

Lecture 9 — February 6

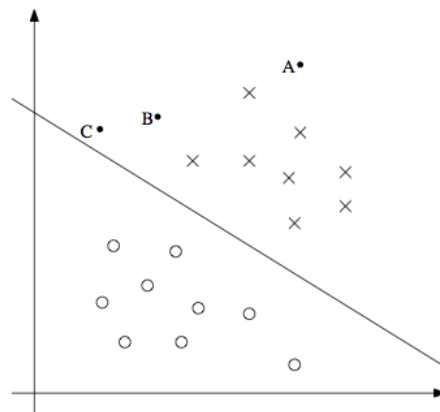
Scribe: Sidharth Shekhar (Student)

Lecturer: Deva Ramanan

9.1 Support Vector Machines

9.1.1 Motivation

Recall Linear Classification



Let us assume we have m data points $(\{x^{(i)}, y^{(i)}\}, \dots, \{x^{(m)}, y^{(m)}\})$ that are linearly separable as shown in the figure. It is easier to classify point A as belonging to class 1 since it is farther away from the boundary than points B and C.

Let $\gamma^{(i)}$ be the distance from data point $x^{(i)}$ to the boundary.

Let γ be defined as $\min_{i=1, \dots, m} \gamma^{(i)}$

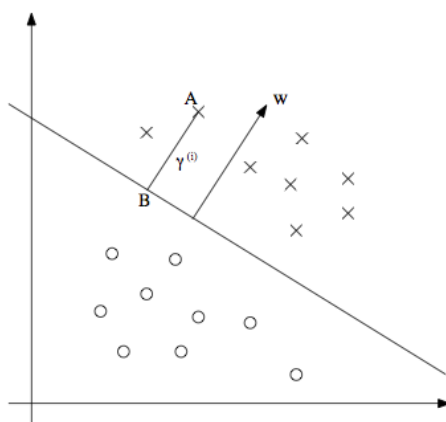
If the data is linearly separable, consider the set of all decision boundaries $w^T x + b$ that separate the data. We would like to select from this set the boundary that maximizes γ , which we will call the margin. Intuitively, one can imagine that such an optimization only depends on points near the boundary. Imagine constructing the convex hull of the positive points, and the convex hull of the negative points. Moving, or even removing, all the points within the hull will not change the solution because they do not effect γ . So we expect that only a subset of the data points near the boundary will effect the final solution. This *sparse-ness* property suggests that one might be able to efficiently optimize w and b by focusing the computational effort on the boundary points.

Some Notation

$$y^i \in \{-1, 1\} \quad (9.1)$$

$$h(x) = w^T x + b \quad (9.2)$$

9.1.2 Planar Geometry



The vector w is orthogonal to the separating boundary. The point A represents an input $x^{(i)}$ of a training example with class label $y^{(i)} = 1$. Its distance to the boundary, $\gamma^{(i)}$, is given by the line segment AB . To calculate this value of $\gamma^{(i)}$, we first assume $\|w\| = 1$. Let $x^{(i)}$ be some arbitrary point. If it lies on the line defined by w and b ,

$$(x^{(i)} - r_0)^T w = 0, \text{ where } r_0 \text{ is any reference point on the line} \quad (9.3)$$

$$w^T x^{(i)} - r_0 w = 0 \quad (9.4)$$

$$w^T x^{(i)} + b = 0 \quad (9.5)$$

We have used the fact that the dot product of a vector v with a unit vector w is the projection of v onto w . Note that $r_0 w$ is the perpendicular distance of the line to the origin. Hence, when w is unit-length, one can interpret $-b$ as the distance of the line to the origin.

Following similar arguments, if $x^{(i)}$ does not lie on the boundary line, we have

$$\gamma^{(i)} = w^T x^{(i)} + b \quad (9.6)$$

The distance $\gamma^{(i)}$, as currently defined, can be positive or negative based on which side of the line the point $x^{(i)}$ lies. To make all distances positive we redefine $\gamma^{(i)}$ as follows.

$$\gamma^{(i)} = w^T x^{(i)} + b, \text{ for positive examples} \quad (9.7)$$

$$\gamma^{(i)} = -(w^T x^{(i)} + b), \text{ for negative examples} \quad (9.8)$$

Since $y^{(i)} \in \{1, -1\}$,

$$\gamma^{(i)} = y^{(i)}(w^T x^{(i)} + b) \quad (9.9)$$

Now let us consider the case when $\|w\| \neq 1$. The equation $w^T x^{(i)} + b = 0$ can be scaled arbitrarily. So we scale by the value $\frac{1}{\|w\|}$. Therefore,

$$\gamma^{(i)} = y^{(i)} \left(\frac{w^T}{\|w\|} x^{(i)} + \frac{b}{\|w\|} \right) \quad (9.10)$$

where $\|w\| = \sqrt{\sum_i w_i^2} = \sqrt{w^T w}$

We can now define the problem as

$$\begin{aligned} & \text{maximize} && \gamma \\ & \text{s.t.} && y^{(i)} \left(\frac{w^T}{\|w\|} x^{(i)} + \frac{b}{\|w\|} \right) \geq \gamma, \quad i = 1, \dots, m \end{aligned}$$

Let $\hat{\gamma} = \gamma \cdot \|w\|$

We can always scale (w, b) such that $\hat{\gamma} = 1$. Also $\max \frac{1}{\|w\|}$ is equivalent to $\min \|w\|^2$. Therefore our problem can now be stated as

$$\begin{aligned} & \min && \frac{1}{2} \|w\|^2 \\ & \text{s.t.} && y^{(i)} (w^T x^{(i)} + b) \geq 1, \quad i = 1, 2, \dots, m \end{aligned} \quad (9.11)$$

The above problem is a quadratic program (QP), and can be solved with an off-the-shelf QP solver. We are optimizing a quadratic function subject to linear constraints. In principle, the exact solution can be found. Hence, in some sense, we are done with the problem definition of an SVM - at least for the linearly separable case. However, it will be useful to analyze the methods used for optimizing such QPs. In particular, it is computationally more convenient to solve the dual problem because one can exploit our earlier *sparseness* intuition.

9.1.3 Lagrangian Duality

The problem which we have to solve is a constrained optimization problem. It is of the form

$$\begin{aligned} & \min && f(w) \\ & \text{s.t.} && g(w) \leq 0 \end{aligned}$$

By convention, we write that $g(w) \leq 0$, so this means we multiply the constraints from (9.11) with a negative one. Notice also that in (9.11) we have a constraint for *each* data point. The general problem can be defined as

$$\min f(w) \quad (9.12)$$

$$\text{s.t. } g_i(w) \leq 0, \quad i = 1, \dots, k \quad (9.13)$$

$$h_j(w) = 0, \quad j = 1, \dots, l \quad (9.14)$$

$$(9.15)$$

To solve this we use the method of Lagrange multipliers. We define the Lagrangian to be the original objective function added to a weighted combination of the constraints. The weights are called lagrange multipliers. It will be helpful to focus on the simpler case with one inequality constraint and one lagrange multiplier.

$$L(w, \alpha) = f(w) + \alpha g(w) \quad (9.16)$$

Theorem 9.1. *The original minimization problem can be written as*

$$\min_w \max_{\alpha \geq 0} L(w, \alpha) \quad (9.17)$$

Proof: Looking at the inner term we get

$$\max_{\alpha \geq 0} L(w, \alpha) = \begin{cases} f(w) & g(w) \leq 0 \\ \infty & g(w) > 0 \end{cases} \quad (9.18)$$

This is because when $g(w) \leq 0$, we maximize (9.16) by setting $\alpha = 0$. When $g(w) > 0$, one can drive the value to infinity by setting α to a large number. Minimizing the outer loop, one sees that we obtain the minimum value of $f(w)$ such that the constraint $g(w) \leq 0$ holds. Thus we can say that the two problems are equivalent. \square

9.1.4 Dual Optimization

The primal solution to the problem is given by

$$p^* = \min_w \max_{\alpha \geq 0} L(w, \alpha) \quad (9.19)$$

The dual solution to the problem is given by

$$d^* = \max_{\alpha \geq 0} \min_w L(w, \alpha) \quad (9.20)$$

We claim that $d^* \leq p^*$. Let w^* be the w value that corresponds to the optimal primal solution p^* . We can write, for all $\alpha \geq 0$

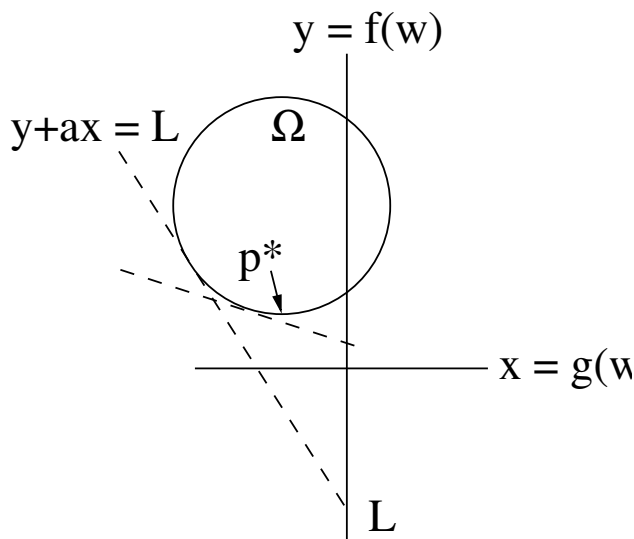
$$\max_{\hat{\alpha} \geq 0} L(w^*, \hat{\alpha}) \geq L(w^*, \alpha) \geq \min_w L(w, \alpha) \quad (9.21)$$

The left-hand-side (LHS) of the above is p^* . This means we can interpret the RHS as a lower bound on p^* for all $\alpha \geq 0$. One obtains the best lower bound when maximizing over set of α - this yields d^* . Hence $d^* \leq p^*$ for any $f(w)$ and $g(w)$. However, if certain conditions are met, namely

1. $f(w)$ is convex
2. $g(w)$ is affine. i.e. $g(w) = a^T w + b$

then $d^* = p^*$.

For the SVM problem, both these conditions hold.



9.1.5 (Optional) geometric interpretation

For those that want an geometric intuition to the above conditions, consider all the possible values of w . For each one, compute the pair of values of $(f(w), g(w))$ and plot them in a 2D space where the x-axis is $g(w)$ and y-axis is $f(w)$. This defines a set of reachable $(f(w), g(w))$ pairs - call this set Ω .

One can compute p^* for this visualization by inspection. Look at the part of Ω that is left of the y axis - this is region of feasible w values for which $g(w) \leq 0$. Amongst this set, select the point with the lowest x value.

We can also compute d^* in this visualization. Consider the set of all f and g values for which the Lagrangian from (9.16) equals some constant L for a fixed α . This is a line of slope $\frac{1}{\alpha}$ that intercepts the y axis at a value of L (since $L = f + \alpha g$ and $g = 0$ at the y intercept).

As we vary L for a fixed α , we obtain a family of parallel lines. Therefore, for a fixed α , $\min_w L(w, \alpha)$ picks the line that passes through Ω (the set of realizable (f, g) values) with the lowest y intercept. This is a lower tangent line of the set Ω .

If we vary α , we will obtain a different lower tangent of Ω . Now, maximizing $[\min_w L(w, \alpha)]$ over all positive α corresponds to picking the lower tangent line with the largest y intercept. d^* corresponds to the y intercept of this particular tangent line.

The above construction shows that $d^* < p^*$ for any set Ω . When Ω is a convex region, the above construction shows that $d^* = p^*$. A convex region satisfies the property that any line that connects 2 points in the region also lies within the region - there are no holes or “crevices”. In particular, the lower envelope of a convex region can be represented as the maximum of a set of lines - this is precisely what the dual formulation is doing. When $f(w)$ is convex and $g(w)$ is affine, Ω will be a convex set. Hence for our SVM problem,

$$\min_w \max_{\alpha \geq 0} L(w, \alpha) = \max_{\alpha \geq 0} \min_w L(w, \alpha) \quad (9.22)$$

9.1.6 Conditions for Optimality (Karush-Kuhn-Tucker Conditions)

Lagrangian duality theory also states a number of necessary and sufficient conditions hold at the optimum solution.

1. w, α are feasible
2. $\alpha g(w) = 0$

Condition 1 means that $g(w) \leq 0$ and $\alpha \geq 0$. Condition 2 is called “complimentary slackness”. It follows from the fact that the constraint $g(w)$ may or may not affect the final solution. If the minimum of $f(w)$ lies within the region $\{w : g(w) < 0\}$, then one can optimize $f(w)$ without regard to the constraint (ie, let $\alpha = 0$). If the minimum of $f(w)$ lies outside this set, then the constraint is turned “on”, and the final solution must satisfy $g(w) = 0$. In this case, α behaves like a typical Lagrange multiplier for an equality constraint.

9.1.7 SVM recap

The Lagrangian corresponding to (9.11) problem is defined as

$$L(w, b, \alpha) = \frac{1}{2}w^T w - \sum_{i=1}^m \alpha_i [y^{(i)}(w^T x^{(i)} + b) - 1] \quad (9.23)$$

Note that we have added a separate lagrange variable α_i for each constraint from (9.11), and that the constraints have been written in the form $g_i(w) \leq 0$. The solution to the original problem is

$$\min_{w, b} \max_{\alpha} L(w, b, \alpha) \quad (9.24)$$

The solution of the dual problem is

$$\max_{\alpha} \min_{w, b} L(w, b, \alpha) \quad (9.25)$$

Since we are optimizing a convex function with linear constraints, the dual solution will equal the primal solution (see Section 9.1.4).

To optimize the dual (9.25), we need to maximize $L(w, b, \alpha)$ with respect to w and b for a fixed value of α . We know that the optimal w and b must satisfy the condition that the partial derivatives of L w.r.t w and b are 0.

$$\frac{\partial}{\partial w} L(w, b, \alpha) = w - \sum_{i=1}^m \alpha_i y^{(i)} x^{(i)} = 0 \quad (9.26)$$

This implies that

$$w = \sum_{i=1}^m \alpha_i y^{(i)} x^{(i)} \quad (9.27)$$

Similarly,

$$\frac{\partial}{\partial b} L(w, b, \alpha) = \sum_{i=1}^m \alpha_i y^{(i)} = 0 \quad (9.28)$$

Hence for a fixed value of α , we have a closed form solution for the w that maximizes $L(w, b, \alpha)$. We also have a condition on the sum of $\alpha_i y^{(i)}$. We can plug them back into the dual expression.

$$L(w, b, \alpha) = \frac{1}{2} \sum_{i,j=1}^m \alpha_i \alpha_j y^{(i)} y^{(j)} (x^{(i)})^T x^{(j)} - \sum_{i,j=1}^m \alpha_i \alpha_j y^{(i)} y^{(j)} (x^{(i)})^T x^{(j)} - b \sum_{i=1}^m \alpha_i y^{(i)} + \sum_{i=1}^m \alpha_i \quad (9.29)$$

But since $\sum_{i=1}^m \alpha_i y^{(i)} = 0$, we get

$$L(\alpha) = \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i,j=1}^m \alpha_i \alpha_j y^{(i)} y^{(j)} (x^{(i)})^T x^{(j)} \quad (9.30)$$

Finally, we are left with a function of α what we wish to maximize. Putting this together with the constraints $\alpha_i \geq 0$ and the constraint $\sum_{i=1}^m \alpha_i y^{(i)} = 0$, we obtain the following optimization problem

$$\begin{aligned} \max_{\alpha} L(\alpha) &= \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i,j=1}^m \alpha_i \alpha_j y^{(i)} y^{(j)} (x^{(i)})^T x^{(j)} \\ \text{s.t. } &\alpha_i \geq 0, i = 1, \dots, m \\ &\sum_{i=1}^m \alpha_i y^{(i)} = 0 \end{aligned} \quad (9.31)$$

This is dual optimization problem corresponding to the primal one from (9.11). Our claim is that the dual problem is more computationally convenient. We validate this claim by considering the KKT conditions, which must hold at the solution. In particular, the complementary slackness conditions can be written as

$$\alpha_i = 0 \implies y^{(i)}(w^T x^{(i)} + b) \geq 1 \quad (9.32)$$

$$\alpha_i > 0 \implies y^{(i)}(w^T x^{(i)} + b) = 1 \quad (9.33)$$

$$(9.34)$$

These conditions mathematically validate our original *sparseness* intuition. Points that lie beyond the margin will have $\alpha_i = 0$, and so will not effect the final solution $w = \sum \alpha_i y^{(i)} x^{(i)}$ (9.27). Consider an iterative optimization algorithm that gradually updates the α_i . If it could quickly determine which of the α_i are zero, it could then focus computation on those points near the decision boundary. The final set of points with nonzero α_i , or alternatively, the set of points with margin 1, are called the *support vectors*.