Notes on Support Vector Machines

Dan Sheldon

October 22, 2012

1 Geometry of Linear Functions

- A linear function in $f : \mathbb{R}^d \to \mathbb{R}$ has the form

$$f(\mathbf{x}) = \mathbf{w}^T \mathbf{x}.$$

  The vector $\mathbf{w}$ is the weight vector.

- Consider the case $d = 2$. Then

$$y = f(x_1, x_2) = w_1 x_1 + w_2 x_2. \tag{1}$$

  The *graph* of this function is the set

$$\{(x_1, x_2, f(x_1, x_2)\} \subseteq \mathbb{R}^3,$$

  which is a *surface* in three-dimensional space (just like the graph of a function of one variable is a line in two-dimensional space). We visualize this as (x_1, x_2) defining the location on the horizontal plane (e.g., the latitude and longitude on the map), and $y = f(x_1, x_2)$ as defining the height.

- **Fact:** the graph of a linear function of two variables is a plane.

- One way to see this is by showing that the the **contours** are parallel lines. The contours are the sets of input values with the same height:

$$\{(x_1, x_2) : f(x_1, x_2) = b\} \subseteq \mathbb{R}^2$$

  They have the exact same meaning as contour lines on a topographic map — they are curves[1] in input space along which the height $f(x_1, x_2)$ is constant.

- To see that the contours are lines, solve for the values of x_1 and x_2 such that $f(x_1, x_2) = b$:

$$w_1 x_1 + w_2 x_2 = b \tag{2}$$

$$w_2 x_2 = b - w_1 x_1 \tag{3}$$

$$x_2 = -\frac{w_1}{w_2} x_1 + \frac{b}{w_2}. \tag{4}$$

  Equation (4) defines a line in $\mathbb{R}^2$. Furthermore, the slope is $-w_1/w_2$, which does not depend on the height b of the contour. This means that all the contours are parallel lines.

- If you think about a three dimensional surface with contours that are parallel lines, it's not hard to convince yourself the surface must be a plane.

[1] We assume that f is continuous.

- Also note that the contour lines are perpendicular to the vector $\mathbf{w}$. E.g., when $b = 0$, points on the contour line have the form $\mathbf{x} = (x_1, -\frac{w_1}{w_2}x_1)^T$. We can check that the inner product between $\mathbf{w}$ and $\mathbf{x}$ is
$$\mathbf{w}^T\mathbf{x} = w_1x_1 - w_2\frac{w_1}{w_2}x_1 = w_1x_1 - w_1x_1 = 0.$$
Visually, the fact that $\mathbf{w}$ is perpendicular to the contours reveals that $\mathbf{w}$ is the "steepest" direction.

- This reasoning all generalizes to higher dimensions, but for our purposes right now it suffices to think about the two-dimensional case.

Slope.

- The slope of a linear function $f(x) = wx$ of one variable is equal to w, which is the "rise" over "run". If we move x units from the origin, the value of $f(x)$ changes from 0 to wx, so
$$\text{slope} = \frac{wx}{x} = w.$$

- How can we generalize this to higher dimensions? We can still measure the amount that the function value increases ("rise") compared with a certain distance moved in input space ("run"), but the issue is that we can now move in many directions in input space. Some directions are steep (cause the function to increase a lot), and some are not. We will define slope as the "rise over run" when moving in the steepest direction, which is the direction of $\mathbf{w}$.

- Lets consider the difference in f when we move from the origin $\mathbf{0} = (0, \ldots, 0)$ to the point $\mathbf{w} \in \mathbb{R}^d$. (Note that $\mathbf{w}$ is the weight vector which defines f, but it is also a point in the input space.) The distance moved is
$$\|\mathbf{w} - \mathbf{0}\| = \|\mathbf{w}\|.$$
The increase in the function value is
$$f(\mathbf{w}) - f(\mathbf{0}) = \mathbf{w}^T\mathbf{w}.$$
The slope is
$$\frac{\mathbf{w}^T\mathbf{w}}{\|\mathbf{w}\|} = \frac{\|\mathbf{w}\|^2}{\|\mathbf{w}\|} = \|\mathbf{w}\|.$$

- To summarize: for a linear function $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$, the slope in the steepest direction is $\|\mathbf{w}\|$, the norm of the weight vector.

2 Support Vector Machine Intuition

- Binary classification problem, training data $\{(\mathbf{x}_i, y_i)\}$

- Assume for now that the the data is *separable*: there is a linear decision boundary that perfectly classifies the training data

- Long history of methods to try to find a *linear separator*, i.e., a weight vector $\mathbf{w}$ such that
$$\begin{aligned} y_i = 1 &\Rightarrow \mathbf{w}^T\mathbf{x}_i \geq 0 \\ y_i = 0 &\Rightarrow \mathbf{w}^T\mathbf{x}_i < 0 \end{aligned}$$

- Rosenblatt's perceptron algorithm (1957). Iterative updates to $\mathbf{w}$. Guaranteed to converge to a linear separator if the data is separable. Updates look like gradient descent.

- Intuition: some separators are better than others. (Illustration)
- **Definition.** *margin* = minimum distance of any correctly classified example to decision boundary.
- **Idea**: find separator with widest margin. Intuition that this will generalize better because it does not go too close to current examples. Minimize the chance the unobserved examples fall on the wrong side of the separator.
- How do we formalize this idea mathematically?

3 Formulating the maximum margin problem

- Goal: find separator $\mathbf{w}$ to maximize the minimum distance from the decision boundary.
- Revisit separator. Visualize in three dimensions.

$$\begin{aligned} y_i = 1 \quad &\Rightarrow \quad \mathbf{w}^T\mathbf{x}_i \geq 0 && (\text{“above sea level”}) \\ y_i = 0 \quad &\Rightarrow \quad \mathbf{w}^T\mathbf{x}_i < 0 && (\text{“below sea level”}) \end{aligned}$$

Margin. First, fix $\mathbf{w}$ and formalize the notion of margin

- Elevation above/below sea level of example i is

$$\begin{aligned} \gamma_i = \gamma_i(\mathbf{w}) &= \begin{cases} \mathbf{w}^T\mathbf{x}_i & y_i = 1 \\ -\mathbf{w}^T\mathbf{x}_i & y_i = 0. \end{cases} \\ &= y_i\mathbf{w}^T\mathbf{x}_i - (1-y_i)\mathbf{w}^T\mathbf{x}_i \end{aligned}$$

- Let $d_i = d_i(\mathbf{w})$ be the horizontal distance $\mathbf{x}_i$ from the decision boundary defined by $\mathbf{w}$. Use slope = $\|\mathbf{w}\|$ = rise over run.

$$\begin{aligned} \|\mathbf{w}\| &= \frac{\gamma_i}{d_i} \\ d_i &= \frac{\gamma_i}{\|\mathbf{w}\|}. \end{aligned}$$

- The margin is the largest value d such that $d_i \geq d$ for all i. (Picture: fix $\mathbf{w}$ and grow the tube around the decision boundary).

Optimization problem.

- Formalize the problem of finding the $\mathbf{w}$ with the best margin as an optimization problem. Include d as a variable to be optimized.
- In words: find the weight vector $\mathbf{w}$ and find the *biggest* value of d such that $d_i(\mathbf{w}) \geq d$ for all i
- In math:

$$\max_{\mathbf{w},d} \; d \tag{5}$$

$$\text{subject to } \; d_i(\mathbf{w}) \geq d \qquad \text{for all } i. \tag{6}$$

Rewriting the problem.

- First, substitute definition for $d_i(\mathbf{w}) = \gamma_i(\mathbf{w})/\|\mathbf{w}\|$.

$$\max_{\mathbf{w},d} \; d \tag{7}$$

$$\text{subject to } \frac{y_i \mathbf{w}^T \mathbf{x}_i - (1-y_i)\mathbf{w}^T \mathbf{x}_i}{\|\mathbf{w}\|} \geq d \qquad \text{for all } i \tag{8}$$

- Bring $\|\mathbf{w}\|$ to right-hand side of constraint

$$\max_{\mathbf{w},d} \; d \tag{9}$$

$$\text{subject to } \; y_i \mathbf{w}^T \mathbf{x}_i - (1-y_i)\mathbf{w}^T \mathbf{x}_i \geq d\|\mathbf{w}\| \qquad \text{for all } i \tag{10}$$

- Make change of variable $\gamma = d\|\mathbf{w}\|$. (Thus, $d = \gamma/\|\mathbf{w}\|$).

$$\max_{\mathbf{w},\gamma} \; \frac{\gamma}{\|\mathbf{w}\|} \tag{11}$$

$$\text{subject to } \; y_i \mathbf{w}^T \mathbf{x}_i - (1-y_i)\mathbf{w}^T \mathbf{x}_i \geq \gamma \qquad \text{for all } i \tag{12}$$

- Make change of variable $\mathbf{v} = \mathbf{w}/\gamma$. (Thus, $\mathbf{w} = \gamma\mathbf{v}$, and $\|\mathbf{w}\| = \gamma\|\mathbf{v}\|$.)

$$\max_{\mathbf{v},\gamma} \; \frac{1}{\|\mathbf{v}\|} \tag{13}$$

$$\text{subject to } \; y_i \gamma \mathbf{v}^T \mathbf{x}_i - (1-y_i)\gamma \mathbf{v}^T \mathbf{x}_i) \geq \gamma \qquad \text{for all } i \tag{14}$$

- Cancel γ in the constraint. (Note that γ disappears entirely!)

$$\max_{\mathbf{v}} \; \frac{1}{\|\mathbf{v}\|} \tag{15}$$

$$\text{subject to } \; y_i \mathbf{v}^T \mathbf{x}_i - (1-y_i)\mathbf{v}^T \mathbf{x}_i \geq 1 \qquad \text{for all } i \tag{16}$$

- Note that maximizing $1/\|\mathbf{v}\|$ is equivalent to minimizing $\|\mathbf{v}\|$.

$$\min_{\mathbf{v}} \; \|\mathbf{v}\| \tag{17}$$

$$\text{subject to } \; y_i \mathbf{v}^T \mathbf{x}_i - (1-y_i)\mathbf{v}^T \mathbf{x}_i \geq 1 \qquad \text{for all } i \tag{18}$$

- Now rename $\mathbf{v}$ back to $\mathbf{w}$ and rewrite the constraints in a more readable form

$$\min_{\mathbf{w}} \; \|\mathbf{w}\| \tag{19}$$

$$\text{subject to } \; \mathbf{w}^T \mathbf{x}_i \geq 1 \qquad y_i = 1 \tag{20}$$

$$\mathbf{w}^T \mathbf{x}_i \leq -1 \qquad y_i = 0. \tag{21}$$

4 Interpretation

- The last version is the widely accepted SVM optimization problem.
- It has a simple geometric interpretation. The constraints specify that we require our linear function $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$ to be at least one for positive examples ($f(\mathbf{x}_i) = \mathbf{w}^T\mathbf{x}_i \geq 1$), and no more than -1 for negatives examples ($f(\mathbf{x}_i) = \mathbf{w}^T\mathbf{x}_i \leq 1$).
- Out of all functions that satisfy this requirement, we are finding the one with the minimum slope $\|\mathbf{w}\|$.
- It is easy to check with some concrete examples that this indeed finds the separator with the biggest margin. (Demo.)

5 Non-Separable Data: Soft Margin

- If the training data is not separable, we need to modify the problem.
- Idea: allow positive example "slack", so it may be s_i units below one:

$$\mathbf{w}^T x_i \geq 1 - s_i$$

Charge cost of s_i in the objective. Do something similar for negative examples.

$$\min_{\mathbf{w},\mathbf{s}} \ \|\mathbf{w}\| + C\sum_{i=1}^{N} s_i \tag{22}$$

$$\text{subject to} \quad \mathbf{w}^T\mathbf{x}_i \geq 1 - s_i \qquad y_i = 1 \tag{23}$$

$$\mathbf{w}^T\mathbf{x}_i \leq -1 + s_i \qquad y_i = 0 \tag{24}$$

$$s_i \geq 0 \qquad i = 1, \ldots, N \tag{25}$$

6 Cost Function and Comparison with Logistic Regression

To be continued...