Machine Learning and Computational Statistics, Fall 2014

Problem Set 2: Support vector machines

Due: Friday, February 14, 2014 at 5pm (as a PDF document – scanning hand written solutions is acceptable – sent to akshaykumar@nyu.edu)

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:

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

   (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 (i.e., $\vec{w}$) as a function of $\vec{x}_1$ and $\vec{x}_2$.

3. (5 points) The primal SVM always has a unique solution because of the strict convexity of the optimization objective. By contrast, the dual SVM solution may not be unique. This question will explore the dual hard-margin SVM optimization problem to explain the non-uniqueness of the solution.

   The setting we consider is the following. There are three data points in the training data: $\{(x_1, +1), (x_2, -1), (x_2 - 1)\}$, i.e. one data point of class +1, and two *identical* data points of class -1. We assume that $x_1 \neq x_2$. For this question you should use the Gaussian kernel, $K(\mathbf{x}_i, \mathbf{x}_j) = e^{\frac{-\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}}$. This will simplify the dual objective, although any other valid kernel would also show the non-uniqueness of the dual solution.

   (a) Write down the dual optimization problem for the setting considered. It will have three dual variables $\alpha_1, \alpha_2, \alpha_3$ for each of the three data points. Do not forget the constraints!

   (b) Simplify the dual objective so that it is a function of only one of the three dual variables. Solve the (now 1-dimensional) optimization problem, and then find the optimal $\alpha_1^*, \alpha_2^*, \alpha_3^*$ associated with the three data points. Use this to explain why the dual solution is non-unique.

   (c) Recall that the optimal primal solution can be obtained from the optimal dual solution using $w^* = \sum_j \alpha_j^* y_j x_j$. Using this, show that the non-unique dual solution still describes a unique weight vector (primal solution).

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:

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

5. (5 points) The multi-class SVM generalizes the binary SVM to multi-class classification. This involves introducing a weight vector $\vec{w}^{(k)}$ and $b^{(k)}$ for each class $k = 1, \ldots, K$ (where K is the number of classes). Learning solves the following optimization problem, where there is still only one slack variable ξ_j for each data point, but now there are $K - 1$ constraints per data point:

   $$\min_{\{\vec{w}^{(k)}, b^{(k)}\}} \sum_{k=1}^{K} ||\vec{w}^{(k)}||_2^2 + C \sum_j \xi_j$$

   subject to

   $$\begin{aligned} \vec{w}^{(y_j)} \cdot \vec{x}_j + b^{(y_j)} &\geq \vec{w}^{(k)} \cdot \vec{x}_j + b^{(k)} + 1 - \xi_j \qquad \forall j \text{ and } k \neq y_j \\ \xi_j &\geq 0 \qquad \forall j. \end{aligned}$$

   Prediction for a new data point $\vec{x}$ is performed using the rule

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

   This problem compares the binary prediction rule $\text{sign}(\vec{w} \cdot \vec{x} + b)$ to the multi-class prediction rule in the case that $K = 2$, and shows how to reduce between the two of them.

   (a) Demonstrate $\vec{w}$ and b as a function of $\vec{w}^{(1)}$, $b^{(1)}$, $\vec{w}^{(2)}$ and $b^{(2)}$ such that the predictions made for all data points $\vec{x}$ using the new binary prediction rule are the same as what would have been made using the multi-class prediction rule with $\vec{w}^{(1)}$, $b^{(1)}$, $\vec{w}^{(2)}$.

   (b) Next you should show the converse. Given $\vec{w}$ and b, demonstrate $\vec{w}^{(1)}$, $b^{(1)}$, $\vec{w}^{(2)}$ and $b^{(2)}$ (as a function of $\vec{w}$ and b) such that the predictions made for all data points $\vec{x}$ using the multi-class prediction rule are the same as what would have been made using the binary prediction rule with $\vec{w}$ and b.

   As always, you must show all of your work to obtain full credit.