Support Vector Machines and an intro to convexity

Consider classification of linearly separable examples. We saw that the perceptron could find a separating hyperplane. But not all hyperplanes are equivalent, even if they classify all examples properly. Intuitively (and formally, via several generalization results) the larger the distance between the examples and the separating hyperplane, the more desirable it is.

We will formulate an optimization problem for training that not only tries to get a separating hyperplane, but also one that will ensure that the examples are as far away from it as possible. To do this, first note the following exercise.

Exercise Show that the distance of any point $\mathbf{z}_i \in \mathbb{R}^p$ from a hyperplane $\mathbf{w}^T\mathbf{x} - b = 0$ (where $\mathbf{w}$ and $\mathbf{x}$ are in $\mathbb{R}^p$, and $b \in \mathbb{R}$ is a number) is given by

$$\frac{\mathbf{w}^T\mathbf{z}_i - b}{||\mathbf{w}||}.$$

Training for maximum margin

Suppose $\mathbf{z}_1, \ldots, \mathbf{z}_n$ are n training examples in $\mathbb{R}^p$ given to us with labels $y_1, \ldots, y_n$ respectively (each label is either $+1$ or -1). Let $\mathbf{w} \in \mathbb{R}^p$ and b be a number, and define

$$\gamma_i(\mathbf{w}, b) = \mathbf{w}^T\mathbf{z}_i - b.$$

Therefore, the distances of the n points to the plane $\mathbf{w}^{\mathbf{x}} - b = 0$ are respectively $\gamma_1/||\mathbf{w}||, \ldots, \gamma_n/||\mathbf{w}||$. In addition, let

$$\gamma(\mathbf{w}, b) = \min_{1 \le i \le n} \gamma_i(\mathbf{w}, b) = \min_{1 \le i \le n} \mathbf{w}^T\mathbf{z}_i - b \tag{1}$$

so that the smallest distance between the examples and the hyperplane is $\gamma(\mathbf{w}, b)/||\mathbf{w}||$. This is called the *margin* of the classifier $\mathbf{w}^T\mathbf{x} - b = 0$.

From our training data, we want to obtain that plane $\mathbf{w}^T\mathbf{x} - b = 0$ which classifies all examples correctly, but in addition has the largest margin. This plane is what we will learn from the training example, and what we will use to predict on the test examples.

So for training, we first set up an optimization. Note that γ is some complicated function of $\mathbf{w}$ and b. Different values of $\mathbf{w}$ and b yield potentially different orientations and intercepts of the separating hyperplane, and their margin is determined by different examples (*i.e.,* the minimizer in (1) is

different). Even though we may not have $\gamma(\mathbf{w}, b)$ in a simple form, we can still ask for

$$\mathbf{w}^*, b^* = \arg\max \mathbf{w}, b \frac{\gamma(\mathbf{w}, b)}{||\mathbf{w}||}$$
$$\text{subject to } y_i(\mathbf{w}^T\mathbf{z}_i - b) \geq 0 \text{ for all } 1 \leq i \leq n.$$

In the optimization above, the first line asks to maximize the margin, while the constraints (there are n of them) ensure that each example is classified properly.

So far so good, but we don't really want to compute $\gamma(\mathbf{w}, b)$ or try expressing it in any closed/numerical form. But there is a simple conceptual way around it. Suppose $\mathbf{w}$ and $\mathbf{b}$ classified all examples such that every example, $\mathbf{z}_1, \ldots, \mathbf{z}_n$ satisfied

$$y_i(\mathbf{w}^T\mathbf{z}_i - b) \geq \nu, \qquad 1 \leq i \leq n. \tag{2}$$

For a given $\mathbf{w}$ and b, since $\gamma(\mathbf{w}, b)/||\mathbf{w}||$ happens to be the distance of the closest point to the plane $\mathbf{w}^T\mathbf{x} - b = 0$, we could satisfy all n constraints of (2) above for every value of ν in the range $0 \leq \nu \leq \gamma(\mathbf{w}, b)$ and for no other.

Therefore, we ask to find the maximum number ν such that all the constraints in˜(2) are satisfied. Note the shift now—we treat ν as just a number (not a function of $\mathbf{w}$ and b) and see which is the largest combination of the number ν, the vector $\mathbf{w}$ and b that satisfies

$$\mathbf{w}^*, b^*, \nu^* = \arg\max_{\nu, \mathbf{w}, b} \frac{\nu}{||\mathbf{w}||}$$
$$\text{subject to } y_i(\mathbf{w}^T\mathbf{z}_i - b) \geq \nu \text{ for all } 1 \leq i \leq n.$$

We can make one more simplification. There is no distinction between the plane $\mathbf{w}^T\mathbf{x} - b = 0$ and the plane $k(\mathbf{w}^T\mathbf{x} - b) = 0$ for any real number $k \neq 0$ (because if $\mathbf{x}$ satisfies the equation $\mathbf{w}^T\mathbf{x} - b = 0$, it automatically satifies the other and vice versa). So all the candidates $(k\mathbf{w}, kb)$, $k \neq 0$, yield exactly the same plane (and hence same margin). We may choose just one candidate among these while searching for the optimum. To make our life simpler, we can choose k such that

$$\min_{1 \leq i \leq n} k(\mathbf{w}^T\mathbf{z}_i - b) = 1$$

or equivalently, given any $\mathbf{w}$ and b, we scale it by $k = \frac{1}{\gamma}$, where γ is as defined as in˜(1), to get $\tilde{\mathbf{w}}$ and $\tilde{b}$, and optimize over only the $\tilde{\mathbf{w}}$ and $\tilde{b}$.

Then, we will have

$$\min_{1 \le i \le n} (\tilde{\mathbf{w}}^T \mathbf{z}_i - \tilde{b}) = 1$$

and the margin of the hyperplane $\tilde{\mathbf{w}}^T \mathbf{x} - \tilde{b} = 0$ is $1/||\tilde{\mathbf{w}}||$. So we can rewrite our training goal to be the optimization

$$\begin{aligned} \mathbf{w}^*, b^*, \nu^* &= \arg\max_{\nu, \tilde{\mathbf{w}}, b} \frac{1}{||\tilde{\mathbf{w}}||} \\ \text{subject to } & y_i(\tilde{\mathbf{w}}^T \mathbf{z}_i - \tilde{b}) \ge 1 \text{ for all } 1 \le i \le n. \end{aligned}$$

Clearly, the νs are now superflous—they don't exist in either the objective or the constraints—we can discard them. In the above, the $\tilde{\mathbf{w}}$ and $\tilde{b}$ are just dummy variables, we can call them by any other name and nothing will really change. Furthermore, maximizing $1/||\mathbf{w}||$ is the same as minimizing $||\mathbf{w}||$, which is in turn the same as minimizing $\frac{1}{2}||\mathbf{w}||^2$. We can therefore write our training objective as obtaining the hyperplane $(\mathbf{w}^*)^T \mathbf{x} - b^* = 0$, where

$$\begin{aligned} \mathbf{w}^*, b^* &= \arg\min_{\mathbf{w}, b} \frac{1}{2}||\mathbf{w}||^2 \\ \text{subject to } & y_i(\mathbf{w}^T \mathbf{z}_i - b) \ge 1 \text{ for all } 1 \le i \le n. \end{aligned} \tag{3}$$

You may wonder why we transformed maximizing $1/||\mathbf{w}||$ to minimizing $\frac{1}{2}||\mathbf{w}||^2$. The reason is that we want our objectives and constraints to be *convex* functions. We will have a little digression here to define convex functions and sets, but practically every large constrained optimization we can solve is convex (or we just fake the steps of a convex optimization if we are stuck with non-convex optimization). Often, even convex optimization does not look that way to begin with—we need to tweak the formulation as above to get to the correct form.

Convex functions

There are two closely related concepts here: a *convex set* and and a *convex function*. Please read this section carefully—we did not exhaustively go through these definitions at this time. But we will encounter them again shortly when looking at gradient descent in more detail.

Suppose $C \subset \mathbb{R}^d$ is a set of vectors with d coordinates. Then we say C is a convex set if given $\mathbf{x}$ and $\mathbf{x}'$ in C, all points between $\mathbf{x}$ and $\mathbf{x}'$ are also in C. Formally, if $\mathbf{x} \in C$ and $\mathbf{x}' \in C$, we must have for $0 \le \alpha \le 1$, the point $\alpha\mathbf{x} + (1 - \alpha)\mathbf{x}' \in C$,

A *convex function* of d variables is any function f that satisfies for all points $\mathbf{x}$ and $\mathbf{x}'$, and all $0 \leq \alpha \leq 1$ that

$$f(\alpha\mathbf{x} + (1-\alpha)\mathbf{x}') \leq \alpha f(\mathbf{x}) + (1-\alpha)f(\mathbf{x}'), \tag{4}$$

namely the chord connecting the points $(\mathbf{x}, f(\mathbf{x}))$ and $(\mathbf{x}', f(\mathbf{x}))$ lies *above* the surface $g(\mathbf{x}, y) = f(\mathbf{x}) - y = 0$ when we set the arguments of f between $\mathbf{x}$ and $\mathbf{x}'$.

There are many other ways to identify convex functions in some restricted cases. You should think of the following as properties rather than definitions in the strict sense.

Tangents: If f is also differentiable (or in multiple dimensions, the gradient exists), then the tangent plane at any point $(\mathbf{x}_0, f(\mathbf{x}_0))$ (the hyperplane perpendicular to the gradient) lies completely below the surface $g(\mathbf{x}, y) = f(\mathbf{x}) - y = 0$. The tangent interpretation is not a definition since there is no requirement that convex functions have to be differentiable (they are just defined through (4)). This characterization only applies to those convex functions that happen to be differentiable as well, absence of a derivative of a function is not any evidence for convexity/absence thereof.

Exercise Let x be a real number. Is the function $|x|$ (absolute value of x) convex? Is it differentiable everywhere?

(You can skip this derivation and proceed directly to (5) if you wish, but I recommend you try to understand the following.) Mathematically, consider the $d+1$ dimensional space (where we plot the arguments of f in the first d dimensions, followed by the value of f in the last dimension[1]). In this $d+1$-dimensional space, let us plot tangents of the surface $g(\mathbf{x}, y) = f(\mathbf{x}) - y = 0$, where $\mathbf{x}$ corresponds to the $d-$dimensional argument and y is the last dimension that will represent the magnitude of the function (so the surface $f(\mathbf{x}) - y = 0$ sets $y = f(\mathbf{x})$). Specifically, let us look at the tangent to the surface $g(\mathbf{x}, y) = 0$ at the point $\mathbf{z}_0 = (\mathbf{x}_0, f(\mathbf{x}_0)$. This is a plane that is perpendicular to the gradient of g, and which passes through the point above, *i.e.,* all points $\mathbf{z} = (\mathbf{x}, y)$ satisfying

$$(\nabla_{\mathbf{x},y} g)^T_{\mathbf{z}_0} (\mathbf{z} - \mathbf{z}_0) = 0,$$

[1]This is like a 3d plot for a function of 2 variables, the argument of the function is on the $x-y$ plane, and the value $f(x, y)$ is along the z dimension

where $\nabla_{\mathbf{x},y}$ is the gradient with respect to all arguments of g, *i.e.,* all coordinates of $\mathbf{x}$ *and* y. Note that

$$\nabla_{\mathbf{x},y} g = \begin{bmatrix} \nabla_x g \\ \frac{\partial g}{\partial y} \end{bmatrix} = \begin{bmatrix} \nabla_x f \\ -1 \end{bmatrix}.$$

and therefore the tangent is all points $(\mathbf{x}, y)$ satisfying

$$(\nabla_{\mathbf{x},y} g)^T_{\mathbf{z}_0}(\mathbf{z} - \mathbf{z}_0) = (\nabla_{\mathbf{x}} f)^T_{\mathbf{x}_0}(\mathbf{x} - \mathbf{x}_0) - (y - f(\mathbf{x}_0)) = 0,$$

or, reorganizing the above, the tangent plane is all points $(\mathbf{x}, y_{\mathbf{x}})$ satisfying

$$y_{\mathbf{x}} = f(\mathbf{x}_0) + (\nabla_{\mathbf{x}} f)^T_{\mathbf{x}_0}(\mathbf{x} - \mathbf{x}_0).$$

$f(\mathbf{x})$ is the value of the function at any point $\mathbf{x}$. If we require the tangent plane to be below the function, it means that any point on the tangent plane $(\mathbf{x}, y_{\mathbf{x}})$ must be below the point $(\mathbf{x}, f(\mathbf{x}))$. That means, if f is convex with the first derivative, we have for all $\mathbf{x}$ and $\mathbf{x}_0$ that

$$f(\mathbf{x}_0) + (\nabla_{\mathbf{x}} f)^T_{\mathbf{x}_0}(\mathbf{x} - \mathbf{x}_0) \leq f(\mathbf{x}) \tag{5}$$

Hessians: Convex functions that have the second derivatives can be characterized by their Hessians. Looking at (5), and because the quadratic approximation of $f(\mathbf{x})$ from the Taylor series (look at notes for Mar 2) around $\mathbf{x}_0$

$$f(\mathbf{x}_0) + (\nabla_{\mathbf{x}} f)^T_{\mathbf{x}_0}(\mathbf{x} - \mathbf{x}_0) + (\mathbf{x} - \mathbf{x}_0)^T \left(\nabla \nabla^T f\right)_{\mathbf{x}_0} (\mathbf{x} - \mathbf{x}_0),$$

we can conclude that

$$(\mathbf{x} - \mathbf{x}_0)^T \left(\nabla \nabla^T f\right)_{\mathbf{x}_0} (\mathbf{x} - \mathbf{x}_0) \geq 0$$

no matter what $\mathbf{x}$ and $\mathbf{x}_0$ are. In other words the Hessian of f at any point $\mathbf{x}_0$,

$$\left(\nabla \nabla^T f\right)_{\mathbf{x}_0}$$

must be positive-definite (or all eigenvalues are ≥ 0) for f to be convex.

Exercise Let $\mathbf{w} = (w_1, w_2)$ be a vector with two coordinates. Recall that the length of $\mathbf{w}$ is $||\mathbf{w}|| = \sqrt{w_1^2 + w_2^2}$.

1. Compute the Hessians of the function $f(\mathbf{w}) = ||\mathbf{w}||^2$ and the function $h(\mathbf{w}) = ||\mathbf{w}||$.

2. Show that the Hessian of $||\mathbf{w}||^2$ is positive definite (so $||\mathbf{w}||^2$ is convex) but the Hessian of $||\mathbf{w}||$ is NOT positive definite (so $||\mathbf{w}||$ is not a convex function).

Now do you see why we minimize $\frac{1}{2}||\mathbf{w}||^2$ and not $||\mathbf{w}||$ in our formulation (3)? Again, the above characterization only applies to those convex functions that happen to have a second derivative. In general, convex functions need not even have a first derivative leave alone the second—absence of derivatives must not be construed as evidence that the function is not convex.

Level sets: If f is a convex function of $\mathbf{x}$, then all level sets of $\mathbf{x}$, *i.e.,* for all L, the sets

$$f_L = \left\{ \mathbf{x} \in \mathbb{R}^d : f(\mathbf{x}) \leq L \right\}$$

are convex *sets*. The converse need not generally hold, but this is often a quick test that helps you rule out functions that are not convex.