150AML: Advanced Topics in Machine Learning	**Spring 2008**
Course Topic: Computational Learning Theory	
Department of Computer Science, Tufts University	
Lecture 21-22	**April, 10**
Instructor: Roni Khardon	**Scribe: Bilal**

Deriving Support Vector Machines: Optimizing functions under constraints

1 Maximum Margin Classification

For a given training set, the margin of the i^{th} sample is given as,

$$\gamma_i = y_i(w \cdot x_i - \theta) \tag{1}$$

The problem of finding the classifier with the maximum margin can be formulated as an optimization problem,

$$\begin{aligned} &\max_{w,\theta} \quad \min_{i} \quad \gamma_i \\ &s.t. \quad \|w\| = 1 \end{aligned} \tag{2}$$

Let γ be the maximum margin of the training set, then by (1),

$$\begin{aligned} &\forall i \quad y_i(w \cdot x_i - \theta) \geq \gamma \\ &\Leftrightarrow \quad y_i(\frac{w}{\gamma} \cdot x_i - \frac{\theta}{\gamma}) \geq 1 \\ &\Rightarrow \quad y_i(V \cdot x_i - T) \geq 1 \end{aligned}$$

By this formulation we have,

$$\begin{aligned} \|V\|^2 &= \left\| \frac{1}{\gamma} w \right\|^2 \\ &= \frac{1}{\gamma^2} \|w\|^2 \\ &= \frac{1}{\gamma^2} \quad (\because \|w\|^2 = 1) \end{aligned}$$

Therefore, maximizing $\frac{1}{\gamma^2}$ is equivalent to minimizing $\|V\|^2$. Hence, the original problem (2) is equivalent to,

$$\begin{aligned} &\min \quad \frac{1}{2} \|V\|^2 \\ &s.t. \quad y_i(V \cdot x_i - T) \geq 1 \end{aligned} \tag{3}$$

This formulation of the optimization problem is an example of a constrained convex optimization problem. The objective function (3) is quadratic and the constraints are linear. The next section outlines the basic fundamentals of convex optimization.

2 Optimization Theory

2.1 Unconstrained minimization of a function

Consider the function,

$$\begin{aligned} & f(x) = x^2 - 5x + 6 \\ \Leftrightarrow \quad & f(x) = (x-2)(x-3) \end{aligned} \tag{4}$$

To find the minimum/maximum of this function, find the first derivative of the function and solve for x,

$$\begin{aligned} \frac{d}{dx} f(x) &= 2x - 5 = 0 \\ \Rightarrow x &= \frac{5}{2} \end{aligned}$$

The second derivative indicates whether the extreme point is a minima or a maxima,

$$\frac{d^2}{dx^2} f(x) = 2$$

since the second derivative is positive, this indicates that the point is a minima.

2.2 Minimizing a function with equality constraints

The form for such problems can be written as,

$$\begin{aligned} min \quad & f(w) \\ s.t. \quad & h(w) = 0 \end{aligned} \tag{5}$$

where $f(w)$ is the objective function and $h(w)$ is a single equality constraint. In this formulation we have given only a single constraint; the model is valid for multiple constraints as well. Such problems can be solved by using Lagrange multipliers. To do this simply write the Lagrangian, which in this case takes the form

$$L(w, \beta) = f(w) + \beta h(w) \tag{6}$$

where β is the Lagrange multiplier.

2.2.1 Theorem

1. if w is a minima then $\frac{\delta}{\delta w} L = 0$ and $\frac{\delta}{\delta \beta} L = 0$
2. if $f()$ is convex then $\frac{\delta}{\delta w} L = 0$ and $\frac{\delta}{\delta \beta} L = 0$ imply that w is a minima.

2.2.2 Example

Consider the case of estimating the maximum likelihood solution for multinomial variables.

- We have a k-sided die, where each face has the probability P_k
- Roll the die N times and get the observations V_i.

The likelihood function is given as,

$$\begin{aligned}\text{Likelihood} = P(data) &= \prod_{i=1}^{N} P(V_i) \\ &= \prod_{j=1}^{k} P_j^{n_i}\end{aligned}$$

where n_i is the number of times the i^{th} face is observed. since maximizing the likelihood is equivalent to maximizing the log-likelihood, we can formulate the overall problem as,

$$\begin{aligned}\max_{P_1,\ldots,P_k} \quad & \sum_{i=1}^{k} n_i log P_i \\ \text{s.t.} \quad & \sum_{i=1}^{k} P_i = 1\end{aligned} \tag{7}$$

To solve the multinomial MLE problem, we have from (7)

$$\begin{aligned}\max_{\vec{P}} \quad & f(\vec{P}) = \sum_{i=1}^{k} n_i log p_i \\ s.t. \quad & h(\vec{P}) = (1 - \sum_{i=1}^{k} p_i) = 0\end{aligned}$$

The Lagrangian therefore takes the form,

$$L(\vec{P}, \beta) = \sum_{i=1}^{k} n_i log p_i + \beta(1 - \sum_{i=1}^{k} p_i) \tag{8}$$

According to the theorem stated previously:

$$\begin{aligned}\frac{\delta}{\delta p_l} L =& \frac{n_l}{p_l} - \beta = 0 \\ \Rightarrow \quad & p_l = \frac{n_l}{\beta} \\ \frac{\delta}{\delta \beta} L =& (1 - \sum_{i=1}^{k} p_i) = 0 \\ \Rightarrow \quad & \sum_{i=1}^{k} p_i = 1 \\ \therefore \quad & \sum_{i=1}^{k} \frac{n_i}{\beta} = 1 \\ \Rightarrow \quad \beta =& \sum_{i=1}^{k} n_i = N \\ \therefore \quad & p_l = \frac{n_l}{N}\end{aligned} \tag{9}$$

2.3 Minimizing with equality and inequality constraints

The general form of an optimization problem with both equality and inequality constraints can be stated as,

$$\begin{aligned} & \min f(w) & \\ s.t.\ \ & h_i(w) = 0 & i = 1, \ldots, m_1 \\ & g_j(w) \leq 0 & j = 1, \ldots, m_2 \end{aligned} \tag{10}$$

The Lagrangian of the above given formulation takes the form,

$$L(w, \alpha, \beta) = f(w) + \sum_{i=1}^{m_1} \beta_i h_i(w) + \sum_{j=1}^{m_2} \alpha_j g_j(w) \tag{11}$$

All the optimization problems we have encountered until now are known as being in the primal form and the variables are known as the primal variables. For example, in the case of (10) the objective function is known as the primal objective function and the variable w is the primal variable. Below we introduce the concept of the dual of an optimization problem.

2.4 Dual Objective Function

The dual objective takes the form,

$$\theta(\alpha, \beta) = \min_{w} L(w, \alpha, \beta) \tag{12}$$

The optimization problem can then be stated as,

$$\begin{aligned} \max_{\alpha,\beta} \ & \theta(\alpha, \beta) \\ s.t.\ \ & \alpha_i \geq 0 \end{aligned} \tag{13}$$

2.4.1 Theorem

1. If w and α satisfy the constraints then, $f(w) \geq \theta(\alpha, \beta)$. This is easy to see from the form of (11) since $h_i(w) = 0$, $g_i(w) \leq 0$ and $\alpha_i \geq 0$ implying that $L(w, \alpha, \beta) = f(w) + 0 -$ (non-negative quantity) $\leq f(w)$.

2. For an optimization problem where the objective function is convex and the constraints are linear, $f(w^*) = \theta(\alpha^*, \beta^*)$, where w^*, (α^*, β^*) are the optimal solution for the primal and the dual respectively.

3. The optimal solution is obtained when:

$$\begin{aligned} \frac{\delta}{\delta w} L = 0 \ \ and \ \ \frac{\delta}{\delta \beta} L = 0 \\ \alpha_i \geq 0, \ \ g_i(w) \leq 0 \ \ and \ \ \alpha_i g_i(w) = 0 \end{aligned}$$

According to this theorem if $\alpha_i > 0$ then $g_i(w) = 0$. As an example consider the constraint $z + 10 \leq 0$. If $\alpha(z + 10) = 0, \ \ \alpha \neq 0 \Rightarrow z = -10$. Such constraints are known as active constraints, since $g_i(w)$ is stretched to its maximum value.

2.4.2 Example

Consider the function $f(x) = (x-2)(x-3)$, and let the optimization problem be,

$$\begin{aligned} \min \quad & f(x) \\ s.t. \quad & x \geq 3 \end{aligned}$$

In standard notation the constraint can be written as,

$$g(x) = 3 - x \leq 0$$

The Lagrangian for this problem is,

$$L(x, \alpha) = (x^2 - 5x + 6) + \alpha(3 - x)$$

To calculate the dual objective function we have,

$$\begin{aligned} \theta(\alpha) &= \min_x L(x, \alpha) \\ \frac{\delta}{\delta x} L &= 2x - 5 - \alpha = 0 \\ \Rightarrow \quad x &= \frac{5+\alpha}{2} \end{aligned} \tag{14}$$

substituting this value of x back in the objective function we get,

$$\begin{aligned} & \theta(\alpha) = (\frac{5+\alpha}{2})^2 - 5\frac{(5+\alpha)}{2} + 6 + 3\alpha - \alpha\frac{(5+\alpha)}{2} \\ \Leftrightarrow \quad & \theta(\alpha) = -\frac{1}{4}(\alpha - 1)^2 \end{aligned}$$

therefore the dual problem takes the form,

$$\begin{aligned} \max \theta(\alpha) &= -\frac{1}{4}(\alpha - 1)^2 \\ s.t. \quad \alpha &\geq 0 \end{aligned}$$

the optimizing value is $\alpha = 1$ since for any other value of α the objective function is negative. So from (14) we have,

$$x = \frac{5+1}{2} = 3$$

Notice that we have $\alpha \neq 0$, an active constraint and indeed the inequality constraint is at the boundary.

For further illustration we can change the problem to

$$\begin{aligned} \min \quad & f(x) \\ s.t. \quad & x \geq 2 \end{aligned}$$

after similar manipulation we get $\theta(\alpha) = -\frac{1}{4}(\alpha + 1)^2$. This time the optimizing value is $\alpha = 0$ (since any other value decreases the objective) and we get

$$x = \frac{5+0}{2} = 2.5$$

Notice that we have $\alpha = 0$, implying the constraint is not active, and indeed the inequality constraint is not at the boundary.

3 Support Vector Machines

3.1 Case-1: Separable Data

This type of SVM is also known as the hard-margin SVM. As outlined in Section 1, the objective function for maximum margin classification can be written as,

$$\begin{aligned} \max_{V,T} \quad & \frac{1}{2}\|V\|^2 \\ s.t. \quad & y_i(V \cdot x_i - T) \geq 1 \end{aligned}$$

In the standard form the constraint takes the form,

$$g_i(V,T) = 1 - y_i(V \cdot x_i - T) \leq 0$$

Here x_i represents the i^{th} sample and y_i is its label.

The Lagrangian for this problem is,

$$L((V,T),\alpha) = \frac{1}{2}V^TV + \sum_i \alpha_i(1 - y_i(V \cdot x_i - T))$$

Forming the dual objective:

$$\begin{aligned} &\frac{\delta}{\delta V}L = V - \sum_i \alpha_i y_i x_i = 0 \\ &\Leftrightarrow V = \sum_i \alpha_i y_i x_i \end{aligned} \tag{15}$$

similarly,

$$\frac{\delta}{\delta T}L = \sum_i \alpha_i y_i = 0 \tag{16}$$

From (15) and (16) we get,

$$\begin{aligned} \theta(\alpha) &= \frac{1}{2}\left\|\sum_i \alpha_i y_i x_i\right\|^2 + \sum_i \alpha_i(1 - y_i((\sum_k \alpha_k y_k x_k)x_i - T)) \\ &= \frac{1}{2}\left\|\sum_i \alpha_i y_i x_i\right\|^2 + \sum_i \alpha_i + \sum_i \alpha_i y_i T - \sum_i \alpha_i y_i (\sum_k \alpha_k y_k x_k)x_i \\ &= \frac{1}{2}\left\|\sum_i \alpha_i y_i x_i\right\|^2 + \sum_i \alpha_i + \sum_i \alpha_i y_i T - (\sum_i \alpha_i y_i x_i)(\sum_k \alpha_k y_k x_k) \end{aligned}$$

since, $\sum_i \alpha_i y_i = 0$ and $(\sum_i \alpha_i y_i x_i)(\sum_k \alpha_k y_k x_k) = \|V\|^2$, the dual objective becomes,

$$\theta(\alpha) = \sum_i \alpha_i - \frac{1}{2}\left\|\sum_i \alpha_i y_i x_i\right\|^2 \tag{17}$$

Notice that we could not substitute (16) completely into $\theta(\alpha)$ and must keep this constraint. Therefore the final form of the optimization required for SVMs is then given by,

$$\begin{aligned} \max \;\; & \theta(\alpha) \\ s.t. \;\; & \alpha_i \geq 0 \\ & \sum_i \alpha_i y_i = 0 \end{aligned}$$

3.2 Case-2: Non-separable Data

The case of non-separable data is depicted in Figure 1. The hard-margin SVM needs to be modified so that it can cater for the misclassification. This can be done by including a penalty term in the objective function, which penalizes the objective function for each mistake it makes. So the overall goal of the optimization is now to find the hyperplane with margin γ such that the number of misclassified instances is minimized. The primal objective function therefore takes the form,

$$\begin{aligned} \min \;\; & \frac{1}{2} V^T V + C \sum_i \zeta_i \\ s.t. \;\; & y_i(w \cdot x_i - T) \geq 1 - \zeta_i \\ & \zeta_i \geq 0 \end{aligned} \tag{18}$$

where the ζs represent the distance of the misclassified instance from the hyplerplane and C represents the penalty amount for each misclassification. This formulation is known as the L1 soft-margin SVM. The Lagrangian can be written as,

$$L((V, T, \zeta), \alpha, \mu) = \frac{1}{2} V^T V + C \sum_i \zeta_i + \sum_i \alpha_i((1 - \zeta_i) - y_i(V \cdot x_i - T)) + \sum_i \mu_i(-\zeta_i) \tag{19}$$

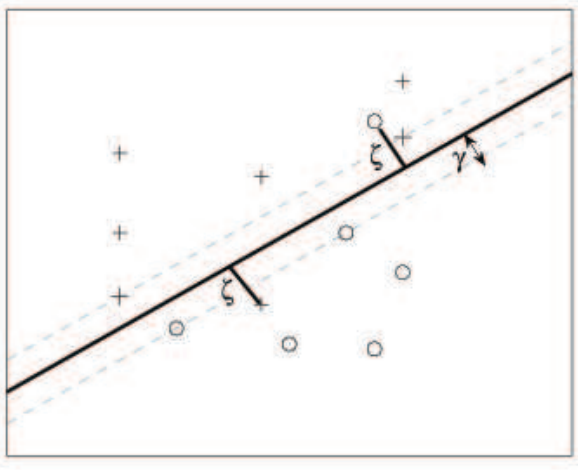

Figure 1: Non-separable data in the case of max-margin classification. The dotted line represents the margin γ and the ζ represent the distance of the misclassified points from the hyperplane.

taking derivative w.r.t the primal variables,

$$\frac{\delta}{\delta V} = V - \sum_i \alpha_i y_i x_i = 0 \quad \Rightarrow V = \sum_i \alpha_i y_i x_i$$
$$\frac{\delta}{\delta T} = \sum_i \alpha_i y_i = 0$$
$$\frac{\delta}{\delta \zeta_k} = C - \alpha_k - \mu_k = 0 \quad \Rightarrow \mu_k = C - \alpha_k$$

Therefore, although the μ_i are dual variables we can substitute them already at this stage. However, the constraint $\mu_i \geq 0$ yields a new constraint $C - \alpha_i \geq 0$ or $\alpha_i \leq C$. Substituting back in (19), the dual objective takes the form,

$$\theta(\alpha) = \sum_i \alpha_i - \frac{1}{2} \left\| \sum_i \alpha_i y_i x_i \right\|^2 \tag{20}$$

Therefore the dual problem for the L1 soft-margin SVM is,

$$\begin{aligned} \max_{\alpha} \quad & \sum_i \alpha_i - \frac{1}{2} \left\| \sum_i \alpha_i y_i x_i \right\|^2 \\ s.t. \quad & 0 \leq \alpha_i \leq C \\ & \sum_i \alpha_i y_i = 0 \end{aligned}$$

If instead of using a linear penalty term, we have a squared penalty term i.e. ζ_i is replace by ζ_i^2 the resulting SVM is known as the L2 soft-margin SVM. The simplification that we get is that we can safely remove the set of inequality constraints which ensure that the ζ_is are positive. The reason lies in the fact that if $\zeta_i \leq 0$ then setting $\zeta_i = 0$ makes the margin constraint easier to satisfy and makes the objective better. Therefore, by using this strategy the optimal solution without the $\zeta_i \geq 0$ constraints would still satisfy the non-negativity constraints.
The optimization problem in this case can be formulated as,

$$\begin{aligned} \min \quad & \frac{1}{2} V^T V + C \sum_i \zeta_i^2 \\ s.t. \quad & y_i(w \cdot x_i - T) \geq 1 - \zeta_i \end{aligned} \tag{21}$$

the Lagrangian is,

$$L((V, T, \zeta), \alpha, \mu) = \frac{1}{2} V^T V + C \sum_i \zeta_i^2 + \sum_i \alpha_i ((1 - \zeta_i) - y_i(V \cdot x_i - T)) \tag{22}$$

taking derivative w.r.t the primal variables,

$$\frac{\delta}{\delta V} = V - \sum_i \alpha_i y_i x_i = 0 \quad \Rightarrow V = \sum_i \alpha_i y_i x_i$$
$$\frac{\delta}{\delta T} = \sum_i \alpha_i y_i = 0$$
$$\frac{\delta}{\delta \zeta_k} = C\zeta_k - \alpha_k = 0 \quad \Rightarrow \zeta_k = \frac{\alpha_k}{C}$$

The dual objective is then given by,

$$\begin{aligned} \max_{\alpha} \quad & \sum_i \alpha_i + \sum_i \sum_k \alpha_i \alpha_k y_i y_k (x_i x_k + \delta_{ik} \frac{1}{2C}) \\ s.t. \quad & \alpha_i \geq 0 \\ & \sum_i \alpha_i y_i = 0 \end{aligned}$$

where δ_{ik} is the delta function.