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 III

In this section, we will apply the duality mentioned above to transform the original problems to a easier dual problem and solve it. Under this framework, we will firstly talk about linearly separable case and later work on the non-separable case.

1 Linearly Separable Case

The separable case means the training data set can be separated by one line, which is shown in Figure 1. We will starts from the original problem:

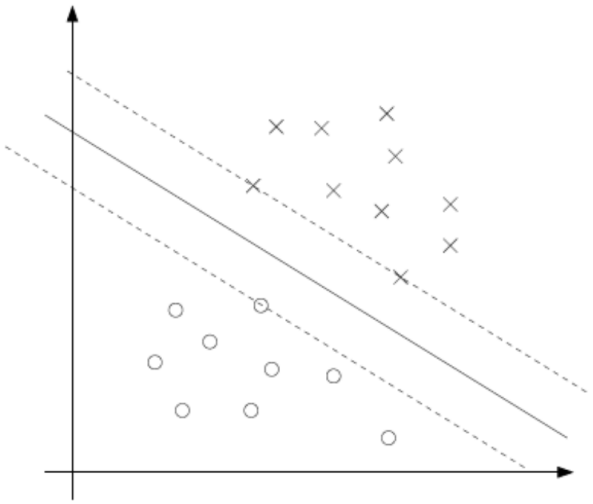

Figure 1: Separable Example

SVM-PRIMAL problem

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

We will first transform the constraints to standard form, and write down the Lagrangian including all constraints

Constraint transformed: $g_i(w, b) = -y_i(w^T x_i + b) + 1 \leq 0$

Lagrangian:

$$\mathcal{L}(w, b, \alpha) = \tfrac{1}{2}||w||^2 - \textstyle\sum_{i=1}^{m} \alpha_i (y_i(w^T x_i + b) - 1)$$

Differentiate $\mathcal{L}$ with respect to w, b , and set the differential to zero:

- For w:

$$\frac{\partial}{\partial w}\mathcal{L}(w,b,\alpha) = w - \sum_{i=1}^{m}\alpha_i y_i x_i = 0$$
$$\implies w = \sum_{i=1}^{m}\alpha_i y_i x_i \tag{1}$$

- For b:

$$\frac{\partial}{\partial b}\mathcal{L}(w,b,\alpha) = 0 - \sum_{i=1}^{m}\alpha_i y_i$$
$$\implies \sum_{i=1}^{m}\alpha_i y_i = 0 \tag{2}$$

Rewrite the Lagrangian objective. Lets put these results back into $\mathcal{L}$ equation in order to eliminate w, b:

$$\begin{aligned}
\mathcal{L}(w,b,\alpha) &= \frac{1}{2}||w||^2 - \sum_{i=1}^{m}\alpha_i(y_i(w^T x_i + b) - 1) \\
&= \frac{1}{2}w^T w - \sum_{i=1}^{m}\alpha_i y_i w^T x_i - \sum_{i=1}^{m}\alpha_i y_i b + \sum_{i=1}^{m}\alpha_i \\
&= \frac{1}{2}\sum_{i,j=1}^{m}\alpha_i\alpha_j y_i y_j x_i^T x_j - \sum_{i,j=1}^{m}\alpha_i\alpha_j y_i y_j x_i^T x_j - \sum_{i=1}^{m}\alpha_i y_i b + \sum_{i=1}^{m}\alpha_i \\
&= \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{m} y_i y_j \alpha_i \alpha_j x_i^T x_j
\end{aligned} \tag{3}$$

We have obtained the Lagrange dual problem for the original SVM-PRIMAL problem. The new variables α, one per data point represent the "force" each point pushes the pane away. The equation stated above $\sum_{i=1}^{m}\alpha_i y_i = 0$ simply states that the plane is in equilibrium as the total force on each side is the same.

It is important to understand the nature of this Lagrangian function: if the linear constraints were equality constraints, typically we'd use the constraints to solve for α-s. But in this case they are inequality constraints (standardized to ≤ 0), which means we cannot simply solve for α by differentiating on α. The KKT theorem (later section) applies to our case (convex objective, linear constraints) and governs the duality with the following rules called **KKT conditions**:

1. the solution for **minimizing** $\mathcal{L}(w,b,\alpha)$ w.r.t. w, b and subject to $\alpha \geq 0$ is the same as the solution of **maximizing** $\mathcal{L}(w,b,\alpha)$ w.r.t. α subject to appropriate constraints.

2. the Lagrangian multipliers are not negative.

3. at solution point, the differential of $\mathcal{L}(w,b,\alpha)$ w.r.t w is zero

4. for equality constraints: at solution point, the differential of $\mathcal{L}(w,b,\alpha)$ w.r.t the Lagrangian multiplier is zero, which is same as saying the constraint is satisfied (we dont have equality constraints here, but we will have them when we introduce slack variables).

5. for inequality constraints: at solution point, either the Lagrangian multiplier is zero and the constraint is satisfied loosely, or multiplier is nonzero and the constrained is satisfied with equality.

The last KKT condition is that for each point $\alpha_i(y_i(w^T x_i + b) - 1) = 0$, or that either $\alpha_i = 0$ or $y_i(w^T x_i + b) = 1$. Thus there are two kinds of training points:

- **support vectors** points for which $\alpha > 0$. These points have an active constraint $y_i(w^T x_i + b) = 1$ which contributes to the equilibrium of the plane and it is **satisfied with equality as the point is on the margin line**. If this point is erased from the training set, the plane will move (equilibrium is changed).
- **non-support vectors** points for which $\alpha = 0$. Such points have a nonactive constraint, which does not contribute to the plane, the constraint is satisfied loosely (perhaps strictly $y_i(w^T x_i + b) > 1$). If this point is erased from the training set, the plane will not move (equilibrium is in the same position).

We will name that last expression of the Lagrangian $\mathcal{L}(w, b, \alpha)$, as a function only of α-s, $W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y_i y_j \alpha_i \alpha_j x_i^T x_j$.

SVM-DUAL OPTIMIZATION PROBLEM

$$\begin{aligned} \max_{\alpha} W(\alpha) &= \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y_i y_j \alpha_i \alpha_j x_i^T x_j \\ \text{s.t.} \quad \alpha_i &\geq 0, \quad , i = 1, \ldots, m \\ \sum_{i=1}^{m} \alpha_i y^{(i)} &= 0. \end{aligned} \tag{4}$$

Recover w, b from α-s. Assuming we have solved the dual problem (next section) and we have the solution on α, let's call it α^*. Then we can calculate the solution in original parameters, call it w^*, b^* as following:

$$w^* = \sum_{i=1}^{m} \alpha^* y_i x_i$$

And as shown in Figure **??**, we can first calculate b_A and b_B, then get b^*:

$b_A = \max_{i:y_i=-1} w^{*T} x_i$ This is the maximum on negative points that b^* has to compensate to -1: $b^* \leq -1 - b_A$

$b_B = \min_{i:y_i=1} w^{*T} x_i$ This is the minimum on positive points that b^* has to compensate to 1: $b^* \geq 1 - b_B$

So $1 - b_B \leq b^* \leq -1 - b_A$. We will take b^* to be the average of these two values:

$b^* = \frac{1-b_B-1-b_A}{2} = -\frac{b_A+b_B}{2}$

2 Linearly Non-separable Case

The derivation of the SVM as presented so far assumed that the data is linearly separable. In some cases, it is not clear that finding a separating hyperplane is exactly what we'd want to do, since that might be susceptible to outliers. For instance, the Figure 2, it causes the decision boundary to make a dramatic swing, and the resulting classifier has a much smaller margin.

Notice that in the dual form of the SVM problem, the objective function is a function of dot product of x_i and x_j, namely, $x_i^T x_j$.

One way to deal with the non-linearly separable data sets is to use **kernels**, which maps the dot product of x_i and $x_j$into a higher dimension space. We will explain the details in the next module.

Alternatively, in order to make the algorithm work for non-linearly separable data sets as well as be less sensitive to outliers, we reformulate the optimization(using $\mathcal{L}_1 \quad regularization$) as following:

$$\begin{aligned} \min_{w,b,\xi} \frac{1}{2}||w||^2 + C \sum_{i=1}^{m} \xi_i \\ \text{s.t.} \quad y_i(w^T x_i + b) &\geq 1 - \xi_i, \quad i = 1, \ldots, m \\ \xi_i &\geq 0, \quad i = 1, \ldots, m \end{aligned}$$

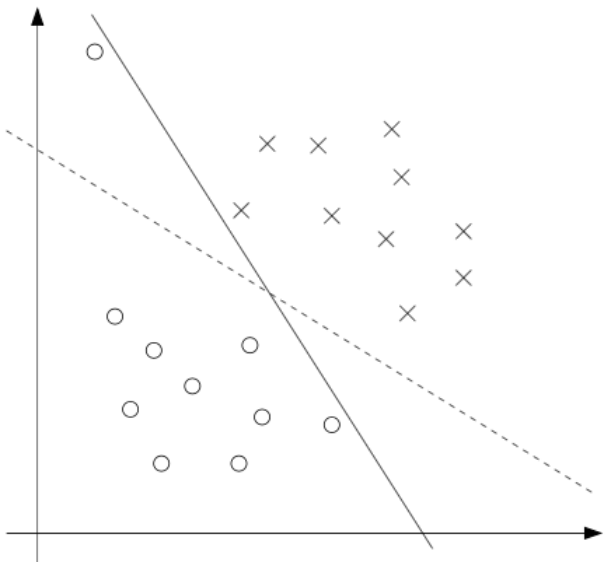

Figure 2: Outlier Example

Thus, examples are now permitted to have margin less than 1, and if an example has functional margin $1-\xi_i$ (with $\xi > 0$), we would pay a cost of the objective function being increased by $C\xi_i$. The parameter C controls the relative weighting between the twin goals of making the $||w||^2$ small and of ensuring that most examples have functional margin at least 1.

SVM-DUAL FORM with SLACK VARIABLES

$$\max_{\alpha} W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y_i y_j \alpha_i \alpha_j x_i^T x_j \tag{5}$$
$$\text{s.t.} \quad 0 \leq \alpha_i \leq C, \quad , i = 1, \ldots, m$$
$$\sum_{i=1}^{m} \alpha_i y(i) = 0.$$

In adding $\mathcal{L}_1$ *regularization*, the only change to the dual problem is that was originally a constraint that $0 \leq \alpha_i$ has now become $0 \leq \alpha_i \leq C$. The calculation for w^* is the same way, but the calculation for b^* has to be modified (b^* calculation discussed as part of SMO solver). In this case there are three types of training points:

- $\alpha = 0$: non interesting points
- $C > \alpha > 0; \beta = 0$: a support vector on the margin line, no slack variable; $y_i(w^T x_i + b) = 1$, $\xi_i = 0$
- $\alpha = C; \beta > 0$: a support vector, inside the side (or even misclassified): $\xi_i > 0$; $y_i(w^T x_i + b) < 1$, $\xi_i > 0$

2.1 Slack variables dual form derivation

Let's derive this non-separable problem like we did before. We will have additional constrains for slack variables $\xi_i \geq 0$

1. Non-separable problem

$$\begin{aligned} \min_{w,b,\xi} \frac{1}{2}||w||^2 + C\sum_{i=1}^{m}\xi_i \\ \text{s.t.} \quad y_i(w^T x_i + b) \geq 1 - \xi_i, \quad i = 1, \ldots, m \\ \xi_i \geq 0, \quad i = 1, \ldots, m \end{aligned}$$

2. Constraint transformed:

$$\begin{aligned} g_i(w,b) = 1 - \xi_i - y_i(w^T x_i + b) \leq 0 \\ h_i(w,b) = -\xi_i \leq 0 \end{aligned}$$

3. Lagrangian:

$$\mathcal{L}(w,b,\xi,\alpha,r) = \frac{1}{2}||w||^2 + C\sum_{i=1}^{m}\xi_i - \sum_{i=1}^{m}\alpha_i(y_i(w^T x_i + b) + \xi_i - 1) - \sum_{i=1}^{m} r_i\xi_i$$

4. Set $\theta_{\mathcal{D}}(\alpha, r) = \min_{w,b} \mathcal{L}(w,b,\xi,\alpha,r)$
   Differentiate $\mathcal{L}$ with respect to w, b, ξ to zero:

   - For w:

$$\begin{aligned} \frac{\partial}{\partial w}\mathcal{L}(w,b,\xi,\alpha,r) = w - \sum_{i=1}^{m}\alpha_i y_i x_i = 0 \\ \implies w = \sum_{i=1}^{m}\alpha_i y_i x_i \end{aligned} \tag{6}$$

   - For b:

$$\begin{aligned} \frac{\partial}{\partial b}\mathcal{L}(w,b,\xi,\alpha,r) = 0 - \sum_{i=1}^{m}\alpha_i y_i = 0 \\ \implies \sum_{i=1}^{m}\alpha_i y_i = 0 \end{aligned} \tag{7}$$

   - For ξ:

$$\begin{aligned} \frac{\partial}{\partial \xi_i}\mathcal{L}(w,b,\xi,\alpha,r) = C - \alpha_i - r_i = 0 \\ \implies C = \alpha_i + r_i \;\; \forall i \in \{1, \cdots, m\} \end{aligned} \tag{8}$$

5. Put the last three equalities back into $\mathcal{L}$, allows for an objective like before only in Lagrangian variables

α:

$$\begin{aligned}
\mathcal{L}(w,b,\alpha) &= \frac{1}{2}||w||^2 + C\sum_{i=1}^{m}\xi_i - \sum_{i=1}^{m}\alpha_i(y_i(w^T x_i + b) + \xi_i - 1) - \sum_{i=1}^{m} r_i\xi_i \\
&= \frac{1}{2}||w||^2 + \sum_{i=1}^{m}\alpha_i\xi_i + \sum_{i=1}^{m} r_i\xi_i - \sum_{i=1}^{m}\alpha_i(y_i(w^T x_i + b) - 1) - \sum_{i=1}^{m}\alpha_i\xi_i - \sum_{i=1}^{m} r_i\xi_i \\
&= \frac{1}{2}||w||^2 - \sum_{i=1}^{m}\alpha_i(y_i(w^T x_i + b) - 1) \\
&= \frac{1}{2}\sum_{i,j=1}^{m}\alpha_i\alpha_j y_i y_j x_i^T x_j - \sum_{i,j=1}^{m}\alpha_i\alpha_j y_i y_j x_i^T x_j - \sum_{i=1}^{m}\alpha_i y_i b + \sum_{i=1}^{m}\alpha_i \\
&= \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{m} y_i y_j \alpha_i\alpha_j x_i^T x_j \qquad (9)
\end{aligned}$$

Now we get (15), which is the same with (10) in previous derivatives. Although we added more parameters, we only have α now.