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?

We use $x = (x_1, x_2)$ to calculate the cubic kernel $K(x_i, x_j) = (x_i \cdot x_j + 1)^3$. For example, we can apply $x_i = (x_{i1}, x_{i2})$ and $x_j = (x_{j1}, x_{j2})$ to the cubic kernel.

$$(x_i \cdot x_j + 1)^3 = (x_{i1}x_{j1} + x_{i2}x_{j2} + 1)^3$$
$$= x_{i1}^3x_{j1}^3 + x_{i2}^3x_{j2}^3 + 3x_{i1}x_{j1}x_{i2}^2x_{j2}^2 + 3x_{i1}^2x_{j1}^2x_{i2}x_{j2} + 3x_{i1}^2x_{j1}^2$$
$$+3x_{i2}^2x_{j2}^2 + 6x_{i1}x_{j1}x_{i2}x_{j2} + 3x_{i1}x_{j1} + 3x_{i2}x_{j2} + 1$$

From the calculation above, we can guess the corresponding Φ function like:

$$\Phi(x_i) \cdot \Phi(x_j) = (x_{i1}^3, x_{i2}^3, \sqrt{3}x_{i1}x_{i2}^2, \sqrt{3}x_{i1}^2x_{i2}, \sqrt{3}x_{i1}^2, \sqrt{3}x_{i2}^2, \sqrt{6}x_{i1}x_{i2}, \sqrt{3}x_{i1}, \sqrt{3}x_{i2}, 1)$$
$$\cdot(x_{j1}^3, x_{j2}^3, \sqrt{3}x_{j1}x_{j2}^2, \sqrt{3}x_{j1}^2x_{j2}, \sqrt{3}x_{j1}^2, \sqrt{3}x_{j2}^2, \sqrt{6}x_{j1}x_{j2}, \sqrt{3}x_{j1}, \sqrt{3}x_{j2}, 1)$$

Since the product of two Φ functions should hold the cubic kernel calculation above. Thus, we finally get:

$$\Phi(\mathbf{x}) = (x_1^3, x_2^3, \sqrt{3}x_1^2x_2, \sqrt{3}x_1x_2^2, \sqrt{3}x_1^2, \sqrt{3}x_2^2, \sqrt{3}x_1, \sqrt{3}x_2, \sqrt{6}x_1x_2, 1)$$

2. (Kernel or not). (5 pts) Suppose that K_1 and K_2 are kernels with feature maps ϕ_1 and ϕ_2 respectively. Is function $K(\mathbf{x}, \mathbf{z}) = c_1K_1(\mathbf{x}, \mathbf{z}) + c_2K_2(\mathbf{x}, \mathbf{z})$ for $c_1, c_2 > 0$ a kernel function? If your answer is yes, write down the corresponding ϕ in terms of ϕ_1 and ϕ_2. If not, provide a proof or explain why.

Your answer goes here.

3. Kernelizing Logistic Regression (10 pts) For this 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 $\mathbf{w} = \mathbf{0}$;
while *not converged* **do**
  for $i = 1, ..., N$ **do**
    $\mathbf{w} \leftarrow \mathbf{w} + \gamma(y_i - \sigma(\mathbf{w}^T\mathbf{x}_i))\mathbf{x}_i$
  end
end

Specifically, please:

(a) (4 pts) 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)
Your answer goes here.

(b) (6 pts) Modify the following stochastic gradient descent algorithm logistic regression algorithm to kernelize it. (Hint: similar to the bottom algorithm on slide 14 of the kernel method lecture, but instead of counter, you will learn a continuous weights for α's)
Your answer goes here.

4. (Hard margin SVM) (8 pts) Apply linear SVM without soft margin to the following problem.

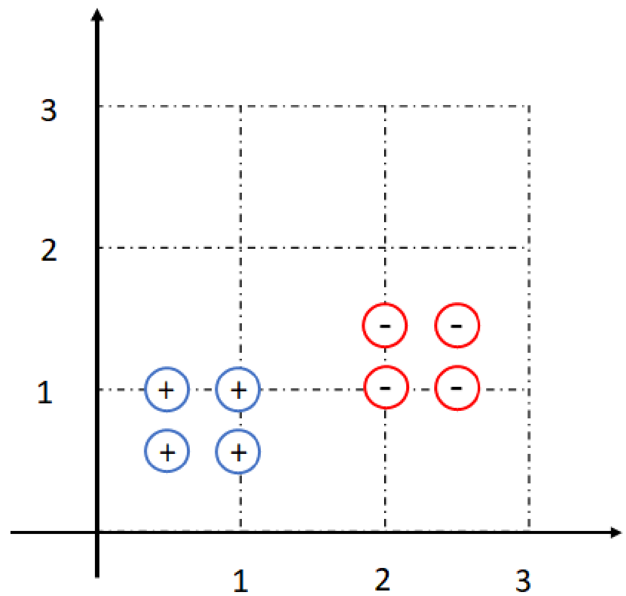

Figure 1:

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.

Your answer goes here.

b. (5 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 $= 1$.
Your answer goes here.

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.

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

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 $\xi_i^* < 0$.)
Your answer goes here

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.
Your answer goes here

c. (8pts) Derive the dual of this problem. How is it different from the standard SVM with hinge loss? Which formulation is more sensitive to outliers?
Your answer goes here