Massachusetts Institute of Technology
6.435 Theory of Learning and System Identification
(Spring 2007)

Prof. Dahleh, Prof. Mitter — Lecture 8
Not scribed — T 3/6

In this lecture, we introduce support vector machines (SVMs). The technique refers to a model class of separating hyperplanes and an algorithm which finds the hyperplane with optimal margin. The terminology comes from the fact that the characterization of the optimal hyperplane consists of a select number of data samples, the so-called support vectors. We will give the setup of the problem, and show how dualizing the optimization gives the aforementioned characterization. Interest in SVMs is due to the non-linear extensions with kernels, and due to its theoretical generalization ability (fast convergence of the empirical risk). We will elaborate these in the next lecture.

1 Support Vector Machines

1.1 Optimal Separating Hyperplane

We are given a sample set z^ℓ, $z_i = (x_i, y_i)$, $x_i \in \mathbf{R}^n$ and $y_i \in \{-1\}$. We assume that the samples are separable, in the sense that a separating hyperplane parameterized by $\phi \in \mathbf{R}^n$ and $c \in \mathbf{R}$ exists such that:

$$\begin{aligned} \phi' x_i > c \quad & \text{if } y_i = 1 \\ \phi' x_i < c \quad & \text{if } y_i = -1 \end{aligned}$$

We denote the index sets by

$$\begin{aligned} I_1 = & \quad \{i : y_i = 1\}, \\ I_2 = & \quad \{i : y_i = -1\}. \end{aligned}$$

The *margin* of the separting hyperplane is denoted by $\rho(\phi)$ and is defined as follows:

$$\begin{aligned} c_1(\phi) &\triangleq \min_{i \in I_1} \phi' x_i \\ c_2(\phi) &\triangleq \max_{i \in I_2} \phi' x_i \\ \rho(\phi) &\triangleq \tfrac{1}{2} \left[c_1(\phi) - c_2(\phi) \right] \end{aligned}$$

We would like to define an optimal hyperplane to be one that maximizes this margin. However, in order for the scale not to be a factor, we would require the norm of ϕ to be bounded. More precisely, an optimal hyperplane is the solution of the following optimization:

$$\begin{aligned} (\phi^\circ, c^\circ) = \quad & \text{argmax}_{\phi, c} \quad \rho(\phi) \\ & \text{s.t.} \quad |\phi| \le 1 \\ & \qquad \phi' x_i > c, \quad i \in I_1 \\ & \qquad \phi' x_i < c, \quad i \in I_2, \end{aligned}$$

1.2 Existence and Uniqueness

Claim. *The optimal ϕ° exists and is unique.*

Proof. Sketch:

- Existence follows from continuity of $\rho(\phi 0)$ and compactness of the domain.
- For uniqueness:
  - · ϕ° is on the boundary, i.e. $|\phi^\circ| = 1$. Indeed, if $|\phi^\circ| < 1$ then $\frac{\phi^\circ}{|\phi^\circ|}$ is a solution (with an appropriately chosen c°), and it produces a larger margin, which cannot be.

- For the sake of contradiction, suppose there exist distinct ϕ_1 and ϕ_2, both optimal. It follows that they both lie on the boundary. But observe that $\lambda\phi_1 + (1-\lambda)\phi_2$ is also optimal, yet it lies in the interior, which is the contradiction we desire.

□

Corollary. $c^\circ = \frac{c_1(\phi^\circ)+c_2(\phi^\circ)}{2}$ *is optimal and unique.*

2 LQP Formulation and Duality

We start by observing the follow:

- When $i \in I_1$, we have that $\phi^{\circ\prime}x_i - c^\circ > c_1(\phi^\circ) - c^\circ = \rho(\phi^\circ)$.
- When $i \in I_2$, we have that $\phi^{\circ\prime}x_i - c^\circ < c_2(\phi^\circ) - c^\circ = -\rho(\phi^\circ)$.

We can write these more compactly as follows:

$$y_i(\phi^{\circ\prime}x_i - c^\circ) > \rho(\phi^\circ), \qquad \forall i.$$

If we normalize by $\rho(\phi^\circ)$, we obtain:

$$y_i\left[\underbrace{\left(\frac{\phi^\circ}{\rho(\phi^\circ)}\right)'}_{\psi} x_i - \underbrace{\frac{c^\circ}{\rho(\phi^\circ)}}_{b}\right] > 1, \qquad \forall i.$$

2.1 LQP Primal Formulation

From above, we have that:

$$y_i(\psi' x_i - b) > 1, \qquad \forall i.$$

Intuitively, maximizing the margin should be equivalent to minimizing $|\psi|$. This is indeed the case, as is formalized in the following proposition, which casts the optimization as a LQP problem.

Proposition 1. *Let:*

$$\begin{array}{rll} (\psi^\circ, b^\circ) = & \mathrm{argmin}_{\psi,b} & \frac{1}{2}\psi'\psi \\ & \text{s.t.} & y_i(\psi' x_i - b) > 1, \quad \forall i. \end{array}$$

Then:

$$\phi^\circ = \frac{\psi^\circ}{|\psi^\circ|}, \quad c^\circ = \frac{b^\circ}{|\psi^\circ|}$$

characterize the optimal hyperplane, and $\rho(\phi^\circ) = |\psi^\circ|^{-1}$.

Proof. Let ϕ° be defined as above. Then:

$$\begin{aligned} \rho(\phi^\circ) &= \frac{1}{2}\left[c_1(\phi^\circ) - c_2(\phi^\circ)\right] \\ &= \frac{1}{2|\psi^\circ|}\left[\underbrace{\min_{I_1} \psi' x_i}_{\geq 1+b} - \underbrace{\max_{I_2} \psi' x_i}_{\leq -1+b}\right] \\ &\geq \frac{1}{|\psi^\circ|} \end{aligned}$$

To show optimality of the margin, it's therefore enough to show that it is impossible to have a strict inequality, i.e. $\rho(\phi^\circ) > |\psi^\circ|^{-1}$. For the sake of contradiction, assume that is indeed the case. Define:

$$\tilde{\psi} = \frac{\phi^\circ}{\rho(\phi^\circ)} = \frac{\psi^\circ}{|\psi^\circ|\rho(\phi^\circ)}.$$

Notice that $\tilde{\psi}$ satisifies the constraints (with properly chosen $\tilde{b}$) since for all i:

$$\begin{aligned}
& y_i(\psi^{\circ\prime} x_i - b) > 1 \\
\therefore \quad & y_i(\phi^{\circ\prime} x_i - c^{\circ}) > \tfrac{1}{|\psi^{\circ}|} > \rho(\phi^{\circ}) \\
\therefore \quad & y_i \left[\left(\tfrac{\phi^{\circ}}{\rho(\phi^{\circ})} \right)' x_i - \tfrac{c^{\circ}}{\rho(\phi^{\circ})} \right] > 1 \\
\therefore \quad & y_i(\tilde{\psi}' x_i - \tilde{b}) > 1.
\end{aligned}$$

It follows that $\tilde{\psi}$ is feasible. But, by hypothesis, $|\psi^{\circ}|\rho(\phi^{\circ}) > 1$, thus $|\tilde{\psi}| < |\psi^{\circ}|$ and therefore ψ° is not optimal, which is the contradiction we desire. □

2.2 LQP Dual Formulation

By introducing the multipliers $a_1, \cdots, a_\ell$, we can write the Lagrangian of the primal problem as follows:

$$\mathcal{L}(\psi, b, a) = \frac{1}{2}\psi'\psi - \sum_{i=1}^{\ell} a_i \left[y_i(\psi' x_i - b) - 1 \right]$$

The constrained optimization is therefore equivalent to:

$$\max_{a_i \geq 0} \min_{\psi, b} \mathcal{L}(\psi, b, a)$$

The conditions for optimality are:

$$\frac{\partial \mathcal{L}}{\partial \psi} = 0 \quad \text{and} \quad \frac{\partial \mathcal{L}}{\partial b} = 0,$$

and the alignment conditions are:

$$a_i \left[y_i(\psi' x_i - b) - 1 \right] = 0.$$

Thus:

$$\frac{\partial \mathcal{L}}{\partial \psi} = \psi^{\star} - \sum_{i=1}^{\ell} a_i y_i x_i = 0 \quad \Rightarrow \quad \psi^{\star} = \sum_{i=1}^{\ell} a_i y_i x_i,$$

$$\frac{\partial \mathcal{L}}{\partial b} = \sum_{i=1}^{\ell} a_i y_i = 0.$$

Observe that it follows that:

$$\psi^{\star\prime} x_j = \sum_{i=1}^{\ell} a_i y_i (x_i' x_j),$$

and after some manipulation the dual problem can be written as:

$$\begin{array}{rll}
a^{\circ} = & \text{argmax}_a & \sum_{i=1}^{\ell} a_i - \frac{1}{2} \sum_{i,j=1}^{\ell} a_i a_j y_i y_j x_i' x_j \\
& \text{s.t.} & a_i \geq 0 \quad \forall i, \\
& & \sum_{i=1}^{\ell} a_i y_i = 0.
\end{array}$$

Finally, the alignment conditions dictate that:

$$a_i^{\circ} \neq 0 \quad \text{iff} \quad y_i(\psi^{\circ\prime} x_i - b) = 1.$$

2.3 Observations

The dual problems reveals a host of interesting properties. As an immediate consequence of the alignment conditions, ψ° ($\psi^\star$ evaluated at a°) is expressed in terms of the support vectors alone. This representation may not be unique, due to multiple dual solutions, but the fact that only a handful of vectors dictate the solution will be crucial in the next section and lecture, when proving the statistical properties of SVMs.

Moreover, from the structure of the dual problem, it is clear that the solution depends only on the inner products $x_i' x_j$, and not on the individual points. This is the main idea used in kernal extensions of SVMs.

Finally, it is worth mentioning that there is no duality gap. The primal optimal cost is $\frac{1}{2}\psi^{\circ\prime}\psi^\circ$, and due to the alignment conditions we have:

$$\mathcal{L}(\psi^\circ, b^\circ, a^\circ) = \frac{1}{2}\psi^{\circ\prime}\psi^\circ.$$

Also, by summing all alignment conditions, we obtain:

$$\sum a_i^\circ y_i(\psi^{\circ\prime} x_i - b) = \sum a_i^\circ y_i \left(\sum a_j^\circ y_j x_j'\right) x_i = \sum a_i^\circ,$$

which gives us the dual cost:

$$\frac{1}{2}\sum_{i=1}^{\ell} a_i^\circ = \mathcal{L}(\psi^\circ, b^\circ, a^\circ) = \frac{1}{2}\psi^{\circ\prime}\psi^\circ.$$

3 Statistical Properties

In this section, we outline a theoretical motivation for using SVMs. In particular, we state a result which shows that the optimal hyperplane has a desirable convergence property.

Let K_ℓ= # of essential support vectors, that is the cardinality of the intersection of all sets of possible support vectors, resulting from different solutions of the dual problem.

Claim. *We have:*

$$K_\ell \leq n$$

Proof. (next time)

The claim follows from using the representation $\psi^\circ = \sum a_i^\circ y_i x_i$. □

In our analysis so far, we have considered convergence properties with respect to the model class itself. In that context, recall that our task was reduced to obtaining a parameter α that minimizes empirical risk. When many solutions are possible, e.g. many separating hyperplanes exist, the details of the algorithm can provide us more information.

Let us specify an algorithm, which is nothing but a family of function $\alpha_\ell(\cdot)$:

$$\begin{array}{rrcl} \alpha_\ell \quad : & \mathcal{Z}^\ell & \rightarrow & \mathcal{C} \\ & z^\ell & \mapsto & (\psi, b) \\ & (x_1, y_1), \cdots, (x_\ell, y_\ell) & \mapsto & \text{optimal hyperplane.} \end{array}$$

It follows that $\alpha_\ell(Z^\ell)$ is a random variable, and we can consider the associated ensemble performance, i.e. the expected value of the risk:

$$\mathbf{E}\left[R\left(\alpha_\ell(Z^\ell)\right)\right] \overset{\triangle}{=} \mathbf{E}\left[L\left(X, Y; \alpha_\ell(Z^\ell)\right)\right].$$

Claim. *We have:*

$$\mathbf{E}\left[R\left(\alpha_\ell(Z^\ell)\right)\right] \leq \frac{\mathbf{E}[K_{\ell+1}]}{\ell+1}.$$

Proof. (next time)

The claim follows from a technique in the style of cross-validation. One data point is removed at a time, and only support vectors affect the solution if removed. □

This convergence result can be shown to surpass that of non-optimal hyperplane algorithms, such as the perceptron.

Finally, this line of investigation is related to the area of active learning, where data is chosen adaptively. By choosing data such that K_ℓ is kept small, faster convergence can be achieved. Despite a gain in speed, disadvantages of the technique include the fact that errors in the data, if allowed, are reinforced.