Homework Set Three

ECE 271B
Department of Computer and Electrical Engineering
University of California, San Diego

Nuno Vasconcelos Spring 2007

Due June 7, 2007

1. In this problem we derive some generic results on local optima of functions defined over convex sets.

a) Consider a convex set X and a continously differentiable function f, and show that the following hold:

1. If x^* is a local minimum of f over X, then
$$\nabla f(x^*)^T(x - x^*) \geq 0, \, \forall x \in X.$$
2. If f is convex over X, then the condition of 1. is also sufficient for x^* to minimize f over X.

(Hint: For 1 you might want to use the mean value theorem, which states that if $f : \mathbb{R} \to \mathbb{R}$ is differentiable over an interval I, $\forall x, y \in I$ there is some $\gamma \in [x, y]$ such that

$$f(y) - f(x) = \frac{\partial f}{\partial x}(\gamma)(y - x).$$

For 2 you may want to start by proving that f is convex if and only if

$$f(z) \geq f(x) + (z - x)^T \nabla f(x), \, \forall x, z \in X.$$

)

b) Using **a)** (and **a)** alone) show that if $X = \{x | x \geq 0\}$, the necessary conditions for $x^* = (x_1^*, \ldots, x_n^*)$ to be a minimum are that

$$\frac{\partial f}{\partial x_i}(x^*) \geq 0, \forall i$$

and

$$\frac{\partial f}{\partial x_i}(x^*) = 0, \text{if } x_i^* > 0.$$

c) Provide a geometric interpretation to the results of **b)** and generalize to the case of **a)**. Why is the constraint that X is convex necessary?

2. In this problem we consider some minimization problems with inequality constraints.

a) Consider a vector $\mathbf{x} \in \mathbb{R}^n$ and find the optimal solution to the problem

$$\min \frac{1}{2}||\mathbf{x}||^2$$

$$\text{subject to} \sum_i x_i \leq -3.$$

b) Prove the inequality

$$(\mathbf{x}^T\mathbf{y})^2 \leq (\mathbf{x}^T\mathbf{Q}\mathbf{x})(\mathbf{y}^T\mathbf{Q}^{-1}\mathbf{y})$$

where $\mathbf{Q}$ is a positive definite symmetric matrix.
(Hint: you may want to consider the problem

$$\max_{\mathbf{x}} \mathbf{y}^T\mathbf{x}$$

$$\text{subject to } \mathbf{x}^T\mathbf{Q}\mathbf{x} \leq 1.$$

)

3. In this problem we study the topic of duality, by considering some simple examples. In all cases, the minimization problems are of the form

$$\min_{x \in X} f(x)$$

$$\text{subject to } g(x) \leq 0$$

. We consider six problems, described by the table below. For each problem answer the following.

Problem	$f(x)$	$g(x)$	X
1	$x_1 - x_2$	$x_1 + x_2 - 1$	$\{(x_1, x_2) \| x_1 \geq 0, x_2 \geq 0\}$
2	x	x^2	$\mathbb{R}$
3	$-x$	$x - \frac{1}{2}$	{0,1}
4	$-x$	$x - \frac{1}{2}$	[0,1]
5	$\frac{1}{2}(x_1^2 + x_2^2)$	$x_1 - 1$	$\mathbb{R}^2$
6	$\|x_1\| + x_2$	x_1	$\{(x_1, x_2) \| x_2 \geq 0\}$

1. Sketch the set of feasible solutions $S = \{g(x), f(x) | x \in X\}$ in (g, f)-space.
2. Determine if there exists a Lagrange multiplier. If so, what is the set of Lagrange multipliers? If not, why is there no multiplier?
3. State the dual problem, and sketch a plot of $q(\mu)$ as a function of μ. Determine if there is a duality gap.

4. In class, we have studied Vapnik's SVM formulation, i.e. the search for the hyperplane that solves the following optimization problem

$$\min_{\mathbf{w},\xi} \frac{1}{2}||\mathbf{w}||^2 + \frac{C}{n}\sum_{i=1}^{n} \xi_i$$

subject to

$$\begin{aligned} y_i(<\mathbf{x}_i, \mathbf{w}> +b) &\geq 1 - \xi_i, \quad i = 1, \ldots, n \\ \xi_i &\geq 0, \quad i = 1, \ldots, n. \end{aligned}$$

One limitation of this formulation is that there is no intuition for what the parameter C means and it can therefore be difficult to find good values for it in practice. In this problem we consider a slightly different, but more intuitive formulation, based on the solution of the following problem

$$\min_{\mathbf{w},\xi,\rho,b} \frac{1}{2}||\mathbf{w}||^2 - \nu\rho + \frac{1}{n}\sum_{i=1}^{n} \xi_i$$

subject to

$$\begin{aligned} y_i(<\mathbf{x}_i, \mathbf{w}> +b) &\geq \rho - \xi_i, \quad i = 1, \ldots, n \\ \xi_i &\geq 0, \quad i = 1, \ldots, n \\ \rho &\geq 0. \end{aligned}$$

a) Determine the dual problem for this case, and the form of the resulting decision function.

b) Given the dual solution how would you determine the values of b and ρ?

c) Define the fraction of margin errors as

$$\epsilon_\rho = \frac{1}{n}|\{i|y_i g(\mathbf{x}_i) < \rho\}|$$

and suppose thet we solve the optimization problem on a dataset with the result that $\rho > 0$. Show that

1. ν is an upper bound on ϵ_ρ.
2. ν is a lower bound on the fraction of vectors that are support vectors.

d) Show that if the solution of the second problem leads to $\rho > 0$, then the first problem with C set a priori to $\frac{1}{\rho}$ leads to the same decision function.

5. In class we have dealt mostly with classification problems. One alternative problem that is quite frequent in practice is the so-called problem of *novelty detection*. This happens when only data from one class is available (there are no negative examples) but the goal is to detect outliers. The novelty detection problem can be solved in at least two ways, using SVM related ideas. They both imply finding the region of support of the data by finding a minimal boundary surface that includes the data from the class. The first strategy is to find the plane the separates, with largest margin, the training points from the origin. It consists of solving the following optimization problem

$$\min_{\mathbf{w},\xi,\rho} \frac{1}{2}||\mathbf{w}||^2 + \frac{1}{\nu n}\sum_{i=1}^{n} \xi_i - \rho$$

subject to

$$\begin{aligned} < \Phi(\mathbf{x}_i), \mathbf{w} > \geq \rho - \xi_i, &\quad i = 1, \dots, n \\ \xi_i \geq 0, &\quad i = 1, \dots, n \end{aligned}$$

The second strategy consists of finding the smallest ball that encircles the data points. This can be done by solving the following problem

$$\min_{R,\xi,c} R^2 + \frac{1}{\nu n}\sum_{i=1}^{n} \xi_i, \; \nu \in [0, 1]$$

subject to

$$\begin{aligned} ||\Phi(\mathbf{x}_i) - c||^2 \; \leq R^2 + \xi_i, &\quad i = 1, \dots, n \\ \xi_i \geq 0, &\quad i = 1, \dots, n. \end{aligned}$$

a) Determine the dual problem associated with each of the problems above, as well as the resulting decision function.

b) How would you recover the ρ parameter in the case of problem 1? Also, in this problem, what happens when ν approaches zero?

c) Determine the set of kernels that make the two problems equivalent.

d) In the case of **c)**, compare the decision function with a Parzen density estimate

$$f(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} \phi(\mathbf{x} - \mathbf{x}_i)$$

where ϕ is a non-negative function that integrates to one.