Introduction to Machine Learning, Fall 2013

Problem Set 2: Support vector machines

Due: Tuesday, September 24, 2013 at 11am (*before* class begins)

Important: *See problem set policy on the course web site.*

---

1. (5 points) Consider a (hard margin) support vector machine and the following training data from two classes:

$$\begin{array}{lccc} +1: & (2,2) & (4,4) & (4,0) \\ -1: & (0,0) & (2,0) & (0,2) \end{array}$$

   (a) Plot these six training points, and construct by inspection the weight vector for the optimal hyperplane. In your solution, specify the hyperplane in terms of $\vec{w}$ and b such that $w_1 x_1 + w_2 x_2 + b = 0$. Calculate what the margin is (i.e., 2γ, where γ is the distance from the hyperplane to its closest data point), showing all of your work.

   (b) What are the support vectors? Explain why.

2. (5 points) Show that, irrespective of the dimensionality of the data space, a data set consisting of just two data points (call them $\vec{x}_1$ and $\vec{x}_2$), one from each class, is sufficient to determine the maximum-margin hyperplane. Fully explain your answer, including giving an explicit formula for the solution to the hard margin SVM as a function of $\vec{x}_1$ and $\vec{x}_2$.

3. (5 points) In Lecture 4 we introduced the Multi-class SVM (slide 9) to generalize the binary SVM to multi-class classification. This involved introducing parameters $\vec{w}^{(k)}$ and $b^{(k)}$ for each class $k = 1, \ldots, K$ (where K is the number of classes), and performing prediction for a new data point $\vec{x}$ using

$$\hat{y} \leftarrow \arg\max_k \ \vec{w}^{(k)} \cdot \vec{x} + b^{(k)}$$

   For this problem, prove that this is equivalent to the binary prediction rule $\text{sign}(\vec{w} \cdot \vec{x} + b)$ in the case that $K = 2$. That is, suppose the data is separable and that $\hat{y} \leftarrow \arg\max_{k \in \{1,2\}} \ \vec{w}^{(k)} \cdot \vec{x} + b^{(k)}$ predicts the correct label for all data points $\vec{x}$. Demonstrate $\vec{w}$ and b (as a function of $\vec{w}^{(1)}$, $b^{(1)}$, $\vec{w}^{(2)}$ and $b^{(2)}$) that gives an equivalent decision rule. As always, you must show all of your work to obtain full credit.

4. (10 points) Kernels

   (a) For any two documents x and z, define $k(x, z)$ to equal the number of unique words that occur in both x and z (i.e., the size of the intersection of the sets of words in the two documents). Is this function a kernel? Justify your answer. (Hint: $k(x, z)$ is a kernel if there exists $\phi(x)$ such that $k(x, z) = \phi(x)^T \phi(z)$).

   (b) Assuming that $\vec{x} = [x_1, x_2], \vec{z} = [z_1, z_2]$ (i.e., both vectors are two-dimensional) and $\beta > 0$, show that the following is a kernel:

$$k_\beta(\vec{x}, \vec{z}) = (1 + \beta \vec{x} \cdot \vec{z})^2 - 1$$

   Do so by demonstrating a feature mapping $\Phi(\vec{x})$ such that $k_\beta(\vec{x}, \vec{z}) = \Phi(\vec{x}) \cdot \Phi(\vec{z})$.

(c) One way to construct kernels is to build them from simpler ones. Assuming $k_1(x, z)$ and $k_2(x, z)$ are kernels, then one can show that so are these:

i. (scaling) $f(x)f(z)k_1(x, z)$ for any function $f(x) \in \mathcal{R}$,
ii. (sum) $k(x, z) = k_1(x, z) + k_2(x, z)$,
iii. (product) $k(x, z) = k_1(x, z)k_2(x, z)$.

Using the above rules and the fact that $k(x, z) = x^T z$ is a kernel, show that the following is also a kernel:

$$\left(1 + \left(\frac{x}{||x||_2}\right)^T \left(\frac{z}{||z||_2}\right)\right)^3.$$