CS 6140: Machine Learning **Spring 2015**
College of Computer and Information Science
Northeastern University
Lecture TBD **TBD, 26**
Instructor: Bilal Ahmed **Scribe: Bingyu Wang, Virgil Pavlu**

Support Vector Machines II

1 Math Background

Before moving forward, we need to mention some math background.

1.1 Lagrange multipliers

Lagrange multipliers can be used to solve the problem of the following form, whose constrain is an equality:

$$\begin{aligned} &\min_w f(w) \\ &\text{s.t.} \quad h_i(w) = 0, \quad i = 1, \ldots, l \end{aligned}$$

We define the Lagrangian as following:

$$\mathcal{L}(w, \beta) = f(w) + \sum_{i=1}^{l} \beta_i h_i(w)$$

We would find and set $\mathcal{L}$'s partial derivatives to zero:

$$\frac{\partial \mathcal{L}}{\partial w_i} = 0; \quad \frac{\partial \mathcal{L}}{\partial \beta_i} = 0$$

And then we can get the w^* to be the solution from the partial derivatives step.

1.2 Primal Problem

Consider the following, which we will call the primal optimization problems, whose has inequality as well as equality constraints.

$$\begin{aligned} &\min_w f(w) \\ &\text{s.t.} \quad g_i(w) \leq 0, \quad i = 1, \ldots, k \\ &\qquad f_i(w) = 0, \quad i = 1, \ldots, l. \end{aligned}$$

Then we can define the generalized Lagrangian

$$\mathcal{L}(w, \alpha, \beta) = f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{l} \beta_i h_i(w)$$

Here consider the quantity

$$\theta_{\mathcal{P}}(w) = \max_{\alpha, \beta : \alpha_i \geq 0} \mathcal{L}(w, \alpha, \beta)$$

In the problem $\theta_{\mathcal{P}}(w)$, if $g_i(w) > 0$ or $f_i(w) \neq 0$, which violates any of the primal constraints given above, then you should be able to verify that

$$\begin{aligned}\theta_{\mathcal{P}}(w) &= \max_{\alpha,\beta:\alpha_i \geq 0} f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{l} \beta_i h_i(w) \\ &= \infty\end{aligned}$$

Conversely, if the constraints are indeed satisfied for a particular value of w, then $\theta_{\mathcal{P}}(w) = f(w)$. Hence,

$$\theta_{\mathcal{P}}(w) = \begin{cases} f(w) & \text{if } w \text{ satisfies primal constraints} \\ \infty & \text{otherwise} \end{cases}$$

Thus, $\theta_{\mathcal{P}}$ takes the same value as the objective in our problem for all values of w that satisfies the primal constraints, and is positive infinity if the constraints are violated. Hence the minimization problem has been transformed to

$$\min_w \theta_{\mathcal{P}}(w) = \min_w \max_{\alpha,\beta:\alpha_i \geq 0} \mathcal{L}(w, \alpha, \beta)$$

For later use, we define $p^* = \min_w \theta_{\mathcal{P}}(w)$ as the value of the primal problem. In fact, we see that primal problem has the same solutions as our original problem.

1.3 Dual Problem

Then we can define

$$\theta_{\mathcal{D}}(\alpha, \beta) = \min_w \mathcal{L}(w, \alpha, \beta)$$

and then pose the dual optimization problem:

$$\max_{\alpha,\beta:\alpha \geq 0} \theta_{\mathcal{D}}(\alpha, \beta) = \max_{\alpha,\beta:\alpha \geq 0} \min_w \mathcal{L}(w, \alpha, \beta)$$

And we also define $d^* = \max_{\alpha,\beta:\alpha \geq 0} \theta_{\mathcal{D}}(\alpha, \beta)$. We can see that dual problem is pretty similar to our primal problem shown above, except that the order of the "max" and the "min" are now exchanged.

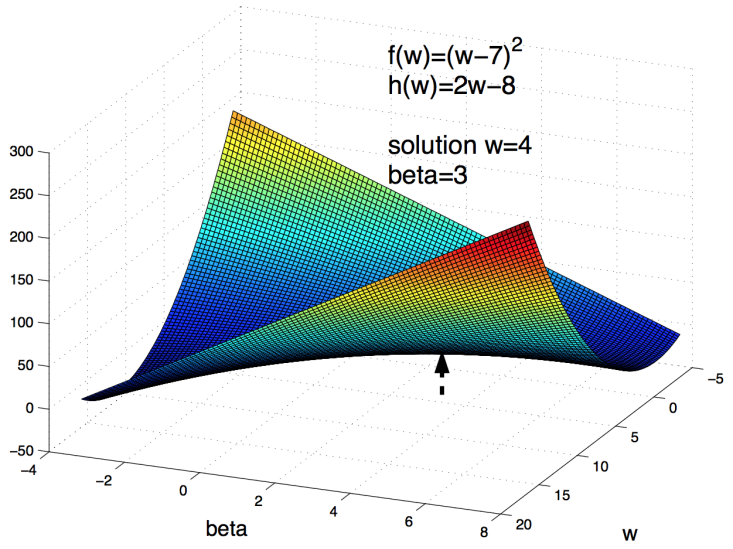

Figure 1: Saddle Point

For problem with convex objectives and linear constraints the **duality gap always closes** (KKT theorem) in the sense that

$\max_{\alpha,\beta:\alpha\geq 0} \min_w \mathcal{L}(w,\alpha,\beta) = \min_w \max_{\alpha,\beta:\alpha_i\geq 0} \mathcal{L}(w,\alpha,\beta)$

The solution is exactly this "saddle point" : maximum of the minimums of each convex slice, same as minimum of the maximums of each concave slice (shown in Figure1.

1.4 Karush-Kuhn-Tucker conditions for duality gap

How are the primal and the dual problems related? And why should we introduce primal and dual problems? We will talk a little bit in this section. Let's start with why. Since we present our original problem as following:

$$\begin{aligned} &\max_{w,b} \frac{1}{2}||w||^2 \\ &\text{s.t.} \quad y_i(w^T x_i + b) \geq 1, \quad i = 1,\ldots,m \end{aligned}$$

By introducing Lagrange multipliers α, the original constrained problem can be expressed as a primal problem:

$$\begin{aligned} w^*, b^* &= \arg p^* = \arg \min_{w,b} \theta_{\mathcal{P}}(w,b) \\ &= \arg \min_{w,b} \max_{\alpha\geq 0} (\frac{1}{2}||w||^2 - \sum_{i=1}^{m} \alpha_i (y_i(w^T x_i + b) - 1)) \end{aligned}$$

this is a saddle point[1]. If we want to solve this primal problem, we can use QP, which is inefficient. We try to transform the primal problem to the dual problem as following[2]:

$$\begin{aligned} \alpha^* &= \arg d^* = \arg \max_{\alpha} \theta_{\mathcal{D}}(\alpha) \\ &= \arg \max_{\alpha} (\sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j x_i^T x_j) \\ \text{s.t.} \quad &\alpha_i \geq 0 \\ &\sum_{i=0}^{m} \alpha_i y_i = 0 \end{aligned}$$

In dual problem, we get rid off two parameters w, b and the constrains are much easier than before. BTW, notice that we have $x_i^T x_j$ in the formula, which gives us a chance apply kernel trick on it. We will talk about it later.

We can notice that the dual problem is much better than primal problem. If we can transform the original problem to primal problem, and then to dual problem, it will be good steps to the solutions. In fact, the there is some relationship between primal and dual problems. Notice a fact that $\max\min(f) \leq \min\max(f)$, thus

$$d^* = \max_{\alpha,\beta:\alpha\geq 0} \min_w \mathcal{L}(w,\alpha,\beta) \leq \min_w \max_{\alpha,\beta:\alpha_i\geq 0} \mathcal{L}(w,\alpha,\beta) = p^*$$

That's $d^* \leq p^*$. Further, $d^* = p^*$ under the KKT conditions. Once the Primal problem and Dual problem equal to each other, the parameters will meet the KKT conditions. We just introduce the five conditions as following:

$$\frac{\partial}{\partial w_i} \mathcal{L}(w^*, \alpha^*, \beta^*) = 0, \quad i = 1,\ldots,n \tag{1}$$

$$\frac{\partial}{\partial \beta_i} \mathcal{L}(w^*, \alpha^*, \beta^*) = 0, \quad i = 1,\ldots,l \tag{2}$$

$$\alpha_i^* g_i(w^*) = 0, \quad i = 1,\ldots,k \tag{3}$$

$$g_i(w^*) \leq 0, \quad i = 1,\ldots,k \tag{4}$$

$$\alpha^* \geq 0, \quad i = 1,\ldots,k \tag{5}$$

[1] http://en.wikipedia.org/wiki/Saddle_point

[2] http://en.wikipedia.org/wiki/Support_vector_machine

Later, we will just apply KKT conditions on primal problem to get the dual form problem.

2 The Optimal Margin Classifier

Now, we get back to the SVM problem.From the intuition of margins before, we try to find a decision boundary that maximizes the geometric margin, since this would reflect a very confident set of predictions on the training set and a good "fit" to the training data. Specifically, this will result in a classifier that separates the positive and the negative training examples with a "gap" (geometric margin).

For now, we will assume that we are given a training set that is linearly separable; i.e., that it is possible to separate the positive and negative examples using some separating hyperplane. How could we find the one that achieves the maximum geometric margin? We will pose the following optimization problem: maximize the margin $\rho = 1/||w||$, such that all points are no closer (on either side) than $|w^T x+b| = 1$ to the separating plane given by $w^T x + b = 0$; thus the constraints reflect our reference choice for scale. Since the labels are the same as the 1,-1 sides of the plane, we can rewrite the constraints as $y(w^T x + b) \geq 1$ for all training points x with label $y \in \{-1, 1\}$ (will have one constraint for each training point).

To make the math nicer we write the objective in terms of $||w||^2$, and we get the following optimization problem:

SVM-PRIMAL OPTIMIZATION PROBLEM

$$\min_{w,b} \frac{1}{2}||w||^2 \tag{6}$$

$$\text{s.t.} \quad y_i(w^T x_i + b) \geq 1, \quad i = 1, \ldots, m \tag{7}$$

We've now transformed the problem into a form that can be efficiently solved. The above is an optimization problem with a **convex quadratic objective (1)** and only **linear constraints (2)**. Its solution gives us the optimal margin classifier. This optimization problem can be solved using commercial quadratic programming(QP) code[3] or (better) with duality formulation.

We will use Lagrange duality to solve the above constrained convex optimization problem. This will allow the use kernels, and it is also more efficient.

[3] http://en.wikipedia.org/wiki/Quadratic_programming