SUPPORT VECTOR CLASSIFIER (ch 9)

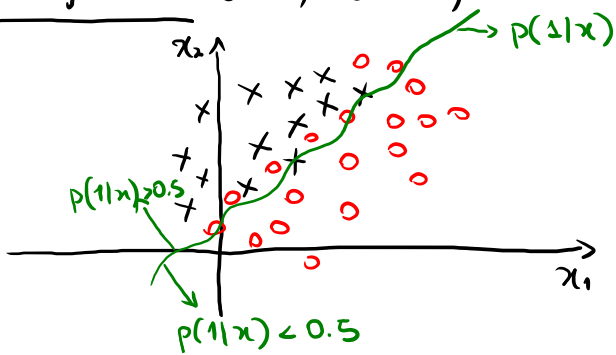
19/4/21

Up to now (logistic regression, discriminant analysis, decision trees, K-nn):

Zero-One Loss → Bayes classifier ($p(c|x)$) → different methods provide different estimates to $p(c|x)$ → then

Assign x to class c with largest ($t=0.5$) $\hat{p}(c|x)$

Geometric view to classification (2D, 2 classes)



$p(1|x)=0.5$ (decision boundary under Bayes classifier)

LR

$$\hat{p}(1|\underline{x}) = \frac{e^{\hat{\beta}^t \underline{x}}}{1+e^{\hat{\beta}^t \underline{x}}}$$

$\hat{p}(1|x) \geq 0.5 \iff \hat{\beta}^t \underline{x} \geq 0$

$\hat{\beta}^t \underline{x} \gg 0 \rightarrow \hat{p}(1|x) \simeq 1$

$\hat{\beta}^t \underline{x} \ll 0 \rightarrow \hat{p}(1|x) \simeq 0$

The further one goes to the decision surface, the more confident one is about the classification

Support vector machines take the geometric view forward and look at the function $g(\underline{x})$ that best separates the two classes. So they focus directly on the decision surface. (1990s, Vapnik)

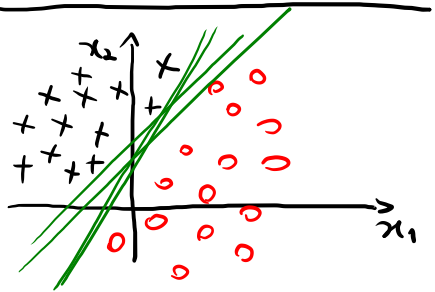
There are different forms of SVMs:

① Earliest and simplest form: maximal margin classifier
→ linearly separable classes

② Extension to non-separable case:
→ linear decision surfaces

③ Extension to support vector machines
→ non-linear class boundaries

MAXIMAL MARGIN CLASSIFIER



Problem: Given two linearly separable classes, find the best "separating hyperplane"

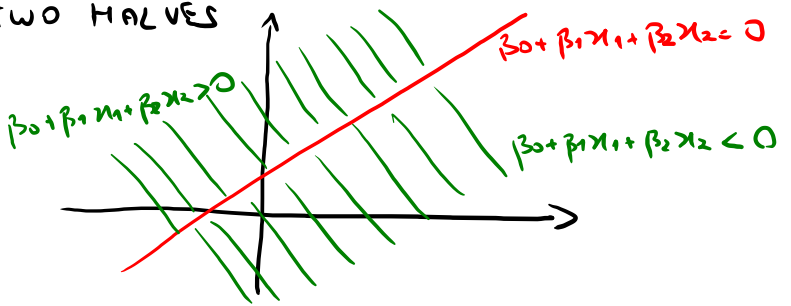
Def: In a p-dimensional space, a hyperplane is a flat affine subspace of dimension $p-1$

$p=2 \rightarrow$ line $\rightarrow \quad \beta_0 + \beta_1 x_1 + \beta_2 x_2 = 0$

$p=3 \rightarrow$ plane $\rightarrow \quad \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 = 0$

$p > 3 \rightarrow$ hyperplane $\rightarrow \quad \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p = 0 \iff \beta_0 + \beta^t \underline{x} = 0$

ONE HYPERPLANE $\rightarrow$ TWO HALVES



For most points:

$$\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p > 0 \quad \text{or} \quad \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p < 0$$

CLASSIFICATION VIA A HYPERPLANE

Data matrix : $X_{n \times p}$ matrix of predictors

Response Observations fall into class -1 or 1

$$y_1, \dots, y_n \in \{-1, 1\}$$

Suppose that it is possible to linearly separate the two classes. Then

$$\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} \geq 0 \quad \text{if } y_i = 1$$

$$\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} \leq 0 \quad \text{if } y_i = -1$$

So a separating hyperplane satisfies:

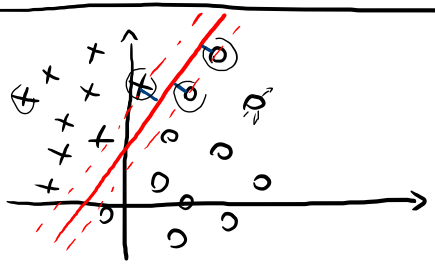
$$\boxed{y_i \left(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} \right) \geq 0 \quad i = 1, \dots, n}$$

This is the reason for denoting the classes as -1 and 1

Then given a new x^*, classed x^* based on sign of

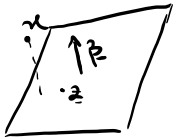
$$\beta_0 + \beta_1 x_1^* + \beta_2 x_2^* + \cdots + \beta_p x_p^*$$

CONSTRUCTING A SEPARATING HYPERPLANE



OPTIMAL SEPARATING HYPERPLANE:
The one that maximises the margin (minimal distance of the points to the hyperplane)

How can we write this into an algorithm?



Without loss of generality, given a hyperplane $\beta_0 + \beta^T x = 0$ choose the representation such that $\|\beta\|_2 = \beta_1^2 + \beta_2^2 + \dots + \beta_p^2 = 1$ (orthonormal vector)

The distance of x to the hyperplane is the orthogonal projection of $x - z$ on β:

$$\text{dist}(x, H) = |\beta^T (x - z)| = |\beta^T x - \beta^T z| = |\beta^T x + \beta_0 - \underbrace{\beta_0 - \beta^T z}_{0}| = |\beta^T x + \beta_0|$$

OPTIMAL HYPERPLANE

$$(\hat{\beta}_0, \hat{\beta}) = \underset{\substack{\beta_0, \beta \\ \|\beta\|=1}}{\text{argmax}} \Big(\underbrace{\min_{i=1,\dots,n} \underbrace{y_i(\beta_0 + \beta^T x_i)}_{\text{distance of } x_i \text{ to hyperplane}}}_{\text{margin}} \Big)$$

Equivalently:

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

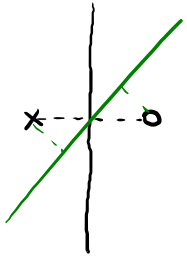
s.t. $\beta_1^2 + \ldots + \beta_p^2 = 1$

$y_i(\beta_0 + \beta^t x_i) \geq M \quad i = 1, \ldots, n$

MAXIMAL MARGIN HYPERPLANE

Some remarks:

① There is at least one observation on either side of the margin.

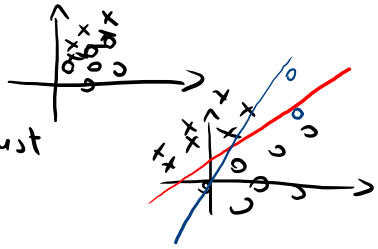


② The hyperplane is completely characterized by the closest points, which are support vectors.

SUPPORT VECTOR CLASSIFIER

An extension for:

1. non-separable case
2. separable case but more robust



Idea Hard margin → Soft margin

Look for maximal margin classifier, but allowing for misclassifications

As an algorithm:

SOFT MARGIN CLASSIFIER

$$\max_{\substack{\beta_0,\dots,\beta_p \\ \varepsilon_1,\dots,\varepsilon_n}} M$$

s.t. $\beta_1^2 + \beta_2^2 + \dots + \beta_p^2 = 1$

slack variables

$$y_i(\beta_0 + \underline{\beta}^T \underline{x}_i) \geq M(1-\varepsilon_i), \quad i = 1,\dots,n$$

$\varepsilon_i \geq 0$ softening the margin $(M(1-\varepsilon_i) \leq M)$

$\sum_{i=1}^{n} \varepsilon_i \leq C, \quad C \geq 0$

C: slack budget

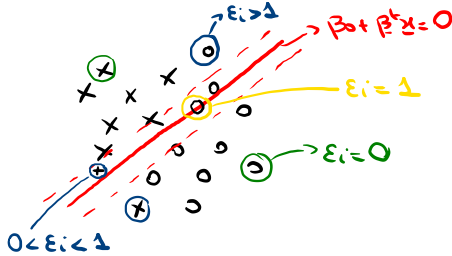
Fix C → Find $\hat{\beta}, \hat{\varepsilon}$ that maximises M → classify according to $\hat{\beta}$, like before

$\underline{x}^* \rightarrow \beta_0 + \underline{\beta}^T \underline{x}^* \begin{cases} \geq 0 \rightarrow \text{classify to } 1 \\ < 0 \rightarrow \text{classify to } -1 \end{cases}$

Some remarks :

① Let's look closely at ε_i :

- $\varepsilon_i = 0$ → Observation x_i is on the correct side of the margin



- $\varepsilon_i > 0$ → Observation x_i is on the wrong side of the margin.

  There are 3 different cases :

  - $0 < \varepsilon_i < 1$ → $1 - \varepsilon_i > 0$ → $M(1-\varepsilon_i) > 0$

    Observation still on correct side of hyperplane
  - $\varepsilon_i = 1$ → $M(1-\varepsilon_i) = 0$
  - $\varepsilon_i > 1$ → Observation on wrong side of hyperplane

② Tuning parameter C gives a bias/variance trade-off

$\sum_{i=1}^{n} \varepsilon_i \leq C$ with $C \geq 0$

In particular:

$C = 0 \rightarrow \varepsilon_1 = \varepsilon_2 = \cdots = \varepsilon_n = 0$ (maximal margin classifier)

$C > 0 \rightarrow$ upper bound on number of violations

Thus, support vectors are now the observations on the margin or on the wrong side of the margin / hyperplane. (This makes SVC "local" methods)

C LARGE: Many support vectors, large margin with many violations
(high bias, low variance)

C SMALL: Few support vectors, small margin with few violations
(low bias, high variance)

So the optimal C is typically chosen by cross-validation!

Let us go back to the optimization problem:

$$\max_{\beta_0,\dots,\beta_p,\ \varepsilon_1,\dots,\varepsilon_n} M$$

s.t. C_1 $\beta_1^2 + \cdots + \beta_p^2 = 1$

C_2 $y_i(\beta_0 + \beta^T x_i) \geq M(1-\varepsilon_i)$ $\quad i=1,\dots,n$

C_3 $\begin{cases} \varepsilon_i \geq 0 \\ \sum_{i=1}^{n} \varepsilon_i \leq C \end{cases}$

C_2 could also be reformulated as

$$\tilde{C}_2: \frac{1}{\|\beta\|} y_i(\beta_0 + \beta^T x_i) \geq M(1-\varepsilon_i)$$

Since $\left\| \frac{\beta}{\|\beta\|} \right\| = 1$

$$\Leftrightarrow \quad y_i(\beta_0 + \beta^T x_i) \geq \underbrace{M\|\beta\|}_{:=1}(1-\varepsilon_i), \quad i=1,\dots,n$$

If (β_0, β) satisfy the inequality, then any positively scaled multiple will also satisfy it. So we scale the parameters so that $\|\beta\| = \frac{1}{M}$, i.e. $M = \frac{1}{\|\beta\|}$.

So the optimisation problem becomes:

$$\max_{\beta_0,\ldots,\beta_p,\ \varepsilon_1,\ldots,\varepsilon_n} \frac{1}{\|\beta\|}$$

s.t.

C_1: $y_i(\beta_0 + \beta^T x_i) \geq (1-\varepsilon_i)$, $i=1,\ldots,n$

C_2: $\varepsilon_i \geq 0$

C_3: $\sum_{i=1}^{n} \varepsilon_i \leq C$

OR

$$\min_{\beta_0,\ldots,\beta_p,\ \varepsilon_1,\ldots,\varepsilon_n} \frac{1}{2}\|\beta\|^2 = \beta_1^2 + \cdots + \beta_p^2$$

s.t.

C_1

C_2

C_3

QUADRATIC CONSTRAINED CONVEX OPTIMIZATION PROBLEM (single global minum)