CS281B/Stat241B: Advanced Topics in Learning & Decision Making

Maximal Margin Classifier

Lecturer: Michael Jordan *Scribes: Jana van Greunen*

Corrected version - 2/12/2004

1 References/Recommended Reading

1.1 Websites

- www.kernel-machines.org

1.2 Books

- Learning with Kernels - MIT press
- Shawe-Taylor and Cristianini (2nd edition)

2 Separating Binary Data with a Hyper-plane

2.1 Observations from Lecture 1

From the discussion of the four classifiers in lecture 1, it can be seen that all methods use the inner product between the data vectors $x_i \cdot x_j \ \ \forall i, j$, instead of the data vectors themselves. This fact is obvious for the Perceptron classifier. In lecture 1 we saw that $\Delta\theta$ at each step is a weighted sum of $x_i's$. The predictions are determined by taking the inner product of $\theta \cdot x_i$. This inner product expands into an inner product between the x_i vectors. The use of the inner product instead of the data vectors themselves will yield great computational savings and allow kernelization (see lecture 3 for more details).

2.2 Maximal Margin cont'

In this lecture we will consider problems for which the input data does not overlap, in other words, there exists a hyper-plane which separates the data into the two sets corresponding to $y_i = 1$ and $y_i = -1$.

Note: In this problem it will be very useful to consider the Lagrangian dual problem. The Lagrangian dual turns out to be an unconstrained quadratic problem that is simpler to solve. The Lagrangian approach also gives insight about the structure of the solution (i.e. the solution is also a weighted linear combination of the data elements x_i.)

First, consider a vector w that is perpendicular to the separating hyper-plane. Define: $\theta = (w, b)^T$ where b is a scalar value.

The primal problem can be formulated in the following way:

$$minimize\ w^T w$$

$$subject\ to\ y_i(w^t x_i + b)\ \geq 1$$

The optimization problem is the minimization of the norm of w. It is a quadratic problem because the constraints are linear while the objective function is quadratic. The fact that minimizing the norm of w solves the problem may be surprising at first but can be shown with the following algebra. Pick z_1 and z_{-1} so that they are the vectors lying on the boundaries defining the margins (see Figure 1 for a depiction of z_1 and z_{-1}). (Note: for the maximal margin problem, all data points that do not lie on either z_1 or z_{-1} can be discarded because they do not contribute to the margins.) It is easy to see that the separating hyperplane needs to lie at mid-distance between z_1 and z_{-1} (otherwise you could move the plane towards the midpoint to increase the minimal margin). And so the minimal margin will simply be half the perpendicular distance between z_1 and z_{-1}, that is half the vector between z_1 and z_{-1} projected on the unit normal to the plane:

$$\text{margin} = \frac{1}{2}\frac{w^T}{||w||}(z_1 - z_{-1}) \tag{1}$$

To find the value of this expression, we do the following manipulations:

$$w^T z_1 + b = 1$$

$$w^T z_{-1} + b = -1$$

Now:

$$(w^T z_1 + b) - (w^T z_{-1} + b) = 2$$

And:

$$w^T (z_1 - z_{-1}) = 2$$

and so comparing with (1), we get the expression for the margin:

$$\text{margin} = \frac{1}{||w||}$$

Thus, minimizing $||w||$ is equivalent to maximizing the minimal margin $1/||w||$.

2.3 Using a Lagrangian

A general convex optimization has the following form:

minimize $f(x)$

s.t. $g(x) \leq 0$

Now, define a Lagrangian:

$$\mathcal{L}(x, \lambda) = f(x) + \lambda^T g(x)$$

Claim: the original problem can be written as follows:

$$\min_x \max_\lambda \mathcal{L}(x, \lambda)\ s.t.\ \lambda \geq 0$$

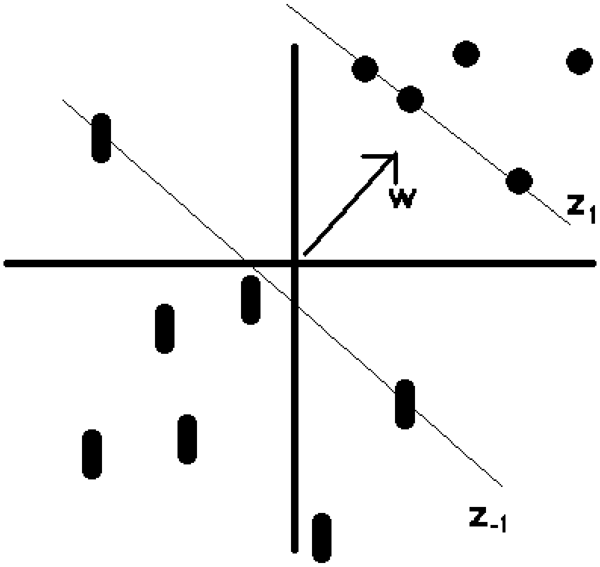

Figure 1: Plot of Data Showing Perpendicular vector w and vectors z_1 and z_{-1}.

This can be seen by taking the inner term:

$$\max_{\lambda} \mathcal{L}(x,\lambda) = \left\{ \begin{array}{ll} f(x) & g(x) \leq 0 \\ \infty & g(x) \geq 0 \end{array} \right\}$$

When we are in the non-feasible region the solutions is ∞ which will never be picked by the outer loop minimization. When we are in the feasible region, the objective function $f(x)$ is minimized.

2.4 Leap of Intuition (Dual Maximal Margin)

Instead of writing the problem in the following way (primal):

$$\min_{x} \max_{\lambda} \mathcal{L}(x,\lambda) \ s.t. \ \lambda \geq 0$$

we swap the two operators and solve

$$\max_{\lambda} \min_{x} \mathcal{L}(x,\lambda).$$

In general, the two problems do not have the same solution. However, if certain *constraint qualifications* hold, the solutions are the same. One example of a constraint qualification is *Slater's* qualification, which means that the problem must be strictly feasible.

Now, why does swapping make sense?

Consider a plot of the image of the domain of x under the mapping $x \to (g(x), f(x))$ (see figure 2). The optimal primal solution lies on the ordinate, on the lower boundary of the image of this mappping.

In the dual problem, the Lagrangian $f(x) + \lambda g(x)$ is being minimized. On the graph this is the y-intercept of the line with slope $-\lambda$ passing through the point $(g(x), f(x))$. The minimization finds the smallest such y-intercept, ranging over all x. This corresponds to the dual function. The subsequent maximization of the dual function takes the maximum of such y-intercepts. This yields the same point as the primal solution.

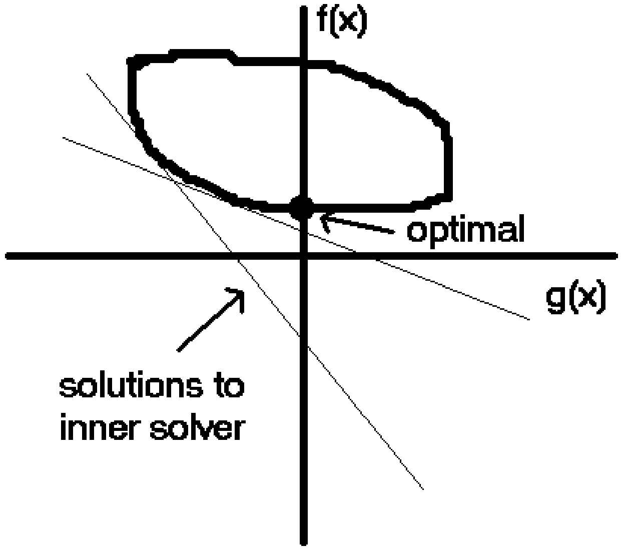

Figure 2: Dual and Primal solutions for Convex Data.

In general, the problem can be solved in the following way: start with a λ the solve to get a lower bound, adjust λ and solve again. Or the problem can be solved by choosing an x as a starting point for the primal problem and a λ as a starting point for the dual-problem and then closing the gap between the solutions. These are called primal-dual algorithms.

When the set is not convex there exists a duality gap. This is demonstrated in figure 3.

2.5 Solving for w and b

We can rewrite the Lagrangian as follows:

$$\mathcal{L}(w, b, \alpha) = \frac{1}{2} w^T w + \sum_{1}^{n} \alpha_i (1 - y_i (w^T x_i + b)$$

We first take the derivative with respect to w and set to zero:

$$\frac{\delta \mathcal{L}}{\delta w} = 0$$

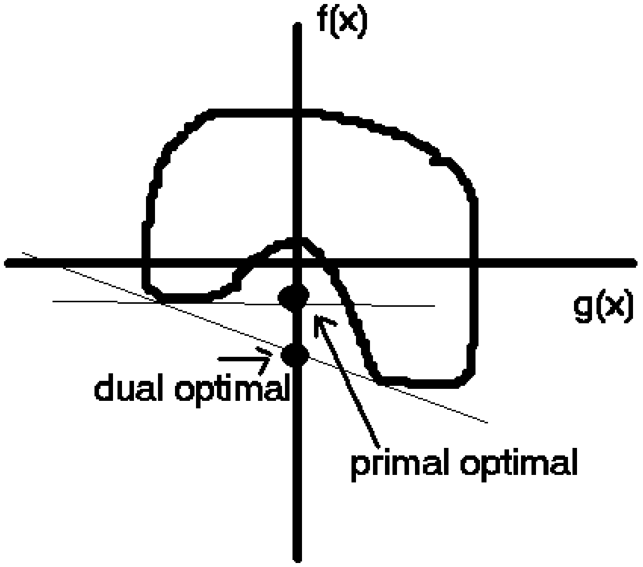

Figure 3: Duality Gap for Non-Convex Data.

$$=> w = \sum_i \alpha_i y_i x_i.$$

Substituting w back into the Lagrangian we get:

$$\sum_i \alpha_i - \sum_i \alpha_i y_i b - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j x_i^T x_j.$$

Thus, for α such that $\sum_i \alpha_i y_i = 0$ we get:

$$\theta(\alpha) = \sum_i \alpha_i - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j x_i^T x_j.$$

On the other hand, for α such that $\sum_i \alpha_i y_i \neq 0$, we get $\theta(\alpha) = -\infty$ (by taking b to infinity).

When maximizing the dual function, it is clear that points α such that $\sum_i \alpha_i y_i \neq 0$ cannot be maxima. Thus we have uncovered an implicit constraint in the problem, namely that $\sum_i \alpha_i y_i = 0$. Thus the dual problem reduces to maximizing:

$$\theta(\alpha) = \sum_i \alpha_i - \sum_i \alpha_i y_i b - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j x_i^T x_j$$

subject to $\alpha \geq 0$ and $\sum_i \alpha_i y_i = 0$.