SVMs

CS 360 Machine Learning
Week 8, Day 2

March 21, 2024

Contents

1 Support Vector Machines (SVMs)

Recall that the support vectors are the vectors supporting the margins, and the margins are defined by the closest points from the training dataset to the separating hyperplane. When we have a maximal margin hyperplane, the distance between the points supporting the vectors and the hyperplane is maximized.

There are some key theoretical aspects of SVMs that we'll focus on today. Recall that the goal of SVMs is to create a hyperplane between positive and negative labeled points in our training data such that the margin of that hyperplane touches the support vector points (the closest points to that hyperplane). Let γ denote the *overall geometric margin* which is the minimum over all of our margin point distances to the hyperplane, i.e.:

$$\gamma = \min_{i=1...n} \gamma_i$$

where γ_i is the distance from point $\vec{x}_i$ to the hyperplane and is known as the *geometric margin* for $\vec{x}_i$. The goal is to maximize γ.

To do this, we first need to compute the γ_i values. Consider some $\vec{x}_i$. We know that the weight vector $\vec{w}$ is perpendicular to the hyperplane. Suppose that $\vec{w}$ is pointing towards the same side of the hyperplane that contains $\vec{x}_i$. Let $\vec{p}$ be the vector from the origin to the hyperplane, such that it connects with the hyperplane at the point where a perpendicular vector from $\vec{x}_i$ would hit the hyperplane. That perpendicular from $\vec{x}_i$ is the same direction as $\vec{w}$.

Then:

$$\vec{p} + \vec{w} = \vec{x}_i$$

The distance from $\vec{x}_i$ to the hyperplane is denoted γ_i, so this can be rewritten as:

$$\vec{p} + \gamma_i y_i \frac{\vec{w}}{||\vec{w}||} = \vec{x}_i$$

The value $||\vec{w}|| = \sqrt{w_1^2 + w_2^2 + ... + w_p^2}$ and is known as the norm of $\vec{w}$. We take the direction of the vector and divide by the weight of the vector to normalize it, which gives us a unit vector version of $\vec{w}$. The norm term is tricky to take the derivative of, so we'll do some work to clean up this equation. Note that the y_i term (which is -1 or 1) is included to handle the cases when $\vec{w}$ doesn't point in the same direction as $\vec{x}_i$.

Now, we want to solve the above for γ_i. Unfortunately we also don't know what $\vec{p}$ is, so we have two unknowns. Recall that our SVM model is $h_{\vec{w}}(\vec{x}_i) = sign(\vec{w} \cdot \vec{x}_i + b)$. Recall that for points on the hyperplane, this value is 0. We know from this that $\vec{w} \cdot \vec{p} + b = 0$ since $\vec{p}$ is defined to be on the hyperplane. Now we have two equations so we can solve for our two unknowns.

First, we'll manipulate the equations to make this easier. We begin by multiplying by $\vec{w}$:

$$\vec{w} \cdot \left(\vec{p} + \gamma_i y_i \frac{\vec{w}}{||\vec{w}||} = \vec{x}_i \right)$$

$$\vec{w} \cdot \vec{p} + \gamma_i y_i \frac{\vec{w} \cdot \vec{w}}{||\vec{w}||} = \vec{w} \cdot \vec{x}_i$$

Rewriting $\vec{w} \cdot \vec{p} + b = 0$ to be $\vec{w} \cdot \vec{p} = -b$ and substituting into the above (and using that $\vec{w} \cdot \vec{w} = ||\vec{w}||^2$) we get:

$$-b + \gamma_i y_i ||\vec{w}|| = \vec{w} \cdot \vec{x}_i$$

Rearranging these terms and using the fact that $y_i = \frac{1}{y_i}$ we can solve for the geometric margin γ_i:

$$\gamma_i = y_i \left(\frac{\vec{w}}{||\vec{w}||} \cdot \vec{x}_i + \frac{b}{||\vec{w}||} \right)$$

Recall that we're trying to solve for $\vec{w}$ while maximizing the minimum γ_i. Usually people do this by trying to optimize the *functional margin* $\hat{\gamma}_i$ determined by multiplying through by the magnitude of $\vec{w}$:

$$\hat{\gamma}_i = y_i(\vec{w} \cdot \vec{x}_i + b)$$

Note that this means that $\gamma_i = \frac{\hat{\gamma}_i}{||\vec{w}||}$. Recall that $\hat{y}_i = h_{\vec{w}}(\vec{x}_i) = sign(\vec{w} \cdot \vec{x}_i + b)$, and that our goal is to maximize the minimum distance between any training example and the hyperplane, i.e. letting $\gamma = \min_{i=1...n} \gamma_i$ we want to:

$$\max_{\gamma, \vec{w}, b} \gamma$$

$$\text{s.t } y_i \Big(\vec{w} \cdot \vec{x}_i + b \Big) \geq \gamma, \text{ for } i = 1, ..., n$$

$$\text{and } ||\vec{w}|| = 1$$

The constraint $||\vec{w}|| = 1$ is arbitrary, added to help solve this optimization problem, however there is still a square root in the constraint, which is non-convex and makes it hard to solve this problem. We can get rid of this by optimizing where we've divided by this norm:

$$\max_{\hat{\gamma}, \vec{w}, b} \frac{\hat{\gamma}}{||\vec{w}||}$$

$$\text{s.t. } y_i \Big(\vec{w} \cdot \vec{x}_i + b \Big) \geq \hat{\gamma}, \text{ for } i = 1, ..., n$$

However this leaves a square root in the denominator of the objective function and we can't optimize that easily. If we instead take the inverse of this equation and minimize instead of maximizing, and add a constraint on the functional margin such that $\hat{\gamma} = 1$, this gives us:

$$\min_{\vec{w}, b} \frac{1}{2} ||\vec{w}||^2$$

$$\text{s.t. } y_i \Big(\vec{w} \cdot \vec{x}_i + b \Big) \geq 1, \text{ for } i = 1, ..., n$$

Now we can add a squared term since minimizing the norm is the same as minimizing the norm squared. We include the 1/2 term so that it'll work out nicely when we take the derivative. Finally, we take the negative of our constraints and subtract 1 so that it's relative to 0.

$$\min_{\vec{w},b} \frac{1}{2}||\vec{w}||^2$$

$$\text{s.t. } -y_i\Big(\vec{w}\cdot\vec{x}_i + b\Big) + 1 \leq 0, \text{ for } i = 1, ..., n$$

Now we have this in a form where we can solve the optimization function! This will give us a separating hyperplane with a good functional margin, and is the algorithm used to train an SVM.