STAT 881
Spring 2008

HOMEWORK 3

1. Given $\mathcal{D}_n = \{(x_i, y_i) \in \mathbb{R}^d \times \{-1, 1\} \mid i = 1, \ldots, n\}$, the perceptron algorithm finds a separating hyperplane by sequentially updating β and β_0 of a linear classifier, $f(x) = sign(\beta^\top x + \beta_0)$. It consists of two steps.

Step 1. Initialize $\beta^{(0)} = 0$ and $\beta_0^{(0)} = 0$.

Step 2. While there is a misclassified point such that $y_i({\beta^{(m-1)}}^\top x_i + \beta_0^{(m-1)}) \leq 0$ for $m = 1, 2, \ldots$, repeat the following.

- Choose a misclassified point (x_i, y_i).
- Update $\beta^{(m)} = \beta^{(m-1)} + y_i x_i$ and $\beta_0^{(m)} = \beta_0^{(m-1)} + y_i$.

Let $R = \max_i \|x_i\|$. Suppose that $\mathcal{D}_n$ is separable so that for some $w \in \mathbb{R}^d$ with $\|w\| = 1$ and $b \in \mathbb{R}$, $\delta = \min_i y_i(w^\top x_i + b) > 0$.

In this case, we can prove that the algorithm terminates within $\lfloor (R^2 + 1)(b^2 + 1)/\delta^2 \rfloor$ iterations, where $\lfloor z \rfloor$ is the largest integer that does not exceed z.

(a) First show that $\|\beta^{(m)}\|^2 + (\beta_0^{(m)})^2 \leq \|\beta^{(m-1)}\|^2 + (\beta_0^{(m-1)})^2 + R^2 + 1$ and conclude that $\|\beta^{(m)}\|^2 + (\beta_0^{(m)})^2 \leq m(R^2 + 1)$.

(b) Similarly, verify that ${\beta^{(m)}}^\top w + \beta_0^{(m)} b \geq m\delta$.

(c) Combine (a) and (b) to complete the proof.

2. Consider the optimization problem for the linear support vector machine in the nonseparable case: find $f(x) = \beta^\top x + \beta_0$ minimizing

$$\sum_{i=1}^{n} \xi_i + \frac{n\lambda}{2} \|\beta\|^2$$

subject to $\quad 1 - y_i(\beta^\top x_i + \beta_0) \leq \xi_i$ and $\xi_i \geq 0$ for $i = 1, \ldots, n$.

(a) Show that its Lagrangian dual problem is

$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2n\lambda} \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i^\top x_j$$

subject to $\quad 0 \leq \alpha_i \leq 1$ and $\sum_{i=1}^n \alpha_i y_i = 0$ for $i = 1, \ldots, n$.

(b) Verify that the complementarity conditions for the solution, $\alpha = (\alpha_1, \ldots, \alpha_n)^\top$, β, β_0, and $\xi = (\xi_1, \ldots, \xi_n)^\top$ are

$$\alpha_i\{1 - y_i(\beta^\top x_i + \beta_0) - \xi_i\} = 0 \text{ and } (1 - \alpha_i)\xi_i = 0 \text{ for } i = 1, \ldots, n.$$

Also, discuss how to determine β_0 given the dual solution α.

(c) Based on the complementarity conditions in (b), explain why the leave-ith case-out solution $\hat{f}_\lambda^{[-i]}$ predicts the class of a non support vector (x_i, y_i) of $\hat{f}_\lambda$ correctly. Thus, conclude that the leave-one-out error $\sum_{i=1}^n I(y_i \hat{f}_\lambda^{[-i]}(x_i) \leq 0)$, does not exceed the number of the support vectors of $\hat{f}_\lambda$. Here $\hat{f}_\lambda$ is the SVM solution for the full data set given λ while $\hat{f}_\lambda^{[-i]}$ is defined as the minimizer of

$$\sum_{j=1, j\neq i}^{n} \xi_j + \frac{n\lambda}{2}\|\beta\|^2$$

subject to $\quad 1 - y_j(\beta^\top x_j + \beta_0) \leq \xi_j$ and $\xi_j \geq 0$
for $j = 1, \ldots, i-1, i+1, \ldots, n$.

3. Let $\mathcal{X} = \mathbb{R}^d$ and $K(\cdot, \cdot)$ be a nonnegative definite function from $\mathcal{X} \times \mathcal{X}$ to $\mathbb{R}$. Given a kernel function K, define $\mathcal{F}_K$ as a class of kernel-based decision rules of the form

$$f(x) = I(\beta_0 + \sum_{i=1}^{n} \beta_i K(a_i, x) \geq 0),$$

where $n \in \mathbb{N}$ and $a_i \in \mathcal{X}$, $i = 1, \ldots, n$, and $\beta_j \in \mathbb{R}$ for $j = 0, \ldots, n$.

For example, support vector machines are classifiers of this kind.

(a) For the polynomial kernel $K(x_1, x_2) = (x_1^\top x_2)^p$ with a fixed degree p, verify that the V-C dimension of $\mathcal{F}_K$ is $\binom{d+p-1}{p} + 1$.

(b) For the radial basis kernel $K(x_1, x_2) = \exp(-\gamma\|x_1 - x_2\|^2)$ with $\gamma > 0$ (treating γ as a parameter of decision rules), show that the V-C dimension of $\mathcal{F}_K$ is infinity.