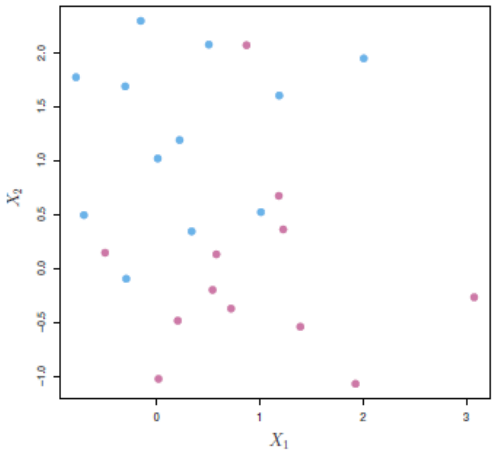
9.2 Support Vector Classifiers

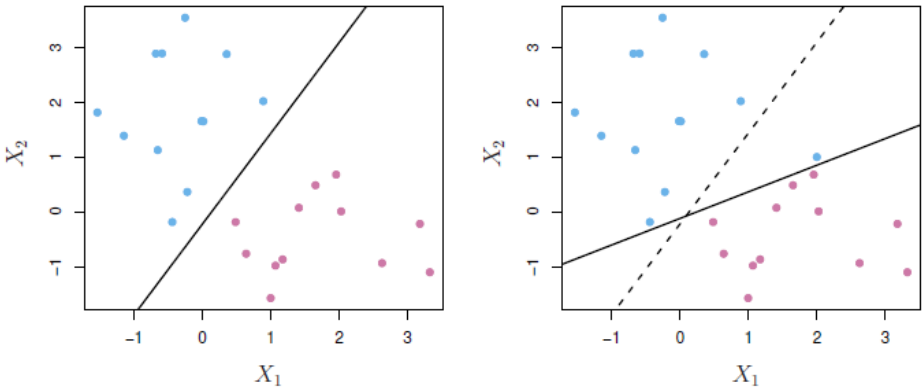
Non-Separable Data



The data on the left are not separable by a linear boundary.

This is often the case, unless $N < p$.

Noisy Data



Sometimes separable data are noisy in a way that generates poor solutions for the maximal-margin classifier.

Noisy Data

- Changing one observation can dramatically change the maximal margin hyperplane.
  - But the distances between points and hyperplane are a measure of confidence.
  - So maximal margin classifiers aren't very robust.

Noisy Data

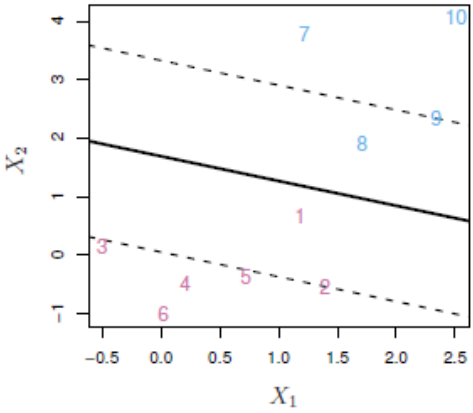
We might consider a classifier based on a hyperplane that doesn't fully separate the two classes for

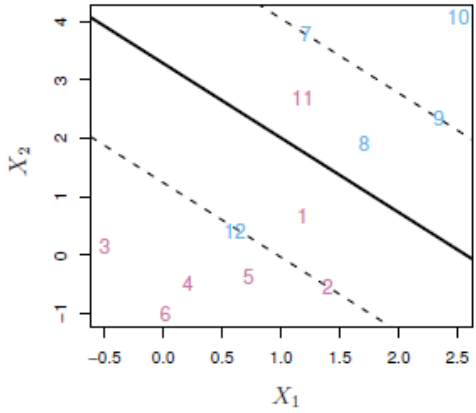
- Greater robustness to individual observations.
- Better classification of *most* of the training observations.
  - (A little bit of bias for the sake of reduced variability!)

Support Vector Classifiers

- Sometimes called *soft margin classifier*, maximize a *soft* margin.
- Instead of requiring the hyperplane to fully separate the two classes, we allow some observations to be on the incorrect side of the margin, or (in the case where the classes are not fully separable) on the wrong side of the hyperplane.

Support Vector Classifiers





Details

- We again classify a test observation depending on which side of the hyperplane it lies.
- But now we soften the constraints.

$$\underset{\beta_0,\beta_1,\ldots,\beta_p,\epsilon_1,\ldots,\epsilon_n}{\text{maximize}} \quad M \quad \text{subject to} \quad \sum_{j=1}^{p} \beta_j^2 = 1,$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}) \geq M(1 - \epsilon_i),$$

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

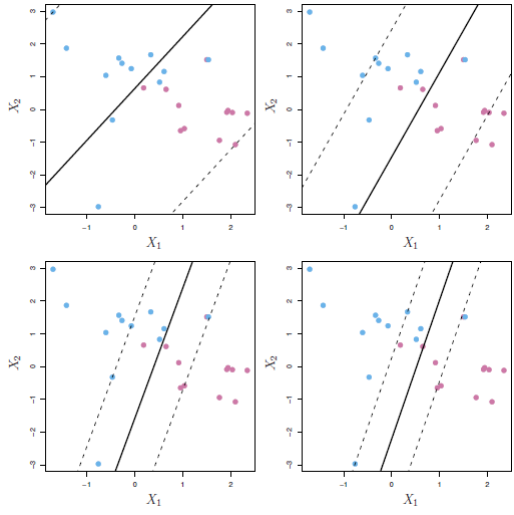
Details

- M is again the width of the margin.
- $\epsilon_1, \ldots, \epsilon_n$ are *slack variables* which allow individual observations to fall on the wrong side of the margin or hyperplane.
  - This tells us where the ith observation is relative to the hyperplane and margin.
  - $\epsilon_i = 0$ if the ith observation is on the correct side of the margin
  - if $\epsilon > 1$, it is on the wrong side of the hyperplane

Details

- C is some nonnegative tuning parameter.
  - This is like a budget for how much the margin can be violated across the n observations.
  - $C = 0$ means no violations (which forces all ϵ_i to 0.
  - No more than C observations can be on the wrong side of the hyperplane.
- We typically choose C using cross-validation.
- C controls the bias-variance trade-off

C is a regulariziation parameter



Properties

- Only observations that fall on the margin or violate the margin will affect the hyperplane.
  - These are the *support vectors* for this method.
- This method is robust to the behavior of observations far from the hyperplane.
  - This is *not* true of methods like linear discriminant analysis.

Example Code

```r
library(e1071)
svmfit <- svm(AHD ~ ., data = Heart, kernel = "linear", scale = FALSE)
summary(svmfit)
```

```
##
## Call:
## svm(formula = AHD ~ ., data = Heart, kernel = "linear", scale = FALSE)
##
##
## Parameters:
##    SVM-Type:  C-classification
##  SVM-Kernel:  linear
##        cost:  1
##
## Number of Support Vectors:  112
##
##  ( 58 54 )
```