AI534 — Written Homework Assignment 2 (45 pts) —

This assignment covers Kernel methods and Support vector machines.

1. (Cubic Kernels.) (8 pts) In class, we showed that the quadratic kernel $K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i \cdot \mathbf{x}_j + 1)^2$ was equivalent to mapping each $\mathbf{x} = (x_1, x_2) \in R^2$ into a higher dimensional space where

$$\Phi(\mathbf{x}) = (x_1^2, x_2^2, \sqrt{2}x_1x_2, \sqrt{2}x_1, \sqrt{2}x_2, 1).$$

   Now consider the cubic kernel $K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i \cdot \mathbf{x}_j + 1)^3$. What is the corresponding Φ function?

2. (Kernel or not). (10 pts) In the following problems, suppose that K, K_1 and K_2 are kernels with feature maps ϕ, ϕ_1 and ϕ_2. For the following functions $K'(x, z)$, state if they are kernels or not. If they are kernels, write down the corresponding ϕ in terms of ϕ, ϕ_1 and ϕ_2. If they are not kernels, prove that they are not.

   - (2 pts) $K'(\mathbf{x}, \mathbf{z}) = cK(\mathbf{x}, \mathbf{z})$ for $c > 0$.
   - (2 pts) $K'(\mathbf{x}, \mathbf{z}) = cK(\mathbf{x}, \mathbf{z})$ for $c < 0$.
   - (2 pts) $K'(\mathbf{x}, \mathbf{z}) = c_1K_1(\mathbf{x}, \mathbf{z}) + c_2K_2(\mathbf{x}, \mathbf{z})$ for $c_1, c_2 > 0$.
   - (4 pts) $K'(\mathbf{x}, \mathbf{z}) = K_1(\mathbf{x}, \mathbf{z})K_2(\mathbf{x}, \mathbf{z})$.

3. Kernelizing Logistic Regression (7pts) For tihs problem you will follow the example of kernelizing perceptron, to kernelize the logistic regression shown below.

Algorithm 1: Stochastic gradient descent for logistic regression

Input: $\{(\mathbf{x}_i, y_i)_{i=1}^N\}$(training data), γ(learning rate)
Output: learned weight vector $\mathbf{w}$

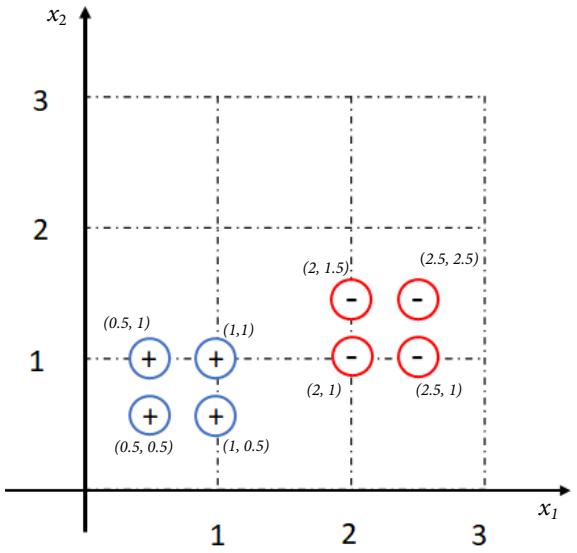
```
Initialize w = 0;
while not converged do
    for i = 1, ..., N do
        w ← w + γ(y_i − σ(w^T x_i))x_i
    end
end
```

   Specifically, please:

   (a) (2pts) Argue that the solution $\mathbf{w}^*$ for logistic regression can be expressed as the weighted sum of training examples (similar to slide 8 of the kernel methods lecture)

   (b) (5 pts) Modify the following stochastic gradient descent algorithm logistic regression algorithm to kernelize it. (Hint: similar to the bottom algorithm on slide 14, but instead of counter, you will learn a continuous weights for α's)

4. (Hard margin SVM) (6 pts) Apply linear SVM without soft margin to the following problem.



a. (3pts) Please mark out the support vectors, the decision boundary ($w_1x_1 + w_2x_2 + b = 0$) and $w_1x_1 + w_2x_2 + b = 1$ and $w_1x_1 + w_2x_2 + b = -1$. You don't need to solve the optimization problem for this, you should be able to eyeball the solution and find the linear separator with the largest margin.

b. (3 pts) Please solve for w_1, w_2 and b based on the support vectors you identified in (a). Hint: the support vectors would have functional margin = 1.

5. L_2 SVM (14 pts)
Given a set of training examples $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$, where $y_i \in \{1, -1\}$ for all i. The following is the primal formulation of L_2 SVM, a variant of the standard SVM obtained by squaring the slacks.

$$\begin{aligned} \min_{\mathbf{w},b,\xi} \ & \frac{1}{2}\mathbf{w}^T\mathbf{w} + c\sum_{i=1}^{N}\xi_i^2 \\ \text{s.t. } & y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i, \ i \in \{1, \cdots, N\} \\ & \xi_i \geq 0, \ i \in \{1, \cdots, N\} \end{aligned}$$

a. (3pts) Show that removing the second constraint $\xi_i \geq 0$ will not change the solution to the problem. In other words, let $(\mathbf{w}^*, b^*, \xi^*)$ be the optimal solution to the problem without this set of constraints, show that $\xi_i^* \geq 0$ must be true, $\forall i \in \{1, \cdots, N\}$. (Hint: use proof by contradiction by assuming that there exists some $\xi_i^* < 0$.)

b. (3 pts) After removing the second set of constraints, we have a simpler problem with only one set of constraints. Now provide the lagrangian of this new problem.

c. (8pts) Derive the dual of this problem following the same procedure as illustrated in the lecture. How does the dual problem differ from that of the standard SVM with hinge loss? In particular, which of the two formulations is more sensitive to outliers? Why?