Machine Learning and Applications

Assignment 1

1. (3) Soft margin hyperplanes. The function of the slack variables used in the optimmization problem for soft margin hyperplanes has the form: $\xi \to \sum_{i=1}^{m} \xi_i$. Instead, we could use $\xi \to \sum_{i=1}^{m} \xi_i^p$, with $p > 1$.

   (a) Give the dual formulation of the problem in this general case.

   (b) How does this more general formulation ($p > 1$) compare to the standard setting ($p = 1$)? In the case $p = 2$ is the optimization still convex?

   Sparse SVM. One can give two types of arguments in favor of the SVM algorithm: one based on the sparsity of the support vectors, another based on the notion of margin. Suppose that instead of maximizing the margin, we choose instead to maximize sparsity by minimizing the L_p norm of the vector $\boldsymbol{\alpha}$ that defines the weight vector $\mathbf{w}$, for some $p \geq 1$. First, consider the case $p = 2$. This gives the following optimization problem:

$$\min_{\boldsymbol{\alpha},b} \frac{1}{2} \sum_{i=1}^{m} \alpha_i^2 + C \sum_{i=1}^{m} \xi_i \tag{1}$$

$$\text{subject to } y_i \left(\sum_{j=1}^{m} \alpha_j y_j \mathbf{x}_i \cdot \mathbf{x}_j + b \right) \geq 1 - \xi_i, i \in [1, m]$$

$$\xi_i, \alpha_i \geq 0, i \in [1, m].$$

   (a) Show that modulo the non-negativity constraint on $\boldsymbol{\alpha}$, the problem coincides with an instance of the primal optimization problem of SVM.

   (b) Derive the dual optimization of problem of 1.

   (c) Setting $p = 1$ will induce a more sparse $\boldsymbol{\alpha}$. Derive the dual optimization in this case

2. (2) Importance weighted SVM. Suppose you wish to use SVMs to solve a learning problem where some training data points are more important than others. More formally, assume that each training point consists of a triplet (x_i, y_i, p_i), where $0 \leq p_i \leq 1$ is the importance of the i-th point. Rewrite the primal SVM constrained optimization problem so that the penalty for mis-labeling a point x_i is scaled by the priority p_i . Then carry this modification through the derivation of the dual solution.

3. (3) Show that the following kernels K are PDS:

   (a) $K(x, y) = \cos(x - y)$ over $\mathbb{R} \times \mathbb{R}$.

   (b) $K(x, y) = (x + y)^{-1}$ over $(0, +\infty) \times (0, +\infty)$.

   (c) $\forall \lambda > 0, K(x, y) = \exp(-\lambda[\sin(y - x)]^2)$ over $\mathbb{R} \times \mathbb{R}$ (*Hint:* rewrite $[\sin(y - x)]^2$) as the square of the norm of the difference of two vectors.)

4. (2) Show that the following kernels K are NDS:

   (a) $K(x, y) = [\sin(x - y)]^2$ over $\mathbb{R} \times \mathbb{R}$.

   (b) $K(x, y) = \log(x + y)$ over $(0, +\infty) \times (0, +\infty)$.

5. (2) Explicit polynomial kernel mapping. Let K be a polynomial kernel of degree d, i.e., $K : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}, K(\mathbf{x}, \mathbf{x}') = (\mathbf{x} \cdot \mathbf{x}' + c)^d$, with $c > 0$. Show that the dimension of the feature space associated to K is

$$\binom{N + d}{d}$$

   Write K in terms of kernels $k_i : (\mathbf{x}, \mathbf{x}') \to (\mathbf{x} \cdot \mathbf{x}')^i, i \in [0, d]$. What is the weight assigned to each k_i in that expression? How does it vary as a function of c?