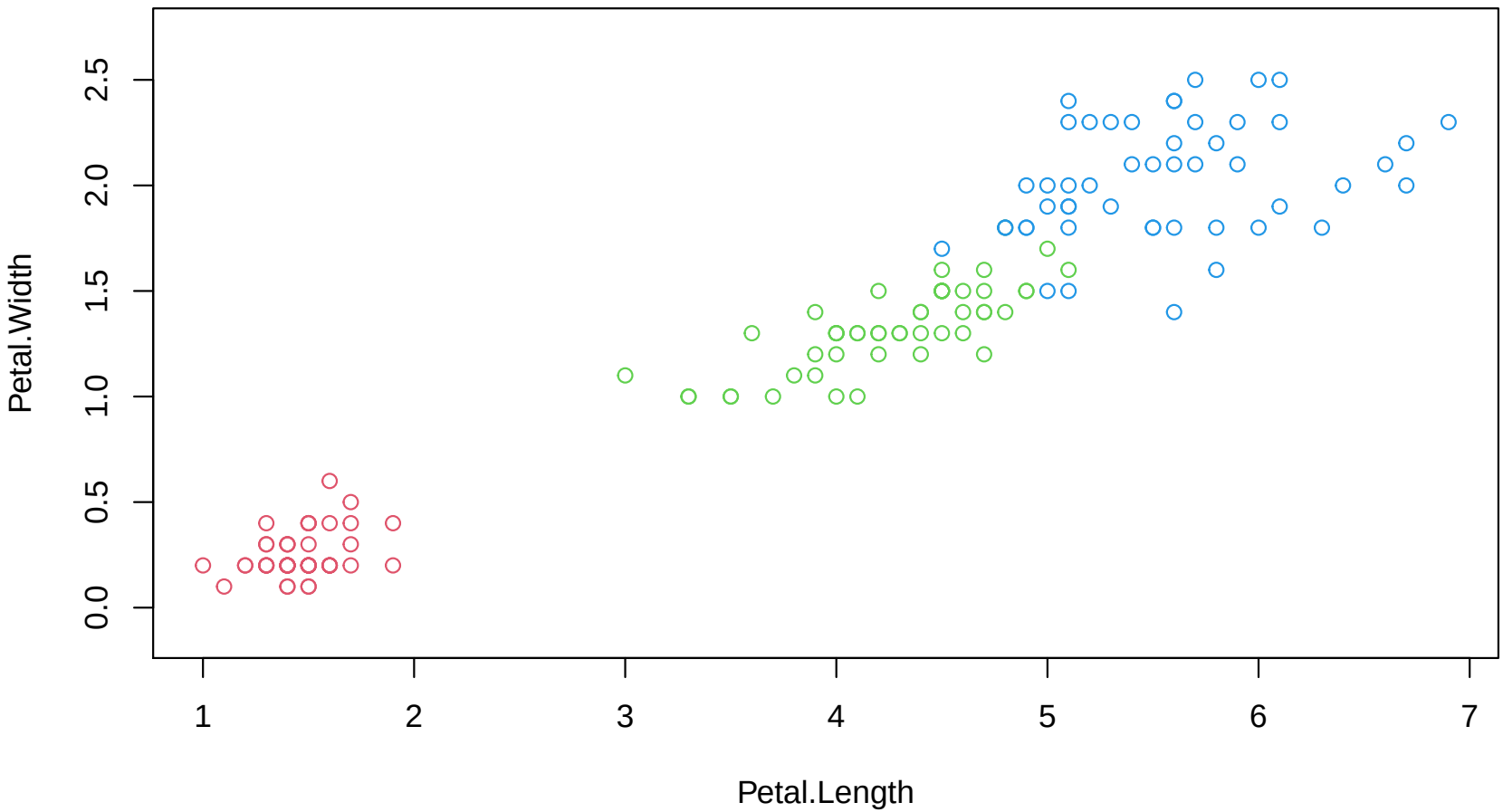
STATS 769

Support Vector Machines

Yong Wang
Department of Statistics
The University of Auckland

2021

Iris Data



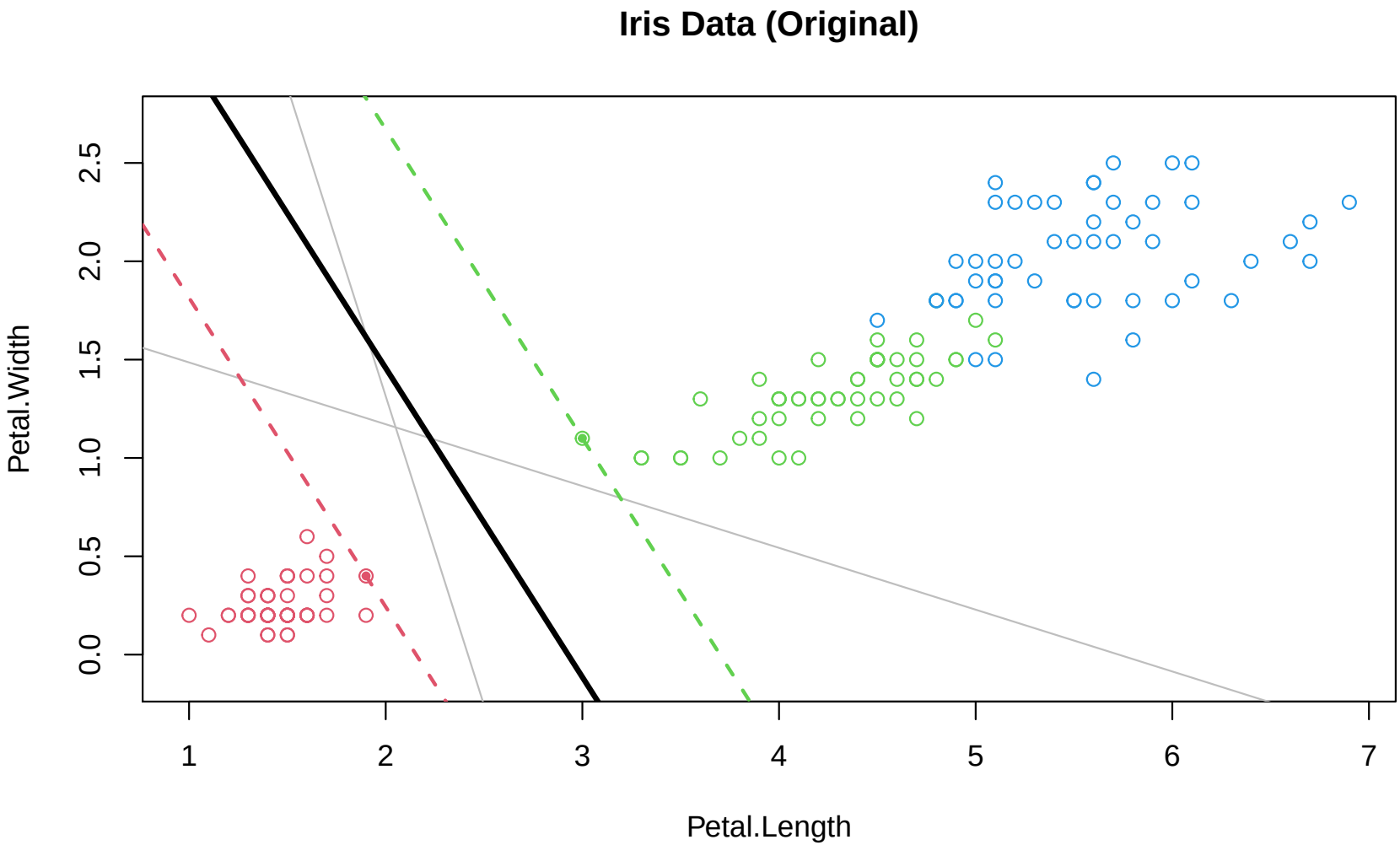
Maximal Margin Classifiers

- A hyperplane in $\mathbb{R}^d$ is defined as

$$\beta_0 + \beta_1 x_1 + \cdots + \beta_d x_d = 0. \tag{1}$$

- Consider a classification problem with $J = 2$ classes
  - $Y = -1$ or 1.
- It is assumed (for now) that the observations can be completely separated according to their class labels, by at least one hyperplane, which is thus known as a separating hyperplane.
- The margin of a separating hyperplane is defined by the smallest perpendicular distance from each of the observations to the hyperplane.
- The maximal margin hyperplane is the separating hyperplane that has the maximal margin.

Maximal Margin Hyperplanes



Maximal Margin Classifiers

- The maximal margin classifier uses the maximal margin hyperplane as its decision boundary.
- There is no observation inside the margin.
- The classifier has a linear discriminant function

$$\delta(x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_d x_d \tag{2}$$

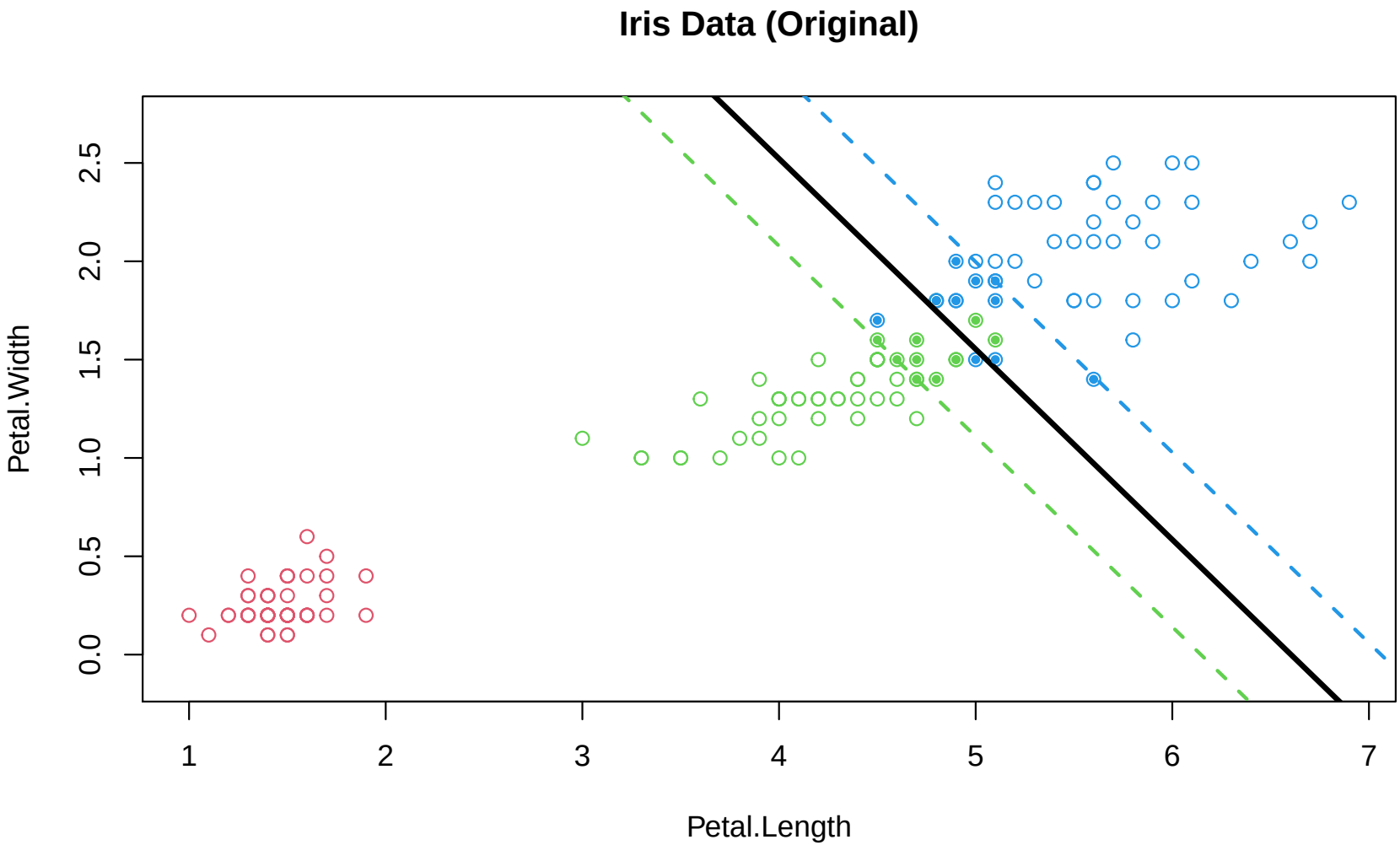
  (just as used for LDA and logistic regression).
- Any observation that attains the maximal margin is known as a support vector.

- Clearly, this classifier does not apply to many practical cases where no separating hyperplane exists.

Support Vector Classifiers

- The support vector classifier (or soft margin classifier) allows an observation to violate the margin or even the hyperplane, with a cost (or penalty).
- The cost for each violating observation is its perpendicular distance to the margin it violates.
- The total cost that observations can at most incur can be used as a tuning parameter C.
- Here, all violating observations, including those directly on the margin, are known as support vectors.

Support Vector Classifiers



Dealing with Nonlinearity

- Using only linear decision boundaries can be inappropriate in many cases.
- To introduce nonlinearity in the decision boundary, one solution is to include nonlinear terms in the function that is used to define the hyperplane (1).
- For example,

$$\beta_0 + \beta_1 x_1 + \cdots + \beta_d x_d + \beta_{d+1} x_1^2 + \cdots + \beta_{2d} x_d^2 = 0. \quad (3)$$

  defines a quadratic surface in $\mathbb{R}^d$.
- This is just like that we have d new predictor variables $x_1^2, \ldots, x_d^2$, and we will use the linear support vector classifier in $\mathbb{R}^{2d}$.

Kernels

- A more general extension is to adopt the concept of kernels.
- A kernel measures the similarity between two observations.
- One can show that:
  - Function (2) can be rewritten as

$$\delta(x) = \beta_0 + \sum_{i=1}^{n} \alpha_i K(x_i, x),$$

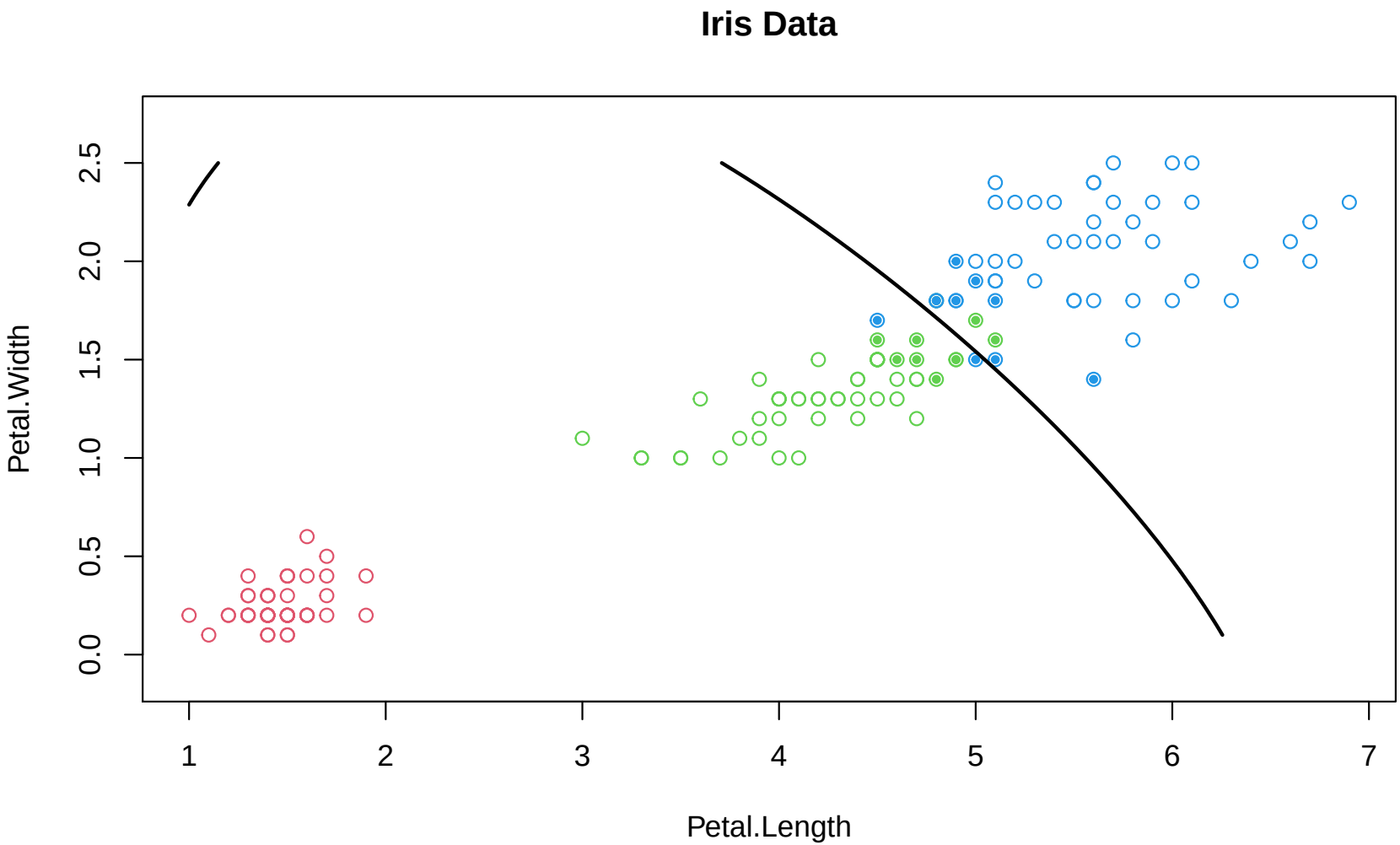
    where $K(x_i, x) = x_i^T x$ is known as a linear kernel and α_i is its coefficient subject to $\alpha_i y_i \geq 0$.
  - At the solution, $\alpha_i \neq 0$ if and only if x_i is a support vector. In particular, $\alpha_i > 0$ if $Y_i = 1$ or $\alpha_i < 0$ if $Y_i = -1$.
- Now consider using nonlinear kernels instead:
  - Polynomial kernels: $K(x_i, x) = (\gamma x_i^T x + 1)^d$, $\gamma > 0$, $d = 1, 2, \ldots$.
  - Radial kernels: $K(x_i, x) = \exp(-\gamma \|x_i - x\|^2)$, $\gamma > 0$.
  - Sigmoid kernels: $K(x_i, x) = \tanh(\gamma x_i^T x + 1)$, $\gamma > 0$.
    - Note $\tanh(x) = (e^{2x} - 1)/(e^{2x} + 1)$.

Support Vector Machines

- A support vector classifier becomes a support vector machine if a nonlinear kernel is used.
- Parameters used in nonlinear kernels, such as d and γ in polynomial kernels and γ in radial and sigmoid kernels, are then tuning parameters.
- We also have the cost as a tuning parameter.
- Posterior probabilities can also be computed by treating the discriminant function as an estimate of the log-odds, i.e.,

$$\mathbb{P}(Y = 1|X = x) = \frac{1}{1 + \exp[-\delta(x)]}.$$

Support Vector Machines



Support Vector Machines: $J > 2$

- One versus one
  - Build all pairwise classifiers
  - Vote from $\binom{J}{2}$ classifiers
  - Not feasible for J large.
- One versus all (of the others)
  - $J-1$ classifiers, each relative to a baseline class 0 and having a discriminant function $\delta_{j0}(x)$.
  - Classify using the largest value of $\delta_{j0}(x)$, $j = 1, \ldots, J-1$; class 0 if it is negative.

- A well-tuned support vector machine is known to have fairly good performance.
- Support vector machines have also been extended to solve regression problems.

Recommended Readings

ISLv2 (basics):

- Sections 9.1, 9.2, 9.3, 9.4.
- Labs: 9.6.1, 9.6.2, 9.6.4

ESL (advanced):

- Sections 12.2, 12.3