

CS281B/Stat241B Homework Assignment 2 (due Tuesday, April 4, 2006)

1. **(Constrained optimization)** The ν -support vector regression method involves the following optimization problem.

$$\begin{aligned} \text{minimize}_{w \in \mathbb{R}^d, \epsilon \in \mathbb{R}} \quad & \frac{1}{2} \|w\|^2 + C \left(\nu \epsilon + \frac{1}{n} \sum_{i=1}^n (|y_i - w'x_i| - \epsilon)_+ \right) \\ \text{subject to} \quad & \epsilon \geq 0. \end{aligned}$$

where $x_1, \dots, x_n \in \mathbb{R}^d$, $y_1, \dots, y_n \in \mathbb{R}$, and $C, \nu > 0$.

The representer theorem shows that the solution satisfies $w = \sum_{i=1}^n \alpha_i x_i$ for some $\alpha_1, \dots, \alpha_n$.

- (a) Show that, if $\epsilon > 0$ at the solution, we have

$$|\{i : |w'x_i - y_i| > \epsilon\}| \leq \nu n \leq |\{i : \alpha_i \neq 0\}|.$$

- (b) Show that the optimization is equivalent to the following.

$$\begin{aligned} \text{minimize}_{\alpha \in \mathbb{R}^n} \quad & \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j x_i' x_j - \sum_i \alpha_i y_i \\ \text{subject to} \quad & \alpha \in S_{C,\nu,n}, \end{aligned}$$

where, for some radius $r(C, \nu, n)$, $S_{C,\nu,n}$ is a certain subset of the l_1 ball,

$$S_{C,\nu,n} \subset \{\alpha : \|\alpha\|_1 \leq r(C, \nu, n)\},$$

with $\|\alpha\|_1 = \sum_i |\alpha_i|$. How does the parameter ν affect $r(C, \nu, n)$?

2. (Kernels)

- (a) Consider the constant function, $k(x, y) = c$ for all x, y . Is k a symmetric positive semidefinite kernel? If not, explain why not. If so, describe its reproducing kernel Hilbert space.
- (b) For two symmetric, positive semidefinite kernels k_1, k_2 defined on the same space, let k be their minimum, $k(u, v) = \min\{k_1(u, v), k_2(u, v)\}$. Is k also a symmetric positive semidefinite kernel?
- (c) If k_1, k_2 are symmetric, positive semidefinite kernels, show that the function k defined by $k(u, v) = k_1(u, v)k_2(u, v)$ is also a symmetric, positive semidefinite kernel.
(Hint: You might use the observation that any kernel matrix is the covariance matrix of some random vector.)

3. **(Implementing an SVM classifier)** Investigate the performance of SVM classifiers on two classification problems. There are several public domain implementations of SVMs available on the web; see, for example, SVMlite (<http://svmlight.joachims.org/>) or many others on the list at <http://www.kernel-machines.org/software.html>. Consider two data sets:

- (a) Simulate a two-class classification problem that is easy to visualize. For example, consider a uniform distribution on a subset of \mathbb{R}^2 , with $\Pr(Y = 1|X = x) = 0.05 + 0.9 \times \mathbf{1}[w^T x > 0]$. Use a soft-margin SVM. Investigate the effects of the kernel and the regularization coefficient (for example, the constant C in the standard formulation) on the decision boundary, the number of support vectors obtained, and the misclassification probability.
- (b) Download a public domain data set for a binary classification problem that interests you. Such data sets are available in many places, such as the Delve (<http://www.cs.toronto.edu/~delve/>) or UCI (<http://www.ics.uci.edu/~mllearn/MLRepository.html>) repositories. Compare the performance of an SVM classifier with that of a simple pattern classification technique (such as logistic regression, linear discriminant analysis, or nearest neighbor). Split the data and use a hold-out set to estimate the misclassification probability of the classifier.

4. **(Greedy optimization methods)** Consider the cost function

$$\phi(\alpha) = (\max\{0, 1 - \alpha\})^3.$$

- (a) Does minimization of $\mathbb{E}(\phi(Yf(X))|X = x)$ for each x lead to a classifier $\text{sign}(f)$ which achieves the Bayes risk?
- (b) Give an upper bound on the excess risk of a function f in terms of its excess ϕ -risk.
- (c) Design an (Adaboost-like) algorithm that chooses a function F_T from $\text{span}(G)$ for some class of classifiers G , and aims to minimize the empirical ϕ -risk, $\hat{\mathbb{E}}(\phi(YF_T(X)))$. That is, find an algorithm that, at iteration t , updates a probability distribution D_t over the training data and a combination F_t of classifiers from G by choosing an $\alpha_t \in \mathbb{R}$ and an $f_t \in G$ so that $(f_t(x_1), \dots, f_t(x_n))$ is in the direction of steepest descent of

$$J(v) = \sum_{i=1}^n \phi(y_i(v_i + F_{t-1}(x_i)))$$

and α_t minimizes $\hat{\mathbb{E}}(\phi(Y(F_{t-1}(X) + \alpha_t f_t(X))))$.

- (d) Implement this algorithm, and investigate its performance on the following two-class classification problem: Let $\mathcal{X} = \mathbb{R}$ and suppose that X is uniform on $[0, 1]$ and $\Pr(Y = 1|X = x) = x$. Let the class G of base classifiers be decision stumps. Use a hold-out set to decide when to stop, by minimizing an estimate of the ϕ -risk. Let f_n be the classifier that you arrive at via this method for a sample of size n . Does the risk of f_n appear to approach the Bayes risk? What does f_n converge to? Why?