Hard-margin SVM

Levent Sagun

New York University

February 11, 2016

Problem setup

Given a set of linearly separable training data, how can one find a good separator? What do we expect from a good separator?

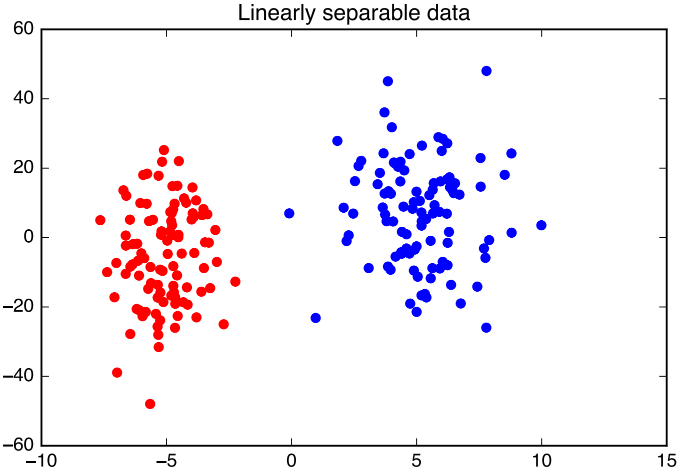
- ... that it actually separates the training points
- ... that it generalizes well

Let $\{x^i, y^i\}_{i=1}^N \in \mathcal{D}$ be the training data, where $x^i \in \mathbb{R}$ and y^i is either $+1$ or -1. What does it mean that the data is linearly separable?

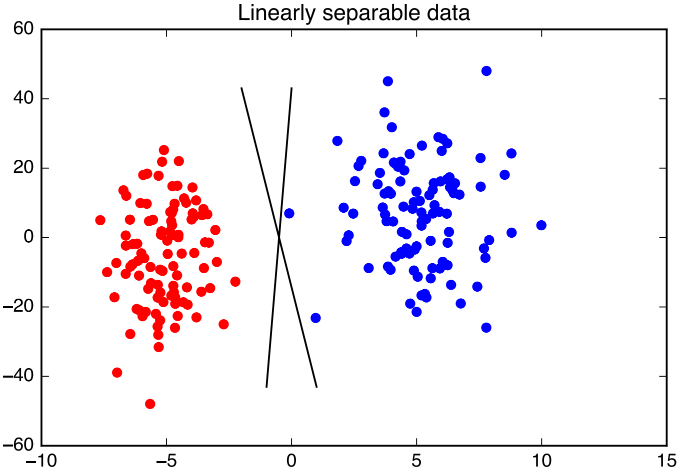
- ... that there is a hyperplane that separates the two clusters
- ... that there is possibly *a lot* of such hyperplanes

How to choose the best one?

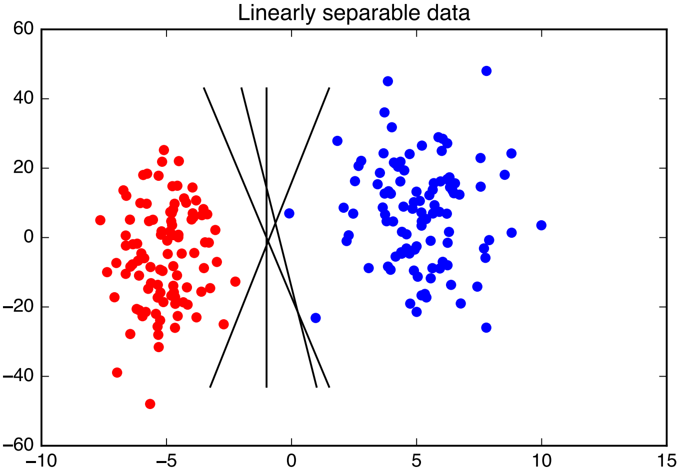
Example



Example



Example



Hyperplane parametrization

Simplest case of real variables, $y = mx + b$ draws a line with slope m that intersects y-axis at the point b:

- Rewrite the above equation: $(m, -1) \cdot (x, y) + b = 0$
- A better notation can be: $(w_1, w_2) \cdot (x_1, x_2) + b = 0$
- $-w_2/w_1 = m$ captures the connection between the two

Hyperplane parametrization

Simplest case of real variables, $y = mx + b$ draws a line with slope m that intersects y-axis at the point b:

- Rewrite the above equation: $(m, -1) \cdot (x, y) + b = 0$
- A better notation can be: $(w_1, w_2) \cdot (x_1, x_2) + b = 0$
- $-w_2/w_1 = m$ captures the connection between the two

Generalize this to higher dimensions, for $w, x \in \mathbb{R}^n$ and $b \in \mathbb{R}$:

- $\ell(x) = w \cdot x + b$ where $L = \{x : \ell(x) = 0\}$ describes a hyperplane.
- w is orthogonal to L (check $w \cdot v = 0$ for v in L)
- What should ℓ assign to the two clusters?

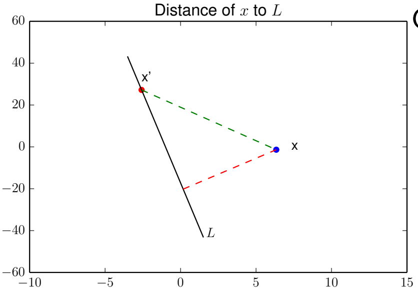
$$\ell(x) \text{ is } \begin{cases} > 0 \text{ if } x \in \text{Blue: } +1 \text{ class} \\ < 0 \text{ if } x \in \text{Red: } -1 \text{ class} \end{cases}$$

- *Note:* $y^i \ell(x^i) > 0$ if $\ell(x) = 0$ separates the data perfectly!

Distance of a point to a line

For a point $x \in \mathbb{R}^n$, how far is x to a given hyperplane L?

- Denote the distance of a point x to L by $d(x, L)$.
- Pick a point on the L, say x', then $d(x, L)$ is the projection of $(x - x')$ onto the normal vector w of L.



Crash course on projections:

- Linear transformations, P, such that $P^2 = P$.
- Unique decomposition into image and kernel of P
- Orthogonal projections: $P = P^T$
- Vector projection: $P_w(v) = \frac{v \cdot w}{||w||^2} w$

Hard-margin SVM

Given two linearly separable clusters, $\mathcal{C}_1$ and $\mathcal{C}_2$, and a hyperplane $L = \{x : \ell(x) = w \cdot x + b = 0\}$ with $||w|| = 1$, suppose $x^{1,L} \in \mathcal{C}_1$ and $x^{2,L} \in \mathcal{C}_2$ are the closest points to L.

- For any i, $y^i \ell(x^i) \geq \min\{d(x^{1,L}, L), d(x^{2,L}, L)\} > 0$
- **GOAL:** Maximize the *margin* around L!
- Since data is linearly separable, the maximizer will be on the set where $d(x^{1,L}, L) = d(x^{2,L}, L)$, let's call this M.
  (note that M depends on data points and the line)

Hard-margin SVM

Given two linearly separable clusters, $\mathcal{C}_1$ and $\mathcal{C}_2$, and a hyperplane $L = \{x : \ell(x) = w \cdot x + b = 0\}$ with $||w|| = 1$, suppose $x^{1,L} \in \mathcal{C}_1$ and $x^{2,L} \in \mathcal{C}_2$ are the closest points to L.

- For any i, $y^i \ell(x^i) \geq \min\{d(x^{1,L}, L), d(x^{2,L}, L)\} > 0$
- **GOAL:** Maximize the *margin* around L!
- Since data is linearly separable, the maximizer will be on the set where $d(x^{1,L}, L) = d(x^{2,L}, L)$, let's call this M.
  (note that M depends on data points and the line)

Procedure:

$$\max\{M : b \in \mathbb{R}, w \in \mathbb{R}^n, ||w|| = 1\} \tag{1}$$

$$\text{subject to } y^i(w \cdot x^i + b) \geq M \tag{2}$$

Equivalent formulation

For any pair of (w, b) we can calculate M and then considering the new pair $(w', b') = (\frac{w}{M}, \frac{b}{M})$ we get $y^i(\frac{w}{M} \cdot x^i + \frac{b}{M}) \geq 1$. Therefore, maximizing M can be rephrased as minimizing $||w'||$.

Equivalent procedure:

$$\min\{||w'|| : b' \in \mathbb{R}, w' \in \mathbb{R}^n\} \tag{3}$$

$$\text{subject to } y^i(w' \cdot x^i + b') \geq 1 \tag{4}$$

- Note that: $||w'|| = ||\frac{w}{M}|| = \frac{||w||}{M} = \frac{1}{M}$
- This is a convex optimization problem: quadratic criterion, linear inequality constraints.
- But, what if the clusters overlap?

Overlapping clusters

For all data points let $t^i > 0$ be the slack variables that represent how wrong the prediction is. We will modify the first formulation first:

Recall the procedure:

$$\max\{M : b \in \mathbb{R}, w \in \mathbb{R}^n, ||w|| = 1\} \tag{5}$$

$$\text{subject to } y^i(w \cdot x^i + b) \geq M \tag{6}$$

Let's modify the second equation to allow each point to have a little more room:

Modified procedure:

$$\max\{M : b \in \mathbb{R}, w \in \mathbb{R}^n, ||w|| = 1\} \tag{7}$$

$$\text{subject to } y^i(w \cdot x^i + b) \geq M(1 - t^i) \tag{8}$$

Overlapping clusters

Now let's find the equivalent version of the modified problem:
Recall the equivalent procedure:

$$\min\{||w'|| : b' \in \mathbb{R}, w' \in \mathbb{R}^n\} \tag{9}$$

$$\text{subject to } y^i(w' \cdot x^i + b') \geq 1 \tag{10}$$

We give a little room for the points to sneak in the margin:
Modified equivalent procedure:

$$\min\{||w'|| : b' \in \mathbb{R}, w' \in \mathbb{R}^n\} \tag{11}$$

$$\text{subject to } y^i(w' \cdot x^i + b') \geq 1 - t^i \tag{12}$$

How much should we allow points to sneak in? Let's put a bound on this: $\sum t^i < C$
Final procedure:

$$\min\{\frac{1}{2}||w||^2 + c\sum t^i\} \tag{13}$$

$$\text{subject to } y^i(w \cdot x^i + b) \geq 1 - t^i \tag{14}$$

Separable vs non-separable

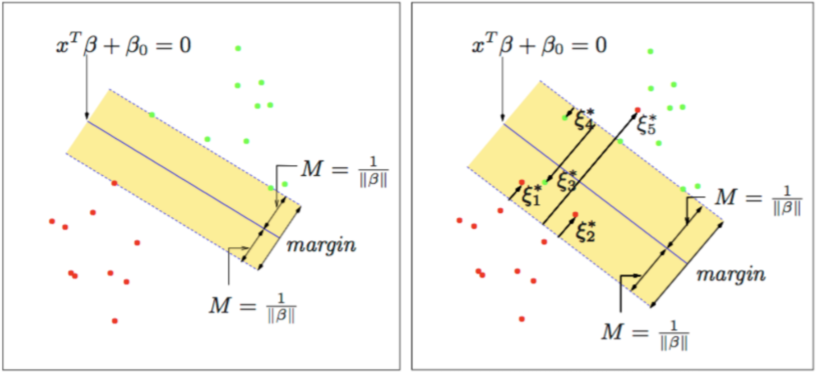


Figure from Hastie's book. Here $\beta = w$ and $\beta_0 = b$.

Exercises

- *Linear regression;* Minimizing sum of squares of errors in $y = X\beta + \epsilon$: Find β such that $||y - X\beta|| = f(\beta)$ is minimized.
- What's the orthogonal projection of y onto the columns of X?
- What's the connection of the two?
- When is $X^T X$ not invertible?
- In the overlapping case, what would happen if you modified the constraint by $y^i(w \cdot x^i - b) \geq M - t^i$