar{w}, t, \bar{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{2}\|\bar{w}\|_2^2 + C \cdot \sum_{i=1}^{n} \xi_i + \sum_{i=1}^{n} \alpha_i (1 - \xi_i - y_i(\bar{X}_i \cdot \bar{w} - t)) - \sum_{i=1}^{n} \beta_i \xi_i \tag{4}$$

The original problem is now equivalent to

$$\min_{\bar{w}, t, \bar{\xi}} \left[\max_{\boldsymbol{\alpha} \geq \mathbf{0}, \boldsymbol{\beta} \geq \mathbf{0}} L(\bar{w}, t, \bar{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \right].$$

For this objective, minmax theorem says that the min max problem is equivalent to the max min problem below. You do not have to prove this, but are encouraged to do so (using the argument given in the discussion earlier). This gives the following problem.

$$\max_{\boldsymbol{\alpha} \geq \mathbf{0}, \boldsymbol{\beta} \geq \mathbf{0}} \left[\min_{\bar{w}, t, \bar{\xi}} L(\bar{w}, t, \bar{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \right].$$

1. For a fixed $\boldsymbol{\alpha}, \boldsymbol{\beta}$ take the gradient of the the Lagrangian with respect to $\bar{w}$, and express $\bar{w}$ in terms of the other variables.

2. Differentiate the Lagrangian with respect to t, and equate to zero to obtain another equation.

3. Differentiate the Lagrangian with respect to ξ_i and show that $\alpha_i + \beta_i = C$ at the optimum. This shows that $\alpha_i \leq C$, since $\beta_i \geq 0$.

4. The expressions from 1, 2, 3 define one of the KKT conditions. Show that under these conditions,

   $$\min_{\bar{w}, t, \bar{\xi}} L(\bar{w}, t, \bar{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j (\bar{X}_i \cdot \bar{X}_j).$$

   Basically, use the expressions from 1, 2, and 3 to "cancel out" $\bar{w}, t, \bar{\xi}$ in the Lagrangian.

5. Combine the results above to argue that the following optimization is equivalent to the soft margin SVM we started with.

   $$\begin{aligned} &\max_{\boldsymbol{\alpha} \geq \mathbf{0}} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j (\bar{X}_i \cdot \bar{X}_j) \\ &\text{such that} \\ &\qquad 0 \leq \alpha_i \leq C,\ i = 1, \ldots, n. \end{aligned} \tag{5}$$

Problem 4 (25 points) SVM Classification. Please refer to the Jupyter Notebook in the assignment, and complete the coding part in it! You can use sklearn SVM package: `https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html#sklearn.svm.SVC`