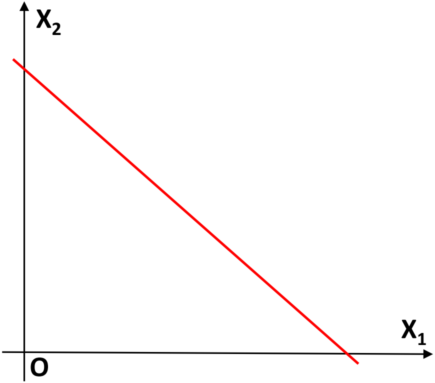
Classification: Support vector machine

Rui Zhu

Overview

Hyperplane

Hyperplane: In a p-dimensional space, a hyperplane is a flat affine subspace of hyperplane dimension $p-1$.



Hyperplane

A hyperplane in a p-dimensional space is defined as

$$\mathbf{w}^T\mathbf{x} + b = 0,$$

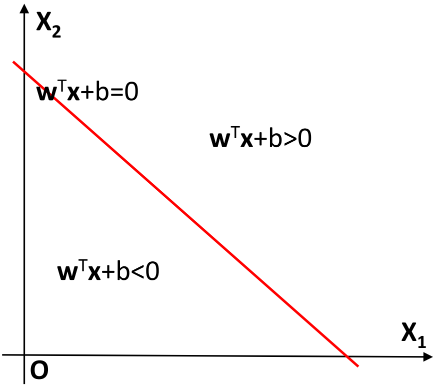
where

- $\mathbf{x} \in \mathbb{R}^{p\times 1}$ is the point on the hyperplane
- $\mathbf{w} \in \mathbb{R}^{p\times 1}$ is the normal vector of the hyperplane
- b is the bias of the hyperplane

Hyperplane

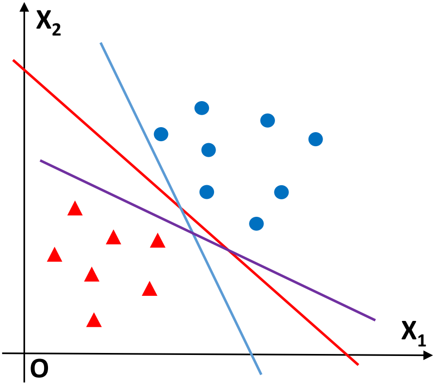
A hyperplane can divide the feature space to two halves:

- $\mathbf{w}^T\mathbf{x} + b > 0$
- $\mathbf{w}^T\mathbf{x} + b < 0$



Separating hyperplane

If we have two classes that can be separated by a linear boundary, then we can find an infinite number of separating hyperplanes that separates the training observations perfectly according to their class labels.



Separating hyperplane

Classification based on the separating hyperplane

Suppose the training instances are $\mathbf{X} \in \mathbb{R}^{N \times p}$ and the class labels are coded as 1 and -1.

Separating hyperplane should satisfy the following conditions:

- $\mathbf{w}^T\mathbf{x}_i + b > 0$ if $y_i = 1$
- $\mathbf{w}^T\mathbf{x}_i + b < 0$ if $y_i = -1$

or

$$y_i(\mathbf{w}^T\mathbf{x}_i + b) > 0, \quad i = 1, 2, \ldots, N$$

For a test instance $\mathbf{x}_t$, we classify it to

- class 1 if $\mathbf{w}^T\mathbf{x}_t + b > 0$
- class -1 if $\mathbf{w}^T\mathbf{x}_t + b < 0$

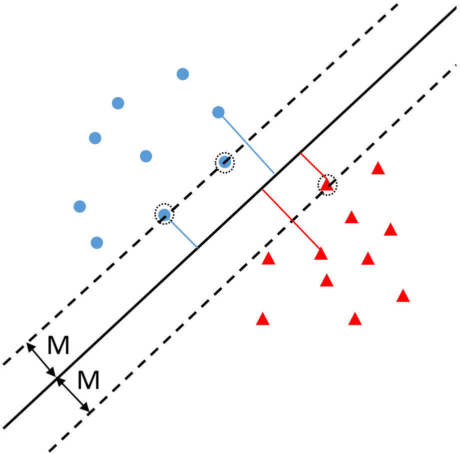
Maximal margin classifier

Margin: the minimal distance from the training observations to the hyperplane.

Maximal margin hyperplane: the separating hyperplane for which the margin is largest.

We can then classify a test observation based on which side of the maximal margin hyperplane it lies.

Maximal margin classifier



Maximal margin classifier

The maximal margin hyperplane depends directly on the support vectors, not on the other observations.

The maximal margin hyperplane depends directly on only a small subset of the observations.

Maximal margin classifier

The distance from a point $\mathbf{x}_i$ to a hyperplane $\mathbf{w}^T\mathbf{x} + b = 0$:

$$y_i(\mathbf{w}^T\mathbf{x}_i + b),$$

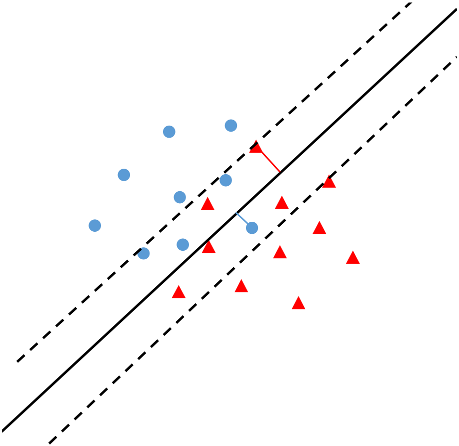
when $||\mathbf{w}||_2 = 1$.

Maximal margin classifier

$$
\begin{aligned}
\max_{M,\mathbf{w},b} \quad & M \\
\text{s.t.} \quad & ||\mathbf{w}||_2 = 1, \\
& y_i(\mathbf{w}^T\mathbf{x}_i + b) \geqslant M, \forall i.
\end{aligned}
$$

- Maximise M, subject to two constraints, $||\mathbf{w}||_2 = 1$ and $y_i(\mathbf{w}^T\mathbf{x}_i + b) \geqslant M$, for all $i = 1, 2, \ldots, N$
- Variables to be solved: $M, \mathbf{w}, b$

Maximal margin classifier



Maximal margin classifier

Maximal margin classifier only works for two classes that are separable.

For non-separable case, there's no solution with margin larger than 0, $M > 0$.

A classifier based on a separating hyperplane will necessarily perfectly classify all of the training observations; this can lead to sensitivity to individual observations, also overfitting.

What to do? Use soft margin.

The generalization of the maximal margin classifier to the non-separable case is known as the support vector classifier.

Support vector classifier

We consider a classifier based on a hyperplane that does not perfectly separate the two classes:

- Greater robustness to individual observations, and
- Better classification of most of the training observations.

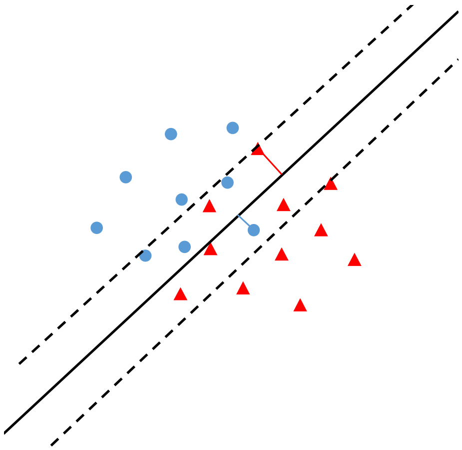
It could be worthwhile to misclassify a few training observations in order to do a better job in classifying the remaining observations.

Support vector classifier

Support vector classifier (Soft margin classifier):

- We allow some observations to be on the incorrect side of the margin, or even the incorrect side of the hyperplane.
- The margin is soft because it can be violated by some of the training observations.
- Observations on the wrong side of the hyperplane correspond to training observations that are misclassified by the support vector classifier.

Support vector classifier



Support vector classifier

$$
\begin{aligned}
\max_{M,\mathbf{w},\xi,b} \quad & M \\
\text{s.t.} \quad & ||\mathbf{w}||_2 = 1, \\
& y_i(\mathbf{w}^T\mathbf{x}_i + b) \geqslant M(1-\xi_i), \\
& \xi_i \geqslant 0, \sum_{i=1}^{N} \xi_i \leqslant C, \quad \forall i.
\end{aligned}
$$

Two new symbols:

- ξ_i: additional variables in the optimisation problem
- C: nonnegative tuning parameter

Support vector classifier

$$y_i(\mathbf{w}^T\mathbf{x}_i + b) \geqslant M(1 - \xi_i)$$

- $\boldsymbol{\xi} \in \mathbb{R}^{N\times 1} = (\xi_1, \xi_2, \ldots, \xi_N)^T$: slack variables that allow individual observations to be on the wrong side of the margin or the hyperplane
- ξ_i tells us where the ith observation is located, relative to the hyperplane and relative to the margin
- If $\xi_i > 0$ then the ith observation is on the wrong side of the margin, and we say that the ith observation has violated the margin
- If $\xi_i > 1$ then it is on the wrong side of the hyperplane

Support vector classifier

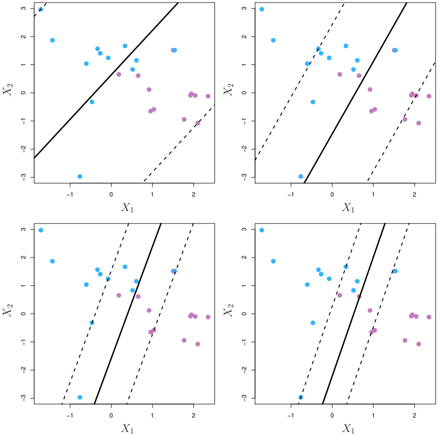
$$\sum_{i=1}^{N} \xi_i \leqslant C$$

- C bounds the sum of the ξ_i's, and so it determines the number and severity of the violations to the margin (and to the hyperplane) that we will tolerate
- If $C = 0$: the maximal margin hyperplane optimization problem
- If $C > 0$: no more than C observations can be on the wrong side of the hyperplane
- C controls the bias-variance trade-off: When C is small, low bias but high variance; When C is large, more biased but may have lower variance

Support vector classifier

- Only observations that either lie on the margin or that violate the margin will affect the hyperplane
- An observation that lies strictly on the correct side of the margin does not affect the support vector classifier
- Support vectors: Observations that lie directly on the margin, or on the wrong side of the margin for their class
- It is quite robust to the behaviour of observations that are far away from the hyperplane

Support vector classifier



1

[1]Some of the figures in this presentation are taken from "An Introduction to Statistical Learning, with applications in R" (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani

Support vector classifier

When C is large: the margin is wide, many observations violate the margin, and so there are many support vectors. Many observations are involved in determining the hyperplane.

Support vector classifier

Now we see how to generalise from separable case to non-separable case, however, both classifiers have linear classification boundary.

What if we need a nonlinear classification boundary?

Enlarge the feature space: make p larger

Support vector machine (SVM)

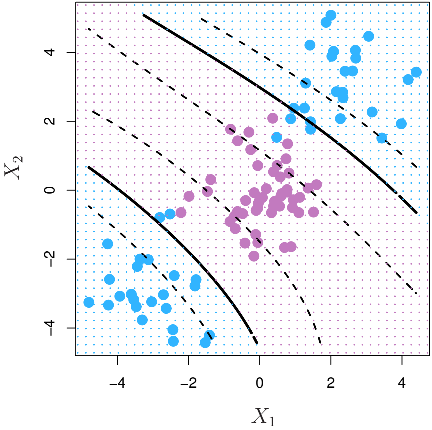
- Enlarge the feature space by adding transformed features: X_1^2, X_2^2, X_1X_2, etc.
- Train the support vector classifier in the enlarged feature space
- Nonlinear classification boundary in the original feature space

Example: From (X_1, X_2) to (X_1, X_2, X_1^2, X_2^2):

$$b + w_1X_1 + w_2X_2 = 0$$

$$b + w_1X_1 + w_2X_2 + w_3X_1^2 + w_4X_2^2 = 0$$

Support vector machine (SVM)



$$b + w_1X_1 + w_2X_2 + w_3X_1^2 + w_4X_2^2 + w_5X_1X_2 + w_6X_1^3 + w_7X_2^3 + w_8X_1^2X_2 + w_9X_1X_2^2 = 0$$

Support vector machine (SVM)

- The computational cost is too high when we have a lot of transformed variables.
- There's a more elegant way by using kernels, which is inspired by the solution of the support vector classifier.

Support vector machine (SVM)

- Inner product:

$$\langle \mathbf{x}_i, \mathbf{x}_{i'} \rangle = \sum_{j=1}^{p} x_{ij} x_{i'j}$$

- The linear support classifier:

$$f(\mathbf{x}) = b + \sum_{i=1}^{N} \alpha_i \langle \mathbf{x}, \mathbf{x}_i \rangle$$

- α_i is nonzero only for the support vectors in the solution:

$$f(\mathbf{x}) = b + \sum_{i \in \mathbb{S}} \alpha_i \langle \mathbf{x}, \mathbf{x}_i \rangle$$

- All we need are inner products!

Support vector machine (SVM)

- A generalisation of the inner product: kernel function

$$K(\mathbf{x}_i, \mathbf{x}_{i'})$$

- Kernel measures the similarity between observations
-

$$f(\mathbf{x}) = b + \sum_{i \in \mathbb{S}} \alpha_i K(\mathbf{x}, \mathbf{x}_i)$$

Support vector machine (SVM)

- Linear kernel:

$$K(\mathbf{x}_i, \mathbf{x}_{i'}) = \langle \mathbf{x}_i, \mathbf{x}_{i'} \rangle = \sum_{j=1}^{p} x_{ij} x_{i'j}$$

  Quantifies the similarity of a pair of observations using Pearson correlation.

- Polynomial kernel:

$$K(\mathbf{x}_i, \mathbf{x}_{i'}) = (1 + \langle \mathbf{x}_i, \mathbf{x}_{i'} \rangle)^d$$

  Fit a support vector classifier in a higher-dimensional space involving polynomials of degree d.
  Tuning parameter: d

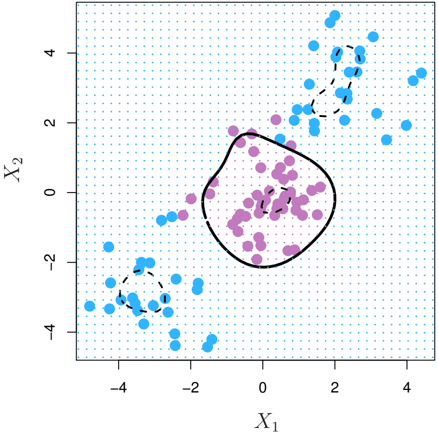
Support vector machine (SVM)

- Radial kernel (radial basis function, RBF)

$$K(\mathbf{x}_i, \mathbf{x}_{i'}) = \exp\left\{-\frac{(\mathbf{x}_i - \mathbf{x}_{i'})^T(\mathbf{x}_i - \mathbf{x}_{i'})}{\gamma}\right\},$$

  $\gamma > 0$ (tuning parameter).
- As γ increases and the fit becomes more non-linear.
- The radial kernel has very local behaviour: only nearby training observations have an effect on the class label of a test observation.
- For the radial kernel, the feature space is implicit and infinite-dimensional. However, we don't have to know about this infinite-dimensional space, because we just need the kernels.

Support vector machine (SVM)



Support vector machine (SVM)

- Support vector classifier with $C = 0$: maximal margin classifier
- SVM with polynomial kernel $d = 1$ or linear kernel: support vector classifier

Support vector machine (SVM)

Multi-class:

- One versus All: Fit C different binary SVM classifiers $\hat{f}_k(\mathbf{x})$, $k = 1, 2, \ldots, C$; each class versus the rest. Classify $\mathbf{x}$ to the class for which $\hat{f}_k(\mathbf{x})$ is the largest.
- One versus One: Fit all $\binom{C}{2}$ pairwise classifiers. Classify $\mathbf{x}$ to the class that wins the most pairwise competitions.
- When C is not too large, use One versus One.

Relationship between SVM and logistic regression

Another formulation of SVM:

$$\min_{\mathbf{w},b} \sum_{i=1}^{N} [1 - y_i f(\mathbf{x}_i)]_+ + \lambda ||\mathbf{w}||_2^2,$$

where $\lambda > 0$.

- $[\ \ ]_+$: hinge loss
- $\lambda ||\mathbf{w}||_2^2$: ridge penalty
- When λ is large: elements in $\mathbf{w}$ are small, more violations to the margin are tolerated, and a low-variance but high-bias classifier will result
- When λ is small: few violations to the margin will occur; this amounts to a high-variance but low-bias classifier.

Relationship between SVM and logistic regression

Loss+Penalty:

$$\min_{\beta} L(\mathbf{X}, \mathbf{y}, \boldsymbol{\beta}) + \lambda P(\boldsymbol{\beta})$$

Relationship between SVM and logistic regression