Paul Martins

Support Vector Machines

Notes from ISLR book

Contents

1. Maximum Margin Classifier

SVMs rely on the separating the predictor space with hyperplanes and assign a response value based on which side of the hyperplane a given observation is.

In the binary classification case with $y_i \in \{-1, 1\}$, a separating hyperplane takes the form

$$y_i(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}) > 0 \tag{1}$$

For a new observation x^*, the sign of $f(x^*) = \beta_0 + \beta_1 x_1^* + \cdots + \beta_p x_p^*$ would determine the value of y^*. In the case where the data are completely separable, there exits several of such planes as in the left panel of Figure 2.

A smart way of selecting a plane among all candidates is to take the one which maximise the margin M, minimal distance from the observation to the hyperplane. The bigger the margin, the more confident we are about the classifier predictions.

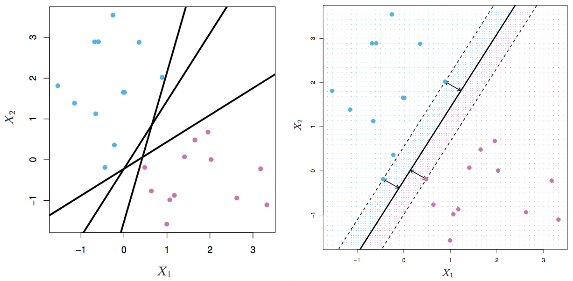

Figure 1. Illustrations of the maximal margin classifier.

The parameters $\beta_{0,\ldots,p}$ and M are found by solving the optimisation problem

Maximize M subject to
$\beta_0, \beta_1, \ldots, \beta_p, M$

- $\sum_{j=1}^{p} \beta_j^2 = 1,$
- $y_i(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}) \geq M$

The obs that define the margin are called support vectors. They are the only observations needed to build the model. The addition of an extra observation on the correct side of the margin would have no effect on the model, but if it is on the wrong side, the hyperplane would need updating.

The main drawback of this model is that most of the time, this problem has no solution: there is no plane that can clearly separate all observations.

2. Support Vector Classifier

To extend the maximum margin classifier, the support vector classifier allows some obs to be on the wrong side of the margin. This brings greater robustness and better classification on most training observations.

Maximize M subject to
$\beta_0, \beta_1, \ldots, \beta_p, \varepsilon_1, \ldots, \varepsilon_p, M$

- $\sum_{j=1}^{p} \beta_j^2 = 1,$
- $y_i(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}) \geq M(1 - \varepsilon_i),$
- $\varepsilon_i \geq 0,$
- $\sum_{i=1}^{n} \varepsilon_i \leq C$ with $C \geq 0$

The slack variable ε_i defines the position of the ith observation (Table 1). Observations with $\varepsilon_i > 0$ are called support vectors.

$\varepsilon_i = 0$	correct side of margin
$0 < \varepsilon_i < 1$	wrong side of margin (violation)
$\varepsilon_i = 1$	on hyperplane
$\varepsilon_i > 1$	wrong side of hyperplane

Table 1. Meaning of slack variables

The tuning parameter C controls the severity of the violations. As C increases, more violations are allowed, hence M increases too. The model has a lower variance but higher bias.

2.1 Link to regression

The above optimisation algorithm can be rewritten with the hinge loss.

Minimize
$\beta_0, \beta_1, \ldots, \beta_p$

$$\underbrace{\sum_{i=1}^{n} \max[0, 1 - y_i(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})]}_{\text{Hinge Loss}} + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{2}$$

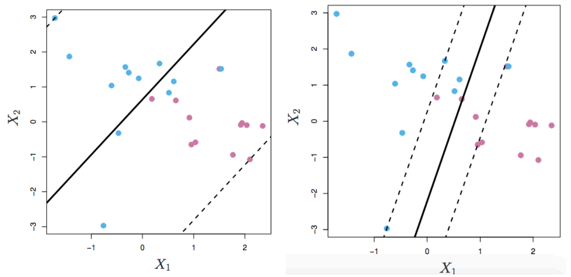

Figure 2. Illustrations of the support vector classifier with 2 values of C: left: large, right: small.

The use of the ridge penalty means that large λ lead to small β_j, hence more violations to the margin are tolerated, thus C should be large too.

Observations with $y_i f(x_i) \geq 1$ are on the correct side of the margin and have a loss of 0, so they do not contribute to improving the model.

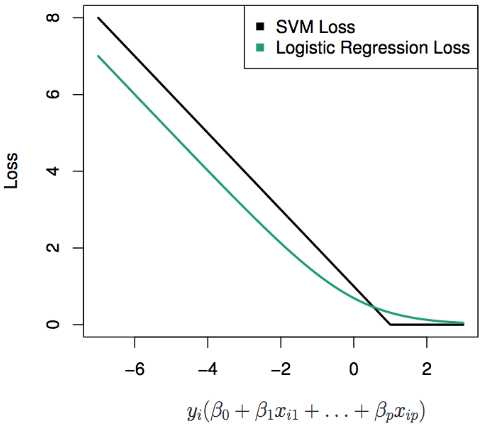

Figure 3. Comparison of Hinge loss and log reg. loss

3. Support Vector Machines

In many cases, we cannot separate the obs with linear borders. To extend the support vector classifier to non linear boundaries, the notion of kernel is introduced.

3.1 Definition

Kernels are functions that quantify the similarity of 2 observations. It can be shown that the solution to the linear support classifier (ie the β_j and M) only depends on the inner product of the observations

$$K(x_i, x_{i'}) = \langle x_i, x_{i'} \rangle = \sum_{j=1}^{p} x_{ij} x_{i'j} \tag{3}$$

Here K is the linear kernel. The support vector classifier can be represented as

$$f(x) = \beta_0 + \sum_{i=i}^{n} \alpha_i K(x, x_i) \tag{4}$$

but α_i is non-zero only if the observations are support vectors, hence

$$f(x) = \beta_0 + \sum_{i \in \mathscr{S}} \alpha_i K(x, x_i) \tag{5}$$

Expanding each kernel gives the relation between the α_i and the β_j.

3.2 Kernels

To expend to non linear boundaries, we just need to use another kernel definition, like polynomial of degree d

$$K(x_i, x_{i'}) = \left(1 + \sum_{j=1}^{p} x_{ij} x_{i'j}\right)^d \tag{6}$$

or a radial kernel as shown in Figure 4.

$$K(x_i, x_{i'}) = \exp\left(-\gamma \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2\right) \tag{7}$$

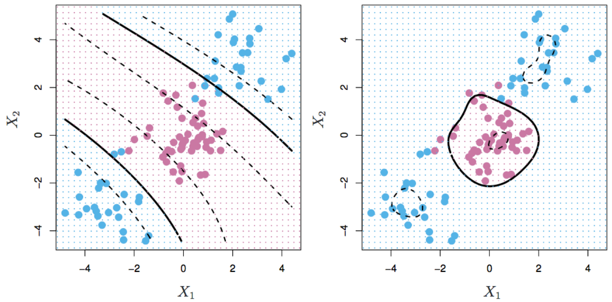

Figure 4. SVM with polynomial kernel of degree $d = 3$ (left) and radial kernel (right).

SVMs are usually for binary classification. The most popular extensions to $K > 2$ classes are one-vs-one classification where we construct $\binom{K}{2}$ SVMs each of of which compares a pair of classes, or one-vs-all classification where we fit K SVMs comparing the kth class to the others.