CS260: Machine Learning Theory
Lecture 15: Convex Optimization and Maximizing the Margin

November 14, 2011

Lecturer: Jennifer Wortman Vaughan

1 Road Map

During our discussion of AdaBoost, we saw that large margin classifiers lead to low generalization error. Given this observation, one might ask if it is possible to design a learning algorithm that explicitly maximizes the margin of the training data. This is the idea behind Support Vector Machines (SVMs), which maximize the margin directly using techniques from convex optimization.

We begin this lecture with a review of the necessary ideas from optimization, and then go on to define the primal form of the SVM algorithm. In the next lecture, we will derive the dual form of the algorithm, which turns out to be useful. We will also see how to extend the algorithm to handle data that cannot be classified perfectly with a linear separator.

For a much more thorough introduction to convex optimization, check out Boyd and Vandenberghe's great book, which is available for free online,[1] or take Vandenberghe's class here at UCLA.

2 Basics of Convex Optimization

We say that an optimization problem is in standard form if it is possible to write it as

$$\begin{aligned} \min_{\vec{w}} \quad & f(\vec{w}) \\ \text{s.t.} \quad & g_i(\vec{w}) \le 0, \quad i = 1, ..., k \\ & h_i(\vec{w}) = 0, \quad i = 1, ..., \ell \end{aligned}$$

for some *objective function* f, *inequality constraints* $g_i(\vec{w}) \le 0, i = 1, ..., k$, and *equality constraints* $h_i(\vec{w}) = 0, i = 1, ..., \ell$.

We say that a standard form problem is *convex* if f is convex, each of the functions g_i is convex, and each of the functions h_i is affine, i.e., can be written as $h_i(\vec{w}) = \vec{z} \cdot \vec{w} + b$ for some $\vec{z}$ and b.[2]

Convex standard form optimization problems are nice because they can be solved efficiently using a variety of different techniques including the subgradient method, the interior point method, and the ellipsoid method (which we unfortunately will not have time to discuss in this class). We will refer to a problem of this form as our *primal* problem.

All CS260 lecture notes build on the scribes' notes written by UCLA students in the Fall 2010 offering of this course. Although they have been carefully reviewed, it is entirely possible that some of them contain errors. If you spot an error, please email Jenn.

[1] `http://www.stanford.edu/~boyd/cvxbook/bv_cvxbook.pdf`

[2] This is different from the definition given in class, where we said that "standard form" implies that f and g_i are convex and h_i affine, but as some people pointed out, it is more common in references to see "standard form" defined as it is here now.

The *Lagrangian* of an optimization problem is defined as

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

and the components of $\vec{\alpha}$ and $\vec{\beta}$ are referred to as Lagrange multipliers. Let's think about the value

$$\max_{\vec{\alpha}, \vec{\beta} : \alpha_i \geq 0} \; L(\vec{w}, \vec{\alpha}, \vec{\beta}) = \max_{\vec{\alpha}, \vec{\beta} : \alpha_i \geq 0} \left(f(\vec{w}) + \sum_{i=1}^{k} \alpha_i g_i(\vec{w}) + \sum_{i=1}^{l} \beta_i h_i(\vec{w}) \right) .$$

First consider the case in which the primal constraints are met. In this case, the Lagrangian is maximized when we set each α_i to 0. Each β_i can be set arbitrarily since $h_i(\vec{w}) = 0$. In this case, the second and third terms of the Lagrangian are 0, and we're left with just $f(\vec{w})$. What if the primal constraints are not met. In this case, it is easy to verify that maximal value is infinite. We therefore have that

$$\max_{\vec{\alpha}, \vec{\beta} : \alpha_i \geq 0} \; L(\vec{w}, \vec{\alpha}, \vec{\beta}) = \begin{cases} f(\vec{w}) & : \text{ if the primal constraints are satisfied} \\ \infty & : \text{ otherwise.} \end{cases}$$

Because of this, the primal optimization problem is *equivalent to* the following:

$$\min_{\vec{w}} \; \max_{\vec{\alpha}, \vec{\beta} : \alpha_i \geq 0} \; L(\vec{w}, \vec{\alpha}, \vec{\beta}) .$$

We can define what is called the *dual* optimization problem, by switching the $\min$ and $\max$:

$$\max_{\vec{\alpha}, \vec{\beta} : \alpha_i \geq 0} \; \min_{\vec{w}} \; L(\vec{w}, \vec{\alpha}, \vec{\beta}) .$$

Let's see how these problems are related. It is easy to show that

$$\max_{\vec{\alpha}, \vec{\beta} : \alpha_i \geq 0} \; \min_{\vec{w}} \; L(\vec{w}, \vec{\alpha}, \vec{\beta}) \leq \min_{\vec{w}} \; \max_{\vec{\alpha}, \vec{\beta} : \alpha_i \geq 0} \; L(\vec{w}, \vec{\alpha}, \vec{\beta}) .$$

This has nothing to do with the form of the Lagrangian; the "max min" is always less than or equal to the "min max," a fact that was used in the proof of the Minimax Theorem.

What is more surprising is that for standard form problems that are convex (i.e., have convex f and g_i and affine h_i) if it is the case that for all i, there exists some $\vec{w}$ such that $g_i(\vec{w}) < 0$, then these two quantities are actually equal.

2.1 KKT Conditions

Suppose we have an optimization problem in standard form that is convex, and that for all i, there exists some $\vec{w}$ such that $g_i(\vec{w}) < 0$. We know in this case that optimal value of the primal problem is equal to the optimal value of the dual problem. It turns out that particular values of $\vec{w}^*$, $\vec{\alpha}^*$ and $\vec{\beta}^*$ are solutions to these optimization problems if and only if they satisfy the following conditions.

The KKT Conditions:

- Stationarity:

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

- Primal Feasibility:

$$\begin{aligned} h_i(\vec{w}^*) &= 0 \quad i = 1, ..., l \\ g_i(\vec{w}^*) &\leq 0 \quad i = 1, ..., k \end{aligned}$$

- Dual Feasibility:

$$\vec{\alpha}^* \geq 0 \quad i = 1, ..., k$$

- Complementary Slackness:

$$\alpha_i^* g_i(\vec{w}^*) = 0 \quad i = 1, ..., k$$

This final condition will be important later when we talk about the dual of the SVM problem.

3 Maximizing the Margin

Consider a linear threshold function of the form $y = \text{sign}(\vec{w} \cdot \vec{x} + b)$, where $\vec{w}$ is the weight vector and b is the intercept (signifying that the threshold function may not pass through the origin). The (L2) margin of a labeled point $(\vec{x}, y)$ with respect to this threshold is given by

$$y \frac{\vec{w} \cdot x + b}{||\vec{w}||} .$$

Intuitively, the larger the margin of a point is, the higher our "confidence" that the point belongs to a certain class should be.

Assume we have m separable points $(\vec{x}_1, y_1), (\vec{x}_2, y_2), \ldots (\vec{x}_m, y_m)$, i.e., m labeled points for which there exists a consistent linear separator. Given these training samples, we would like to find a classifier that maximizes the value of the margin. We can attempt to formulate this problem as a standard form convex optimization problem as follows:

$$\begin{aligned} \max_{\gamma, \vec{w}, b} \quad & \gamma \\ \text{s.t.} \quad & y_i \frac{\vec{w} \cdot \vec{x_i} + b}{||\vec{w}||} \geq \gamma, \quad i = 1, ..., m \end{aligned}$$

We have a $\max$ instead of a $\min$ here, but that problem alone would be easy to fix. The bigger problem is the presence of the term $1/||\vec{w}||$ in the inequality constraints. This term is not convex, meaning that we could not use convex standard form solvers for this problem. So, let's attempt to reformulate this optimization problem to get rid of this annoying constraint.

We can get this term out of the inequality constraints by defining a new variable $\gamma' = \gamma||\vec{w}||$, and rewriting the optimization in terms of this new variable. We have

$$\begin{aligned} \max_{\gamma',\vec{w},b} \quad & \frac{\gamma'}{||\vec{w}||} \\ \text{s.t.} \quad & y_i(\vec{w}\cdot\vec{x_i}+b) \geq \gamma', \quad i = 1,...,m \end{aligned}$$

The inequality constraints are looking nicer, but we still have this annoying term in our objective. However, it can be removed. As a first step, suppose that $\gamma' = \gamma^*, \vec{w} = \vec{w}^*$, and $b = b^*$ is a solution to this problem. Then $\gamma' = 1, \vec{w} = \vec{w}^*/\gamma^*$ and $b = b^*/\gamma^*$ must *also* be a solution. (Verify this for yourself.) Using this, we can reformulate our optimization problem, removing γ' entirely, without missing any potential solutions. We get:

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

This will make our solution unique, which is useful.

To get this in the form we need, we rewrite it one last time as an equivalent problem:

$$\begin{aligned} \min_{\vec{w},b} \quad & \frac{1}{2}\cdot||\vec{w}||^2 \\ \text{s.t.} \quad & 1 - y_i(\vec{w}\cdot\vec{x_i}+b) \leq 0, \quad i = 1,...,m \end{aligned}$$

We now have the optimization problem in the form we want! Note that we could have minimized $||\vec{w}||$ instead, but it turns out this form will be more convenient. Furthermore, since the objective function is quadratic and we have linear inequality constraints, this convex optimization problem could be solved using any standard quadratic programming technique.

If we wanted, we could stop at this point and call ourselves done. However, it turns out that the dual of this problem satisfies a number of nice properties (in particular, it is easy to solve, and will allow us to use the "kernel trick" which will be described in the next lecture). We now derive the dual.

4 Deriving the Dual

As we saw above, the dual problem is defined as

$$\max_{\vec{\alpha}:\alpha_i\geq 0} \min_{\vec{w},b} L(\vec{w}, b, \vec{\alpha}).$$

Note that we dropped the dependence on $\vec{\beta}$ since we only have inequality constraints and no equality constraints.

We will derive a more convenient form in three steps. The first step is to calculate the value of the Lagrangian function by plugging in our objective function and constraints. This gives us

$$\begin{aligned} L(\vec{w}, b, \vec{\alpha}) &= f(\vec{w}, b) + \sum_{i=1}^{m} \alpha_i g_i(\vec{w}, b) \\ &= \frac{1}{2}||\vec{w}||^2 + \sum_{i=1}^{m} \alpha_i(1 - y_i(\vec{w}\cdot\vec{x}_i + b)) \ . \end{aligned} \tag{1}$$

The next step is to find the minimum value of the Lagrangian for any fixed value of $\vec{\alpha}$. Because the Lagrangian is convex, the minimum occurs when the partial derivatives are equal to 0. Setting the partial derivatives with respect to the components of $\vec{w}$ to 0 gives us

$$\vec{w} = \sum_{i=1}^{m} \alpha_i y_i \vec{x}_i .$$

Setting the partial with respect to b to 0 gives us a constraint that

$$\sum_{i=1}^{m} \alpha_i y_i = 0 \; .$$

By plugging these values into Equation 1, we get

$$\min_{\vec{w},b} L(\vec{w}, b, \vec{\alpha}) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} y_i y_j \alpha_i \alpha_j \vec{x}_i \cdot \vec{x}_j .$$

By maximizing this expression with respect to $\vec{\alpha}$ and applying the constraint we derived from the partial derivatives, we obtain the dual form of the problem.

$$\begin{aligned} \max_{\vec{\alpha}} \quad & \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} y_i y_j \alpha_i \alpha_j \vec{x}_i \cdot \vec{x}_j \\ \text{s.t.} \quad & \alpha_i \geq 0 \\ & \sum_{i=1}^{m} \alpha_i y_i = 0 \end{aligned}$$

This version of the problem is easy to solve. Additionally, it only depends on the $\vec{x}_i$ through dot products of the form $\vec{x}_i \cdot \vec{x}_j$, so it can be used with the kernel trick, which we will talk about next time.