Lecture 21: Support vector classifier

Reading: Sections 9.1-9.2

STATS 202: Data mining and analysis

Sergio Bacallado
November 24, 2013

Hyperplanes and normal vectors

- Consider a p-dimensional space of predictors.
- A **hyperplane** is an affine space which separates the space into two regions.
- The normal vector $\beta = (\beta_1, \ldots, \beta_p)$, is a unit vector $\sum_{j=1}^{p} \beta_j^2 = 1$ which is perpendicular to the hyperplane.
- If the hyperplane goes through the origin, the deviation between a point $(x_1, \ldots, x_p)$ and the hyperplane is the dot product:

$$x \cdot \beta = x_1\beta_1 + \cdots + x_p\beta_p.$$

- The sign of the dot product tells us on which side of the hyperplane the point lies.

Hyperplanes and normal vectors

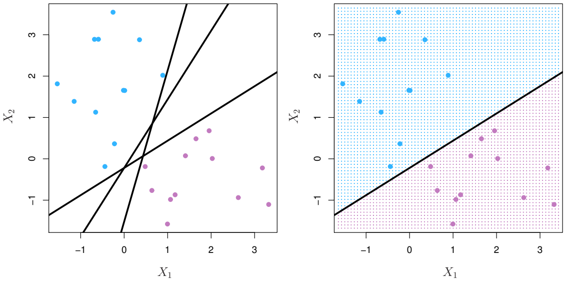
- Consider a p-dimensional space of predictors.
- A **hyperplane** is an affine space which separates the space into two regions.
- The normal vector $\beta = (\beta_1, \ldots, \beta_p)$, is a unit vector $\sum_{j=1}^p \beta_j^2 = 1$ which is perpendicular to the hyperplane.
- If the hyperplane goes through a point $-\beta_0\beta$, i.e. it is displaced from the origin by $-\beta_0$ along the normal vector, the deviation of a point $(x_1, \ldots, x_p)$ from the hyperplane is:

$$\beta_0 + x_1\beta_1 + \cdots + x_p\beta_p.$$

- The sign tells us on which side of the hyperplane the point lies.

Maximal margin classifier

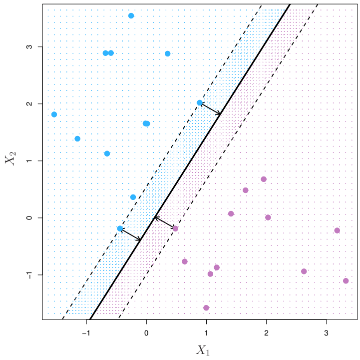
- Suppose we have a classification problem with response $Y = -1$ or $Y = 1$.
- If the classes can be separated, most likely, there will be an infinite number of hyperplanes separating the classes.



Maximal margin classifier

Idea:

- Draw the largest possible empty margin around the hyperplane.
- Out of all possible hyperplanes that separates the 2 classes, choose the one with the widest margin.



Maximal margin classifier

This can be written as an optimization problem:

$$\max_{\beta_0,\beta_1,\ldots,\beta_p} \quad M$$

$$\text{subject to } \sum_{j=1}^{p} \beta_j^2 = 1,$$

$$\underbrace{y_i(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})}_{\text{How far is } x_i \text{ from the hyperplane}} \geq M \quad \text{for all } i = 1, \ldots, n.$$

M is simply the width of the margin in either direction.

Finding the maximal margin classifier

We can reformulate the problem by defining a vector $w = (w_1, \ldots, w_p) = \beta/M$:

$$\min_{\beta_0, w} \quad \frac{1}{2}\|w\|^2$$

subject to

$$y_i(\beta_0 + w \cdot x_i) \geq 1 \qquad \text{for all } i = 1, \ldots, n.$$

This is a quadratic optimization problem.

Finding the maximal margin classifier

$$\min_{\beta_0, w} \ \frac{1}{2}\|w\|^2$$

subject to

$$y_i(\beta_0 + w \cdot x_i) \geq 1 \quad \text{for all } i = 1, \ldots, n.$$

Introducing Lagrange multipliers, $\alpha_1, \ldots, \alpha_n$, this is equivalent to:

$$\max_{\alpha} \ \min_{\beta_0, w} \ \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \alpha_i [y_i(\beta_0 + w \cdot x_i) - 1]$$

subject to $\alpha_i \geq 0$.

Finding the maximal margin classifier

$$\max_{\alpha} \min_{\beta_0, w} \; \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \alpha_i [y_i(\beta_0 + w \cdot x_i) - 1]$$

subject to $\alpha_i \geq 0$.

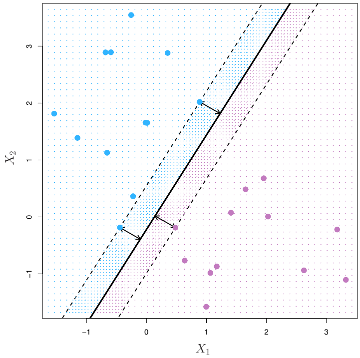
Setting the derivatives with respect to w to 0, we obtain that the solution is of the form:

$$\hat{w} = \sum_{i=1}^{n} \alpha_i y_i x_i$$

Furthermore, $\alpha_i > 0$ if and only if $y_i(\beta_0 + w \cdot x_i) = 1$, that is, if x_i falls on the margin.

Support vectors

The vectors that fall on the margin and determine the solution are called **support vectors**:



Finding the maximal margin classifier

$$\max_{\alpha} \min_{\beta_0, w} \quad \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \alpha_i [y_i(\beta_0 + w \cdot x_i) - 1]$$

$$\text{subject to} \quad \alpha_i \geq 0.$$

The solution is $\hat{w} = \sum_{i=1}^{n} \alpha_i y_i x_i$, so we can plug this in above and simplify the problem of finding the optimal α_i:

$$\max_{\alpha} \quad \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{i'=1}^{n} \alpha_i \alpha_{i'} y_i y_{i'} (x_i \cdot x_{i'})$$

$$\text{subject to} \quad \alpha_i \geq 0, \quad \sum_{i} \alpha_i y_i = 0.$$

Summary

We've reduced the problem of finding w, which describes the hyperplane and the size of the margin, to finding a set of coefficients $\alpha_1, \ldots, \alpha_n$ through:

$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{i'=1}^{n} \alpha_i \alpha_{i'} y_i y_{i'} (x_i \cdot x_{i'})$$

$$\text{subject to} \quad \alpha_i \geq 0, \quad \sum_i \alpha_i y_i = 0.$$

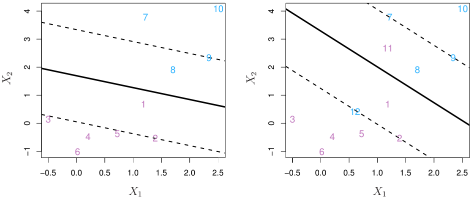
This only depends on the training sample inputs through the inner products $x_i \cdot x_j$ for every pair i, j.

Support vector classifier

Problem: It is not always possible to separate the points using a hyperplane.

Support vector classifier:

- Relaxation of the maximal margin classifier.
- Allows a number of points points to be on the wrong side of the margin or the margin or even the hyperplane.



Support vector classifier

This can be written as an optimization problem:

$$\max_{\beta_0,\beta,\epsilon} \quad M$$

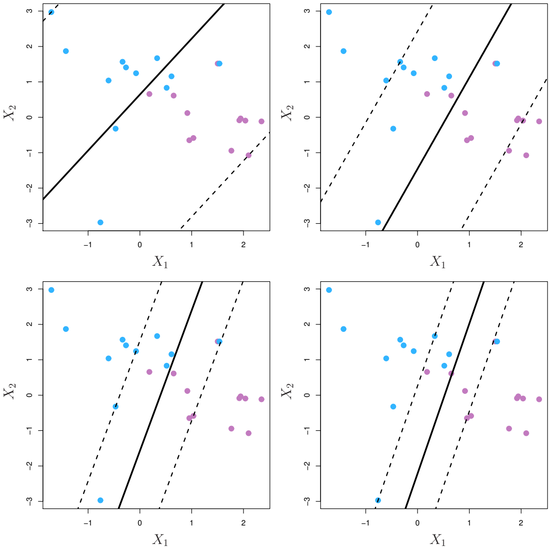
$$\text{subject to } \sum_{j=1}^{p} \beta_j^2 = 1,$$

$$\underbrace{y_i(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})}_{\text{How far is } x_i \text{ from the hyperplane}} \geq M(1 - \epsilon_i) \qquad \text{for all } i = 1, \ldots, n$$

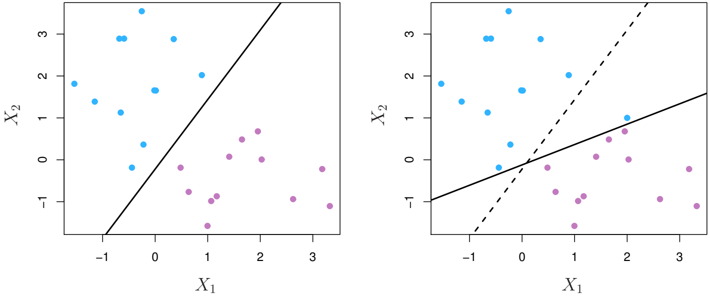
$$\epsilon_i \geq 0 \text{ for all } i = 1, \ldots, n, \quad \sum_{i=1}^{n} \epsilon_i \leq C.$$

M is the width of the margin in either direction.
$\epsilon = (\epsilon_1, \ldots, \epsilon_n)$ are called *slack* variables.
C is called the *budget*.

Tuning the budget, C (high to low)



If the budget is too low, we tend to overfit



Maximal margin classifier, $C = 0$. Adding one observation dramatically changes the classifier.

Finding the support vector classifier

We can reformulate the problem by defining a vector $w = (w_1, \ldots, w_p) = \beta/M$:

$$\min_{\beta_0, w, \epsilon} \quad \frac{1}{2}\|w\|^2 + D\sum_{i=1}^{n} \epsilon_i$$

subject to

$$y_i(\beta_0 + w \cdot x_i) \geq (1 - \epsilon_i) \quad \text{for all } i = 1, \ldots, n,$$

$$\epsilon_i \geq 0 \quad \text{for all } i = 1, \ldots, n.$$

The penalty $D \geq 0$ serves a function similar to the budget C, but is inversely related to it.

Finding the support vector classifier

$$\min_{\beta_0, w, \epsilon} \quad \frac{1}{2}\|w\|^2 + D\sum_{i=1}^{n}\epsilon_i$$

subject to

$y_i(\beta_0 + w \cdot x_i) \geq (1 - \epsilon_i)$ for all $i = 1, \ldots, n.$

$\epsilon_i \geq 0$ for all $i = 1, \ldots, n.$

Introducing Lagrange multipliers, α_i and μ_i, this is equivalent to:

$$\max_{\alpha, \mu} \; \min_{\beta_0, w, \epsilon} \quad \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n}\alpha_i[y_i(\beta_0 + w \cdot x_i) - 1 + \epsilon_i] + \sum_{i=1}^{n}(D - \mu_i)\epsilon_i$$

subject to $\alpha_i \geq 0, \mu_i \geq 0,$ for all $i = 1, \ldots, n.$

Finding the support vector classifier

$$\max_{\alpha,\mu} \min_{\beta_0,w,\epsilon} \quad \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \alpha_i[y_i(\beta_0 + w \cdot x_i) - 1 + \epsilon_i] + \sum_{i=1}^{n}(D - \mu_i)\epsilon_i$$

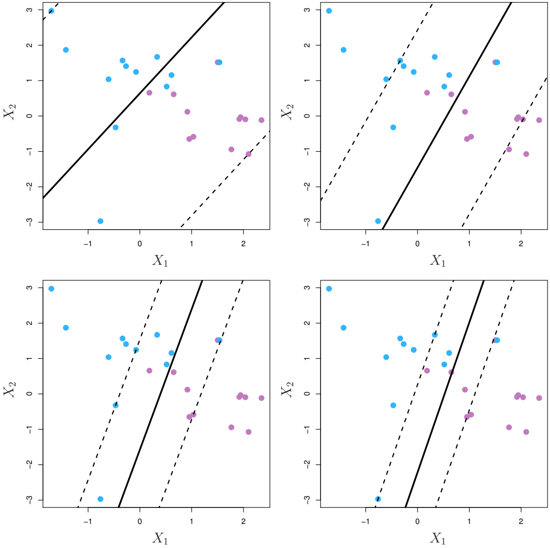
subject to $\alpha_i \geq 0, \mu_i \geq 0,$ for all $i = 1, \ldots, n.$

Setting the derivatives with respect to w to 0, we obtain that the solution is of the form:

$$\hat{w} = \sum_{i=1}^{n} \alpha_i y_i x_i$$

Furthermore, $\alpha_i > 0$ if and only if $y_i(\beta_0 + w \cdot x_i) \leq 1$, that is, if x_i falls on the wrong side of the margin.

Support vectors



The problem only depends on $x_i \cdot x_{i'}$

As with the Maximal Margin Classifier, the problem can be reduced to finding $\alpha_1, \ldots, \alpha_n$:

$$\max_{\alpha} \quad \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{i'=1}^{n} \alpha_i \alpha_{i'} y_i y_{i'} (x_i \cdot x_{i'})$$

$$\text{subject to} \quad 0 \leq \alpha_i \leq D \ \text{ for all } i = 1, \ldots, n,$$

$$\sum_i \alpha_i y_i = 0.$$

As before, this only depends on the training sample inputs through the inner products $x_i \cdot x_j$ for every pair i, j.