Support Vector Machines
(ISLR 9.1-9.4)

Yingbo Li

Southern Methodist University

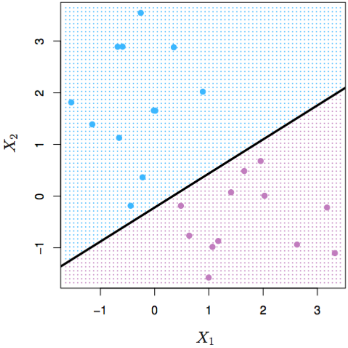
STAT 4399

Outline

Support Vector Machines

Here we approach the two-class classification problem in a direct way:

We try and find a plane that separates the classes in predictor space.



If we cannot, we get creative in two ways:

- Soften what we mean by "separates", and
- Enrich and enlarge the predictor space so that separation is possible.

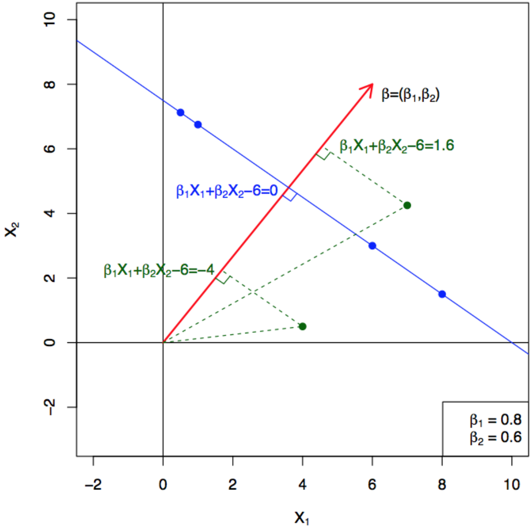
Hyperplane

- A *hyperplane* in $\mathbb{R}^p$ is a flat affine subspace of dimension $p-1$.
- Equation for a hyperplane:

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

  - If $\beta_0 = 0$, the hyperplane goes through the origin, otherwise not.
  - The vector $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)$ points in a direction orthogonal to the surface of a hyperplane.
  - In $p = 2$ dimensions, a hyperplane is a line.
- Let $f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$.
  - If $f(X) > 0$, then the point X lies on one side of the hyperplane,
  - if $f(X) < 0$, then the point X lies on the other side of the hyperplane.

A Hyperplane in $p = 2$ Dimensions



Separating Hyperplane

- For a binary response, we label a “yes” as $+1$, and a “no” as -1.
- If there exists a hyperplane $f(X) = 0$ such that

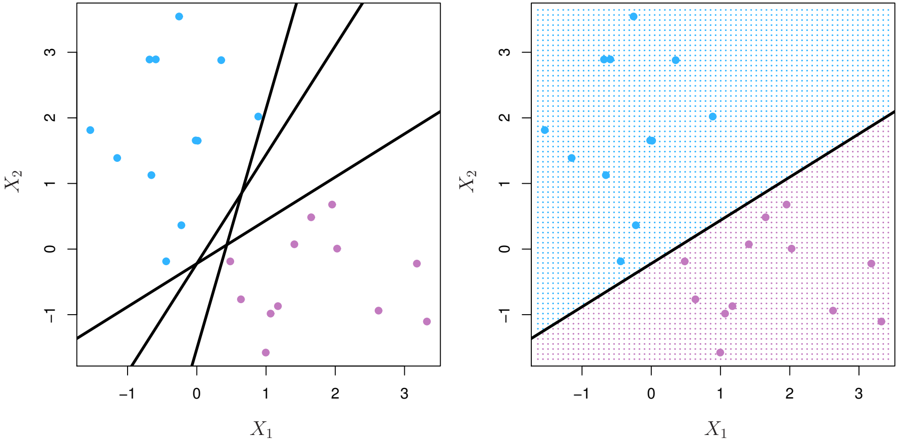
$$f(X_i) \begin{cases} > 0 & \text{if } y_i = 1 \\ < 0 & \text{if } y_i = -1, \end{cases}$$

  then we say $f(X) = 0$ is a *separating hyperplane*.
- Equivalently, $f(X) = 0$ is a separating hyperplane if

$$y_i f(X_i) > 0, \quad \text{for all } i = 1, 2, \ldots, n.$$

- If a separating hyperplane exists, we can use it to construct a very natural classifier: a test observation is assigned a class depending on which side of the hyperplane it is located.
- *Magnitude* of a observation x^* is $f(x^*)$: sign, absolute value

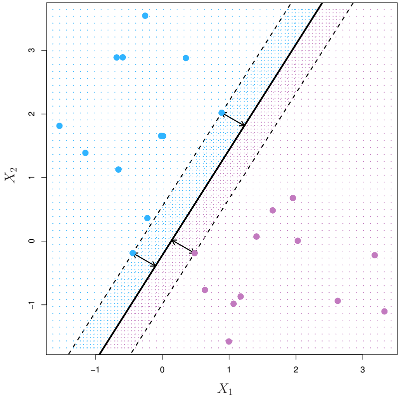
A Example of Separating Hyperplanes in $\mathbb{R}^2$



Blue: $y = 1$, Purple: $y = -1$

The Maximal Margin Classifier

- *Margin*: the minimal distance from the points to the hyperplane.
- *Maximal Margin Classifier*: among all separating hyperplanes, the one that makes the biggest margin (i.e., gap between the two classes).



Terminologies

- Maximal margin hyperplane: the solid line
- Support vectors: points on the dashed lines
- Margin: the distance from either dashed line to the solid line

Finding the Maximal Margin Classifier

For a dataset, if there exist separating hyperplanes, then we can construct of the Maximal Margin Classifier by solving an optimization question:

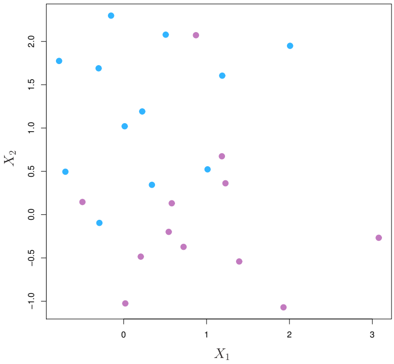
$$\max_{\beta_0,\beta_1,\ldots,\beta_p} M$$

$$\text{subject to} \sum_{j=1}^{p} \beta_j^2 = 1$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}) \geq M, \text{ for } i = 1, 2, \ldots, n.$$

- The perpendicular distance from the ith observation to the hyperplane is
$$y_i(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}).$$
- M represents the margin of the hyperplane

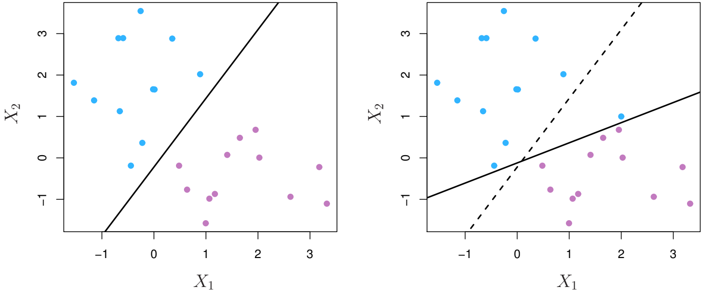
Non-separable Data



- For this dataset, there does not exist a separating hyperplane.
- This is often the case, unless $p > n$.

Maximal Marginal Classifier is Unstable

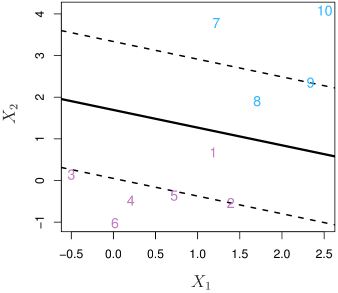
- The maximal margin hyperplane is extremely sensitive to a change in a single observation, so it may have overfit the training data.



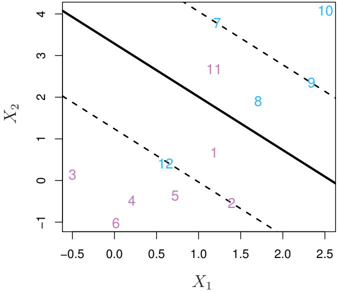
- It is worthwhile to misclassify a few training observations in order to do a better job in classifying the remaining observations.

Support Vector Classifier

- The support vector classifier maximizes a soft margin.
- An observation can be not only on the wrong side of the margin, but also on the wrong side of the hyperplane.



Wrong side of the margin
Observation 1, 8



Wrong side of the hyperplane
Observation 11, 12

Support Vector Classifier Solution

$$\max_{\beta_0,\beta_1,\dots,\beta_p,\epsilon_1,\dots,\epsilon_n} M$$

$$\text{subject to} \sum_{j=1}^{p} \beta_j^2 = 1$$

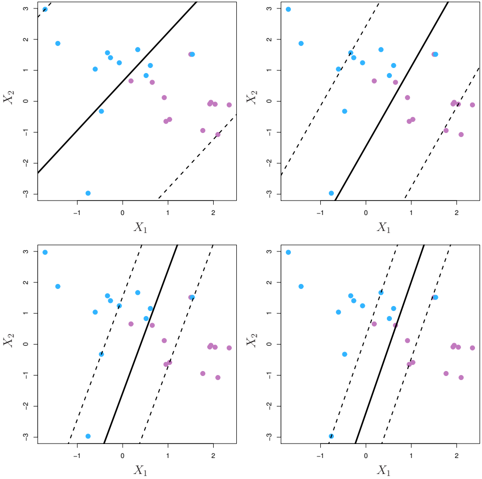
$$y_i(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}) \geq M(1 - \epsilon_i)$$

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

- $\epsilon_1, \dots, \epsilon_n$: slack variables that allow individual observations to be on the wrong side of the margin or the hyperplane
  - $\epsilon_i = 0$: on the correct side of the margin
  - $\epsilon_i > 0$: on the wrong side of the margin
  - $\epsilon_i > 1$: on the wrong side of the hyperplane
- C: a nonnegative tuning parameter.

C is a Regularization Parameter

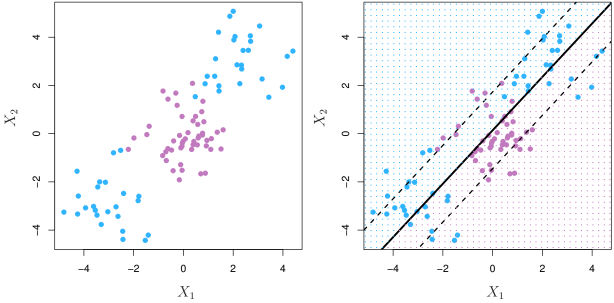
- A budget for the amount that the margin can be violated.
- Small C: narrow margins, few violations, highly fitting



- *Support vectors*: observations that lie directly on the margin, or on the wrong side of the margin.
- Support vector classifier's decision rule is based only on the support vectors: it is quite robust to the observations that are far away from the hyperplane.

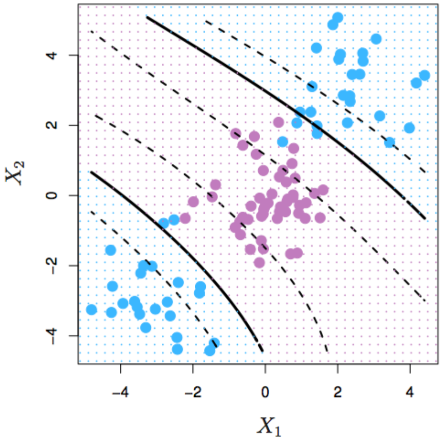
Feature Expansion

Linear boundary can fail



- Enlarge the space of features (predictors) by including transformations; e.g. $X_1^2, X_1^3, X_1X_2, \ldots$.
- Fit a support-vector classifier in the enlarged space.
- This results in non-linear decision boundaries in the original space.

Cubic Polynomials



- We use a basis expansion of cubic polynomials
- The support-vector classifier in the enlarged space solves the problem in the lower-dimensional space
- Polynomials (especially high-dimensional ones) get wild rather fast.
- There is a more elegant and controlled way to introduce nonlinearities in support-vector classifiers — through *kernels*.

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2 + \beta_4 X_2^2 + \beta_5 X_1 X_2 + \beta_6 X_1^3 + \beta_7 X_2^3 + \beta_8 X_1^2 X_2 + \beta_9 X_1 X_2^2 = 0$$

Kernels Representation of the Support Vector Classifier

- *Inner products* between vectors:

$$\langle x_i, x_{i'} \rangle = \sum_{j=1}^{n} x_{ij} x_{i'j}$$

- The linear support vector classifier can be represented as

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

  To estimate the parameters $\beta_0, \alpha_1, \ldots, \alpha_n$, all we need are the inner products $\langle x_i, x_{i'} \rangle$ between all pairs of training observations.

- Actually, it turns out that most of the α_i can be zero:

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

  where $\mathcal{S}$ is the set of support vectors.

Kernels and Support Vector Machines

- We can replace the inner product $\langle x_i, x_{i'} \rangle$ in SV classifer with a generalization of the form $K(x_i, x_{i'})$, i.e., a *kernel* function.
  - For example, a polynomial kernel of degree d:

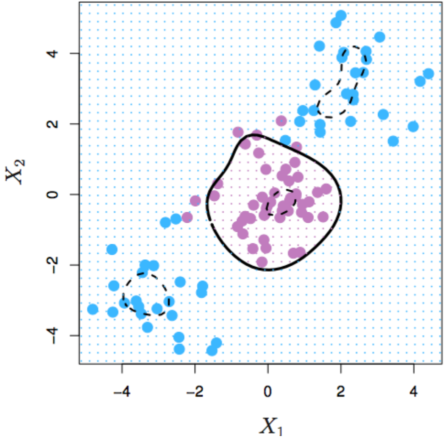
$$K(x_i, x_{i'}) = \left(1 + \sum_{j=1}^{p} x_{ij} x_{i'j}\right)^d$$

- The solution has the form

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

- When the support vector classifier is combined with a non-linear kernel, the resulting classifier is known as a *support vector machine*.

A Common Choice: the Radial Kernel



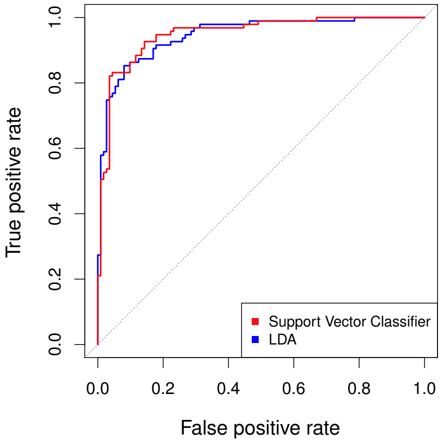
- *Radial kernel*

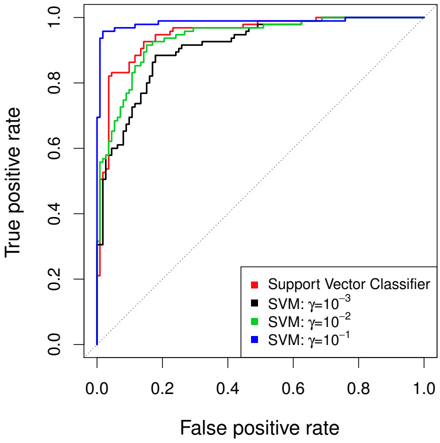
$$K(x_i, x_{i'}) = e^{-\gamma \sum_{j=1}^{p}(x_{ij}-x_{i'j})^2}$$

- Implicit feature space; very high dimensional.
- As γ increases and the fit becomes more non-linear.
- Use CV to decide the tuning parameter γ.

The Heart Data Example

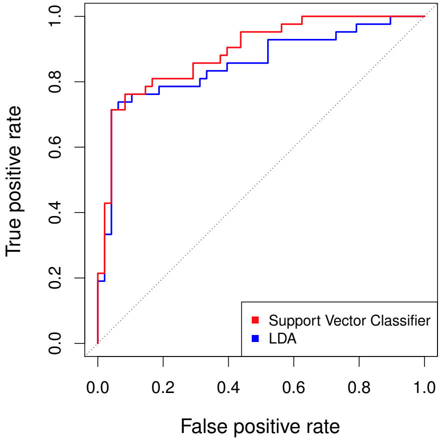
Training data ROC curves

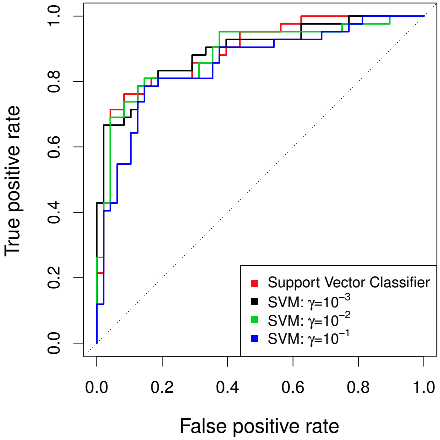




The Heart Data Example

Test data ROC curves





SVMs: More Than Two Classes?

Two ways if we have $K > 2$ classes in the response:

- One versus All.
  Fit K different 2-class SVM classifiers $\hat{f}_k(x), k = 1, \ldots, K$; each class versus the rest. Classify x^* to the class for which $\hat{f}_k(x^*)$ is largest.
- One versus One.
  Fit all $\binom{K}{2}$ pairwise classifiers $\hat{f}_{kl}(x)$. Classify x^* to the class that wins the most pairwise competitions.

Which to choose? If K is not too large, use OVO.