CSCI 5525 Machine Learning Fall 2019

Lecture 5: Support Vector Machine (Part 1)

Sep 17 2019

Lecturer: Steven Wu *Scribe: Steven Wu*

At this point, we have encountered the concept of *margin maximization* several times in the context of classification. Why is maximizing margin a good idea? Suppose the training data is linearly separable. In this case, there are actually infinitely many linear predictors that can achieve zero training error. An intuive solution to break ties is to select the predictor that maximizes the distance between the data points and the decision boundary, which is given by a hyperplane in this case. Now let us write this as an optimization.

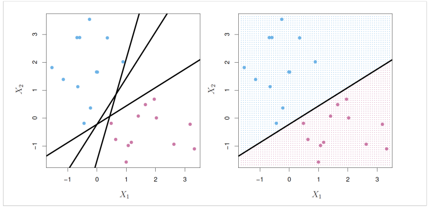

Figure 1: There are inifinitely many hyperplanes that can classify all the training data correctly. We are looking for the one that maximizes the margin. Image source.

For any linear predictor parameterized by a weight vector $\mathbf{w} \in \mathbb{R}^d$, the decision boundary is the hyperplane $H = \{x \in \mathbb{R}^d \mid \mathbf{w}^\intercal x = 0\}$. If the linear predictor perfectly classifies has zero training error, then we know that for all $(x_i, y_i) \in \mathbb{R}^d \times \{\pm 1\}$: $y_i \mathbf{w}^\intercal x_i > 0$. The distance between the point $y_i x_i$ and H is given by

$$\frac{y_i \mathbf{w}^\intercal x_i}{\|\mathbf{w}\|_2}$$

The smallest distance from all training points to the hyperplane is given by

$$\min_i \frac{y_i \mathbf{w}^\intercal x_i}{\|\mathbf{w}\|_2}$$

Then the problem of maximing the margin or distance from the separating hyperplane to the training data is:

$$\max_{\mathbf{w}} \min_i \frac{y_i \mathbf{w}^\intercal x_i}{\|\mathbf{w}\|_2}$$

Observe that rescaling the vector $\mathbf{w}$ does not actually change the relevant hyperplane and the distances. It suffices to consider the set $\mathbf{w}$'s such taht $\min_i y_i \mathbf{w}^\intercal x_i = 1$. Then the problem of margin maximization becomes

$$\max_{\mathbf{w}} \frac{1}{\|\mathbf{w}\|_2} \quad \text{such that} \quad \min_i y_i(\mathbf{w}^\intercal x_i) = 1$$

Or equivalently,

$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|_2^2 \quad \text{such that} \quad \forall i, y_i(\mathbf{w}^\intercal x_i) \geq 1 \tag{1}$$

Note that the objectives are the same, but the constraints are different. (Why are two optimization problems equivalent?)

The optimization problem in (1) computes the linear classifier with the largest margin—the support vector machine (SVM) classifier. The solution is also unique.

Soft-Margin SVM

More generally, the training examples may not be linearly separable. To handle this case, we will introduce *slack variables*. For each of the n training examples, we will introduce an non-negative variable ξ_i, and the optimization problem becomes:

$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^{n} \xi_i \quad \text{such that} \tag{2}$$

$$\forall i, \qquad y_i(\mathbf{w}^\intercal x_i) \geq 1 - \xi_i \tag{3}$$

$$\forall i, \qquad \xi_i \geq 0 \tag{4}$$

This is also called the *soft-margin* SVM. Note that $\xi_i / \|\mathbf{w}\|_2$ is the distance the example i need to move to satisfy the constraint $y_i(\mathbf{w}^\intercal x_i) \geq 1$.

Equivalently, the soft-margin SVM problem can be written as the following unconstrained optimization problem, which replaces the second term with hinge losses:

$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^{n} \underbrace{\max\{0, 1 - y_i \mathbf{w}^\intercal x_i\}}_{\text{hinge loss}}$$

But now here is a more generic idea that turns a constrained optimization problem into an unconstrained optimization problem.

Detour of Duality

In general, consider the following constrained optimization problem:

$$\min_{\mathbf{w}} F(\mathbf{w}) \quad \text{s.t.} \quad h_j(\mathbf{w}) \leq 0 \quad \forall j \in [m]$$

For each of the constraint, we can introduce a Lagrangian multiplier $\lambda_j \geq 0$, and write down the following Lagrangian function:

$$L(\mathbf{w}, \lambda) = F(\mathbf{w}) + \sum_{j=1}^{m} \lambda_j h_j(\mathbf{w})$$

Note thats $\max_\lambda L(\mathbf{w},\lambda)$ is ∞ whenever the $\mathbf{w}$ violates one of the constraints. This means the solution to the following problem

$$\min_{\mathbf{w}} \max_\lambda L(\mathbf{w},\lambda)$$

is exactly the solution to constrained optimization problem. (Why?) Now let's swap min and max, and consider the following problem:

$$\max_\lambda \min_{\mathbf{w}} L(\mathbf{w},\lambda)$$

Let $\mathbf{w}^* = \arg\min_{\mathbf{w}}(\max_\lambda L(\mathbf{w},\lambda))$ and $\lambda^* = \arg\max_\lambda(\min_{\mathbf{w}} L(\mathbf{w},\lambda))$. We can derive the following:

$$\begin{aligned}
\max_\lambda \min_{\mathbf{w}} L(\mathbf{w},\lambda) &= \min_{\mathbf{w}} L(\mathbf{w},\lambda^*) && \text{(definition of } \lambda^*\text{)}\\
&\le L(\mathbf{w}^*,\lambda^*) && \text{(definition of min)}\\
&\le \max_\lambda L(\mathbf{w}^*,\lambda) && \text{(definition of max)}\\
&= \min_w \max_\lambda L(\mathbf{w},\lambda) && \text{(definition of } \mathbf{w}^*\text{)}
\end{aligned}$$

The relationship of maxmin $\le$ minmax is called *weak duality*.

Under "mild" condition (e.g. convex quadratic problem, the so-called Slater's condition), we also have *strong duality*:

$$\max_\lambda \min_{\mathbf{w}} L(\mathbf{w},\lambda) = \min_{\mathbf{w}} \max_\lambda L(\mathbf{w},\lambda)$$

Under strong duality, we can further write:

$$\begin{aligned}
F(\mathbf{w}^*) &= \min_{\mathbf{w}} \max_\lambda L(\mathbf{w},\lambda) && \text{(definition of } \mathbf{w}^*\text{)}\\
&= \max_\lambda \min_{\mathbf{w}} L(\mathbf{w},\lambda) && \text{(strong duality)}\\
&= \min_{\mathbf{w}} L(\mathbf{w},\lambda^*) && \text{(definition of } \lambda^*\text{)}\\
&\le L(\mathbf{w}^*,\lambda^*) && \text{(definition of min)}\\
&= F(\mathbf{w}^*) + \sum_j \lambda_j^* h_j(\mathbf{w}^*) && \text{(definition of } L\text{)}
\end{aligned}$$

Note that since $\mathbf{w}^*$ is a feasible solution such that $h_j(\mathbf{w}^*) \le 0$ for all j, each term $\lambda_j^* h_j(\mathbf{w}^*) \ge 0$, and so the inequality above should also just be equality. This has the following implications:

- (Complementary slackness): last equality implies that $\lambda_j^* h_j(\mathbf{w}^*) = 0$ for all j.
- (Stationarity): $\mathbf{w}^*$ is the minimizer of $L(\mathbf{w},\lambda^*)$ and thus has gradient zero

$$\nabla_{\mathbf{w}} L(w^*,\lambda^*) = \nabla F(w^*) + \sum_j \lambda_j^* \nabla h_j(\mathbf{w}^*) = \mathbf{0}$$

- (Feasibility): $\lambda_j \ge 0$ and $h_j(\mathbf{w}^*) \le 0$ for all j.

These are also called the KKT conditions, which are necessary conditions for the optimal solutions. But they are sufficient when F is convex and the set of h_j are continuously differentiable convex functions.

What is KKT? This was previously known as the KT (Kuhn-Tucker) conditions since the conditions first appeared in publication by Kuhn and Tucker in 1951. However, later people found out that Karush had stated the conditions in his unpublished master's thesis of 1939.