Support Vector Machines

Roberta De Vito

BROWN
Public Health

Support vector machine

- Maximal Marginal Classifier
- It cannot be applied to most data sets
- Separation by a linear boundary

Hyperplane

- Mathematical Definition

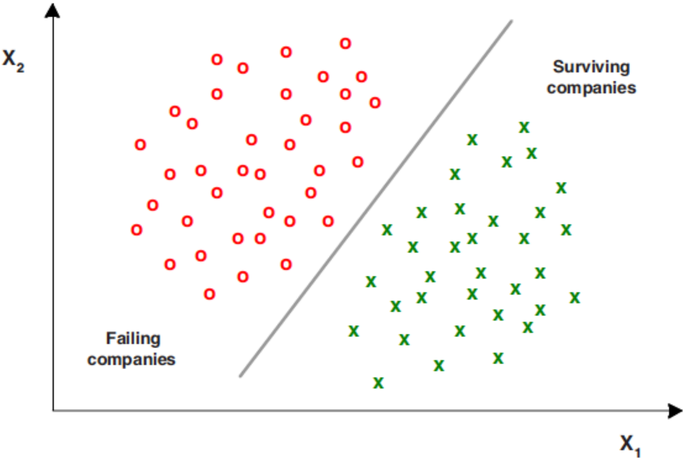
$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p = 0$$

- if X satisfies instead

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p > 0$$

What does it mean (Q1)?

Hyperplane



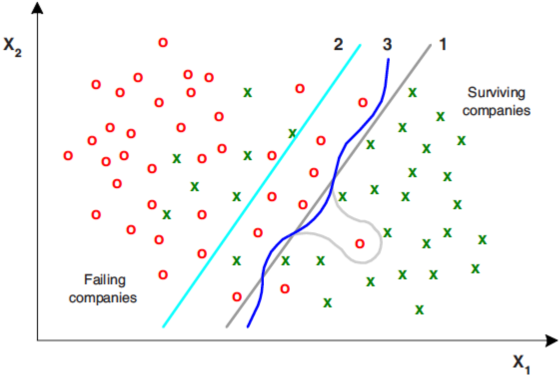
Classification Using a Separating Hyperplane

- Our data matrix

- observations fall into two classes $y_i \in \{-1, 1\}$

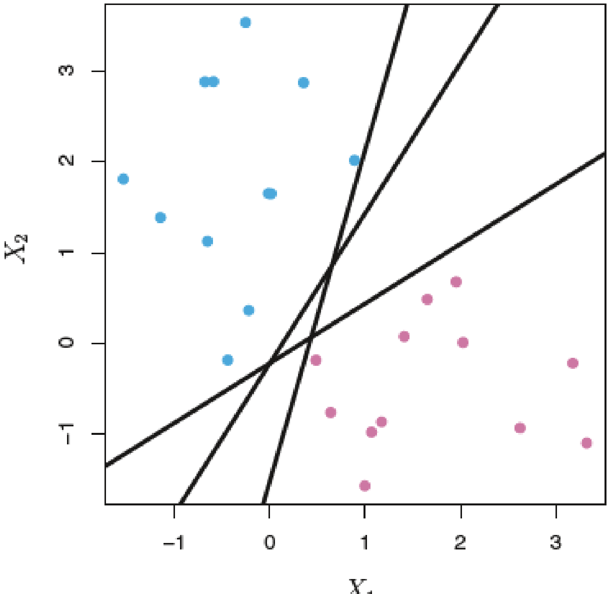
How can we find the best separating hyperplanes

1 It should be linear



How can we find the best separating hyperplanes

Also linear we can find many hyperplanes



How can we find the best separating hyperplanes

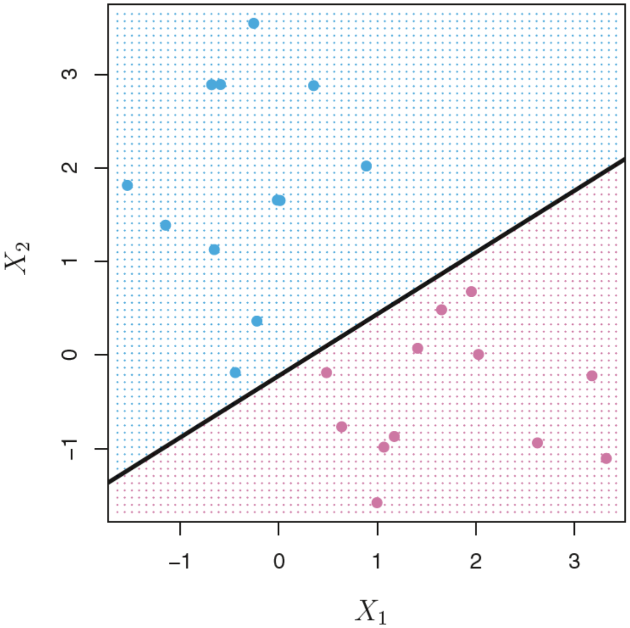
A separating hyperplane has the property that

$$\begin{aligned} \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} &> 0 \text{ if } y_i = 1, \\ \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} &< 0 \text{ if } y_i = -1, \end{aligned}$$

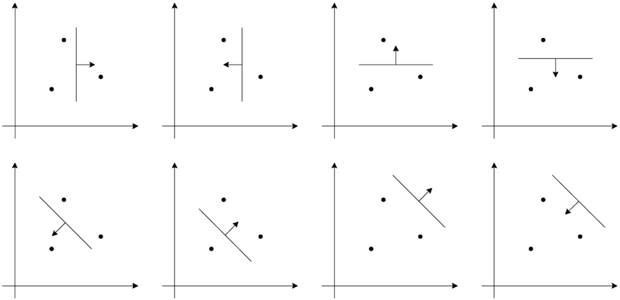
Also we can write

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

How can we find the best separating hyperplanes



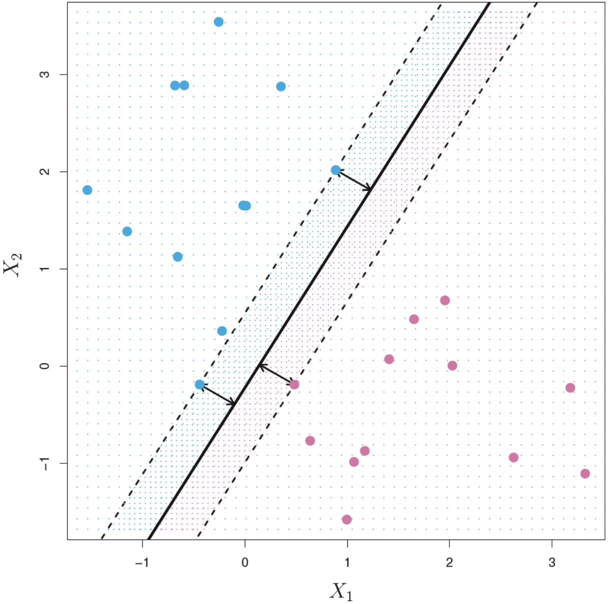
8 Different possible ways to divide three points



The Maximal Margin Classifier

- look for the optimal separating hyperplane
- separating hyperplane that is farthest from the training observation
- perpendicular distance
- then, classify a test observation based the maximal margin hyperplane where it lies

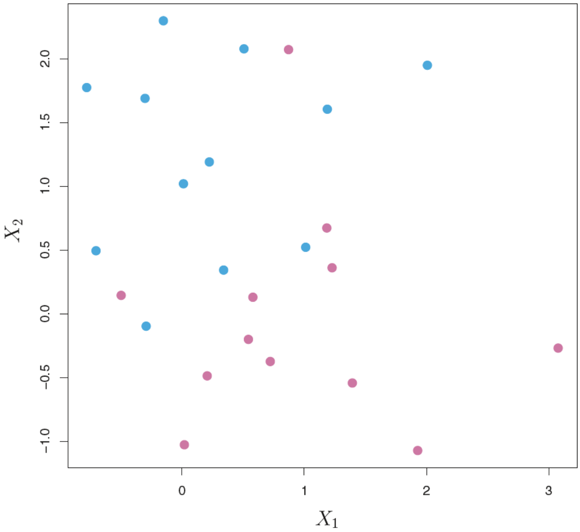
The Maximal Margin Classifier



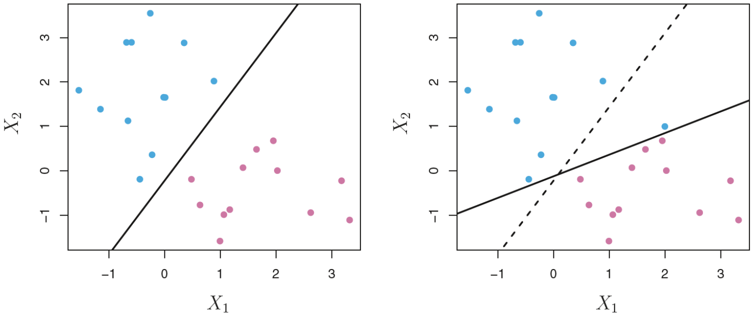
The Maximal Margin Classifier

$$\begin{aligned}
&\textit{maximize } M \\
&\textit{subject to } \sum_{i=1}^{p} \beta_j^2 = 1 \\
&y_i(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}) \geq M
\end{aligned}$$

Observations not separable by a hyperplane



Robustness (Q2)



The soft margin classifier (Q3)

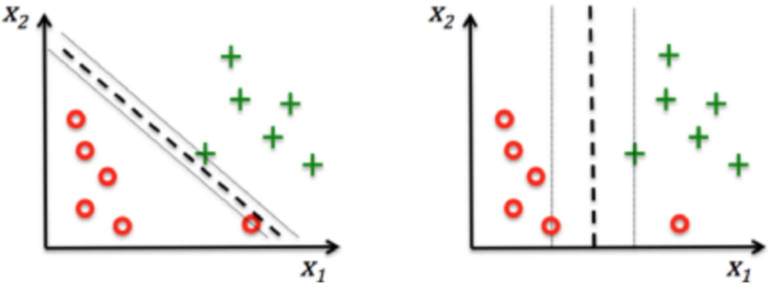
- margin is soft
- separate most of the training observations into the two classes

$$
\begin{aligned}
&\textit{maximize } M \\
&\textit{subject to } \sum_{i=1}^{p} \beta_j^2 = 1 \\
&y_i(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}) \geq M(1 - \epsilon_i) \\
&\epsilon_i \geq 0, \sum_{i=1}^{n} \epsilon_i \leq C
\end{aligned}
$$

The severity of violiations

- C determines the severity of the violations to the margin
- $C = 0$, no margin for violiations $\epsilon_1 = \cdots = \epsilon_n = 0$
- $C > 0$ no more than C observations can be on the wrong side of the hyperplane
- C is chosen via CV
- low-bias tradeoff (Q4)

The severity of violiations Q5



Non-linear Decision Boundaries

- $2p$ features $X_1, X_1^2, \ldots, X_p, X_p^2$
- Now our equations will become

$$\textit{maximize } M$$

$$\sum_{j=1}^{p} \sum_{k=1}^{2} \beta_{jk}^2 = 1$$

$$\textit{subject to } y_i(\beta_0 + \sum_{j=1}^{p} \beta_{j1} x_{ij} + \sum_{j=1}^{p} \beta_{j2} x_{ij}^2) \geq M(1 - \epsilon_i)$$

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

SVM with kernels

- Solution of the equations involves inner products of the observations

$$< x_i, x_{i'} >= \sum_{j=1}^{p} x_{ij} x_{i'}$$

- the linear support vector classifier

$$f(x) = \beta_0 + \sum_{i=1}^{n} \alpha_i < x, x_i >$$

- kernel

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

- the non-linear function now is

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

SVM with kernels (Q6)