Introduction to Machine Learning	**Fall Semester**
Recitation 5	
Lecturer: Regev Schweiger	*Scribe: Regev Schweiger*

5.1 Constrained Optimization

5.1.1 Lagrange Multipliers

Suppose we have the following problem of constrained optimization with equality constraints:

$$\begin{aligned} &\min_{\mathbf{x}\in\mathrm{R}^d} f(\mathbf{x}) \\ &\quad \text{s.t.} \\ &g_i(\mathbf{x}) = 0 \ \text{ for } i = 1, \ldots, N \end{aligned}$$

Define the *Lagrangian* of this problem as follows:

$$\mathcal{L}(\mathbf{x}, \lambda_1, \ldots, \lambda_N) = f(\mathbf{x}) + \lambda_1 g_1(\mathbf{x}) + \ldots + \lambda_N g_N(\mathbf{x})$$

Under some regularity requirements, a necessary condition for $\mathbf{x}^*$ being a critical point for f under the constraints, it that there exist values $\lambda_1^*, \ldots, \lambda_N^*$, such that $(\mathbf{x}^*, \lambda_1^*, \ldots, \lambda_N^*)$ is a critical point of $\mathcal{L}$ (unconstrained); and specifically, that the gradient of the Lagrangian at the point is 0:

$$\nabla\mathcal{L}(\mathbf{x}^*, \lambda_1^*, \ldots, \lambda_N^*) = \mathbf{0}$$

We will not prove this here.

5.1.2 Karush-Kuhn-Tucker Conditions

The KKT conditions extend the ideas of Lagrange multipliers to handle inequality constraints in addition to equality constraints. While there is a general formulation including both equalities and inequalities, we will limit ourselves here1 to provide a first-order optimality condition for the problem:

$$\begin{aligned} &\min_{\mathbf{x}\in\mathrm{R}^d} f(\mathbf{x}) \\ &\quad \text{s.t.} \\ &g_i(\mathbf{x}) \leq 0 \ \text{ for } i = 1, \ldots, N \end{aligned}$$

The Lagrangian is again defined as

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\alpha}) = f(\mathbf{x}) + \sum_{i=1}^{N} \alpha_i g_i(\mathbf{x})$$

where $\boldsymbol{\alpha}$ are called the *Lagrangian multipliers*. The theory (on which we do not elaborate here) tells us that the solution, $\boldsymbol{\alpha}^*$, to the dual program:

$$\begin{aligned} &\max_{\boldsymbol{\alpha}} \min_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\alpha}) \\ &\quad \text{s.t.} \;\; \alpha_i \geq 0 \qquad \forall i = 1, \ldots, N \end{aligned}$$

achieves the same optimal value (in all the cases which we will consider):

$$\min_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\alpha}^*) = f(\mathbf{x}^*)$$

In order to find the optimal point $\mathbf{x}^*$ in which this optimal value $f(\mathbf{x}^*)$ is attained, we use the fact that $\mathbf{x}^*, \boldsymbol{\alpha}^*$ must satisfy the KKT conditions:

$$\begin{aligned} \nabla \mathcal{L}(\mathbf{x}^*, \boldsymbol{\alpha}^*) &= \mathbf{0} \\ \boldsymbol{\alpha}^* &\geq \mathbf{0} \\ g_i(\mathbf{x}^*) &\leq 0 \qquad \forall i = 1, \ldots, N \\ \alpha_i = 0 \vee g_i(\mathbf{x}^*) &= 0 \qquad \forall i = 1, \ldots, N \end{aligned}$$

Example. As an example, let us solve the problem

$$\begin{aligned} &\min_{(x_1, x_2, x_3) \in \mathrm{R}^2} x_1^2 + x_2^2 + x_3 \\ &\quad \text{s.t.} \quad 2x_1 + 2x_2 \geq 1 \\ &\qquad\quad x_3 \geq 1 \end{aligned}$$

The Lagrangian is:

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\alpha}) = x_1^2 + x_2^2 + x_3 + \alpha_1(1 - 2x_1 - 2x_2) + \alpha_2(1 - x_3)$$

For a fixed $\boldsymbol{\alpha}$, we want to calculate $\min_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\alpha})$. To do that, we equate the derivative to zero:

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial x_1} &= 2x_1 - 2\alpha_1 = 0 \Rightarrow x_1 = \alpha_1 \\ \frac{\partial \mathcal{L}}{\partial x_2} &= 2x_2 - 2\alpha_1 = 0 \Rightarrow x_2 = \alpha_1 \end{aligned}$$

the second derivative is positive, so this is indeed a minimum. For x_3, we encounter a different scenario. The derivative is:

$$\frac{\partial \mathcal{L}}{\partial x_3} = 1 - \alpha_2 = 0$$

The derivative should still be 0, but it does not depend on x_3, so we cannot get it as a function of the α-s here. This means, that when the gradient is positive ($1 - \alpha_2 \geq 0 \Rightarrow \alpha_2 \leq 1$), we can take arbitrarily small x_3 and $\mathcal{L}$ will continue to decrease; therefore, if $\alpha \leq 1$, then $\min_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\alpha}) = -\infty$. Similarly, if the gradient is negative, we can take arbitrarily large values of x_3 to decrease that value of $\mathcal{L}$. Therefore, we conclude that a (finite) minimum exists if and only if $\alpha_2 = 1$. Since we are interested in the maximal value out of all of these minima, we conclude that we can limit ourselves only to this value of α_2, and instead solve the problem:

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \min_{\mathbf{x}} & \mathcal{L}(\mathbf{x}, \boldsymbol{\alpha}) \\ \text{s.t.} \quad & \alpha_1 \geq 0 \\ & \alpha_2 = 1 \end{aligned}$$

Note that the additional constraint $\alpha_2 \geq 0$ is satisfied automatically by the more stringent constraint, $\alpha_2 = 1$. When $\alpha_2 = 1$, the value of the Lagrangian is:

$$\begin{aligned} \mathcal{L}(\mathbf{x}, \boldsymbol{\alpha}) &= \alpha_1^2 + \alpha_1^2 + x_3 + \alpha_1(1 - 2\alpha_1 - 2\alpha_1) + \alpha_2(1 - x_3) \\ &= -2\alpha_1^2 + \alpha_1 + \alpha_2 + x_3(1 - \alpha_2) \\ &= -2\alpha_1^2 + \alpha_1 + \alpha_2 \end{aligned}$$

as expected, this expression should not depend on x_3, and it indeed zeros out. Finding the maximum over α_1, subject to $\alpha_1 \geq 0$, gives $\alpha_1^* = 1/4$, along with $\alpha_2^* = 1$, which we already found out. Therefore, the maximal value the dual program attains is $-2\left(\frac{1}{4}\right)^2 + \left(\frac{1}{4}\right) + 1 = \frac{9}{8}$, and this is the minimum value our original optimization problem obtains. To find the optimal point $\mathbf{x}$ in which the optimal value is attained, we utilize the additional KKT condition (complementary slackness), that $\alpha_2 = 0 \vee 1 - x_3 = 0$, which in turn means that, since $\alpha_2 \neq 0$, that $x_3 = 1$. Therefore, the optimal points are $x_1^* = x_2^* = 1/4, x_3^* = 1$. We can verify that indeed $f(\mathbf{x}^*) = 9/8$ as expected.

5.2 SVM - Unrealizable Case

In the lecture we saw the following optimization problem, for a maximum margin classifier with possible margin errors. We have,

$$\begin{aligned} &\min_{\mathbf{w},b,\boldsymbol{\xi}} \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{n=1}^{N}\xi_n \\ &\text{s.t. } y_n(\mathbf{w}^T\mathbf{x}_n + b) \geq 1 - \xi_n \qquad \forall n = 1,\ldots,N \\ &\quad \xi_n \geq 0 \end{aligned}$$

where we have added slack variables ξ_n to ensure feasibility. The first step is to write the Lagrangian:

$$\mathcal{L}(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \mathbf{r}) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{n=1}^{N}\xi_n - \sum_{n=1}^{N}\alpha_n(y_n(\mathbf{w}^T\mathbf{x}_n + b) - 1 + \xi_n) - \sum_{n=1}^{N} r_n\xi_n$$

The first step is to assumed $\boldsymbol{\alpha}$ and $\mathbf{r}$ are fixed, and minimize over $\mathbf{w}, b$ and $\boldsymbol{\xi}$. We now take the derivative of $\mathcal{L}$ and equate it with zero to minimize over $\mathbf{w}, b$ and $\boldsymbol{\xi}$.

$$\nabla_{\mathbf{w}}\mathcal{L} = \mathbf{w} - \sum_{n=1}^{N}\alpha_n y_n \mathbf{x}_n = 0 \quad \Longrightarrow \quad \mathbf{w}^* = \sum_{n=1}^{N}\alpha_n y_n \mathbf{x}_n$$

this gives us a way to compute the $\mathbf{w}$ that achieves the minimal point, given $\boldsymbol{\alpha}$. We call this the $\mathbf{w}$-constraint. For b we have

$$\frac{\partial\mathcal{L}}{\partial b} = -\sum_{n=1}^{N}\alpha_n y_n = 0 \quad \Longrightarrow \quad \sum_{n=1}^{N}\alpha_n y_n = 0$$

We call this the b-constraint. This effectively tells us that there are two classes of $\boldsymbol{\alpha}$, and that the behavior of the Lagrangian's minimal point differs between them. If $\sum_{n=1}^{N}\alpha_n y_n \neq 0$, then there is no minimal point; we can take arbitrarily large (or small, depending on the sign of $\sum_{n=1}^{N}\alpha_n y_n$) values of b. Therefore, the minimum value (technically, infimum), is $-\infty$. However, when $\sum_{n=1}^{N}\alpha_n y_n = 0$, then the value of b doesn't matter, so there is an finite minimum. This stems from the fact that the Lagrangian is a linear function of b. Since we are interested, at the next step, at the maximum over all of these values, we are not interested in the case $\mathcal{L}(\mathbf{w}^*, b^*, \boldsymbol{\xi}^*, \boldsymbol{\alpha}, \mathbf{r}) = -\infty$, so we limit ourselves only to the case $\sum_{n=1}^{N}\alpha_n y_n = 0$.

For ξ_n we have

$$\frac{\partial\mathcal{L}}{\partial\xi_n} = C - \alpha_n - r_n = 0 \quad \Longrightarrow \quad \alpha_n = C - r_n$$

Substituting the constraints in $\mathcal{L}$, we get

$$\begin{aligned}\mathcal{L}(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \mathbf{r}) =& \frac{1}{2}\mathbf{w}^T\mathbf{w} - \mathbf{w}^T \underbrace{\left(\sum_{n=1}^{N} \alpha_n y_n \mathbf{x}_n\right)}_{\mathbf{w}} - b \underbrace{\left(\sum_{n=1}^{N} \alpha_n y_n\right)}_{0} + \left(\sum_{n=1}^{N} \alpha_n\right) + \sum_{n=1}^{N} \xi_n \underbrace{(C - \alpha_n - r_n)}_{0} \\ =& -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j + \sum_{n=1}^{N} \alpha_n\end{aligned}$$

identically to the realizable case. The only difference is that now we have two additional constraints, $r_n \geq 0$ and $\alpha_n = C - r_n$. Since r_n does not appear in the optimization, we can drop it, and join the two constraints to $\alpha_n \leq C$. (For any solution of α_n we can set $r_n = C - \alpha_n$.)

Formally, the dual problem is

$$\begin{aligned}\max_{\boldsymbol{\alpha},\mathbf{r}} \min_{\mathbf{w},b,\boldsymbol{\xi}} \mathcal{L}(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \mathbf{r}) = \min_{\boldsymbol{\alpha},\mathbf{r}} & \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j - \sum_{n=1}^{N} \alpha_n \\ \text{s.t.} & \sum_{n=1}^{N} \alpha_n y_n = 0 \\ & \forall n \;\; C \geq \alpha_n \geq 0\end{aligned}$$

Note that we changed the sign and turned the problem from max to min. This is an instance of quadratic programming, for which there are efficient algorithms, additionally, it is easy to see that the program is convex - see lesson scribe for a proof.

Extracting the optimal point. Suppose we solved the dual problem, and got a solution $\boldsymbol{\alpha}^*, \mathbf{r}^*$. How do we get the solution for the original problem? For $\mathbf{w}^*$, recall we have the w-constraint $\mathbf{w}^*(\boldsymbol{\alpha}) = \sum_{n=1}^{N} \alpha_n y_n \mathbf{x}_n$, which gives us an explicit formula of $\mathbf{w}$ as a function of the Lagrange multipliers.

When $\alpha_n^* > 0$ (for a specific n), this means the constraint $y_n(\mathbf{w}^T\mathbf{x}_n + b) = 1$ must be satisfied, due to the KKT conditions. Therefore, the support vectors are those with $\alpha_n^* > 0$. This allows us also to get the solution for b^* - choose an n for which $\alpha_n^* > 0$; then $b^* = y_n - (\mathbf{w}^*)^T\mathbf{x}_n$.

Support vectors. As mentioend, the support vectors are those with $\alpha_n^* > 0$. Note that when we have an error in classification or in the margin, then $\xi_n > 0$ and therefore $r_n = 0$, which implies that $\alpha_n = C$. If $C > \alpha_n > 0$, this means as before that $y_n(\mathbf{w}^T\mathbf{x}_n + b) = 1$ and $\xi_n = 0$, and thus $\mathbf{x}_n$ is a support vector.