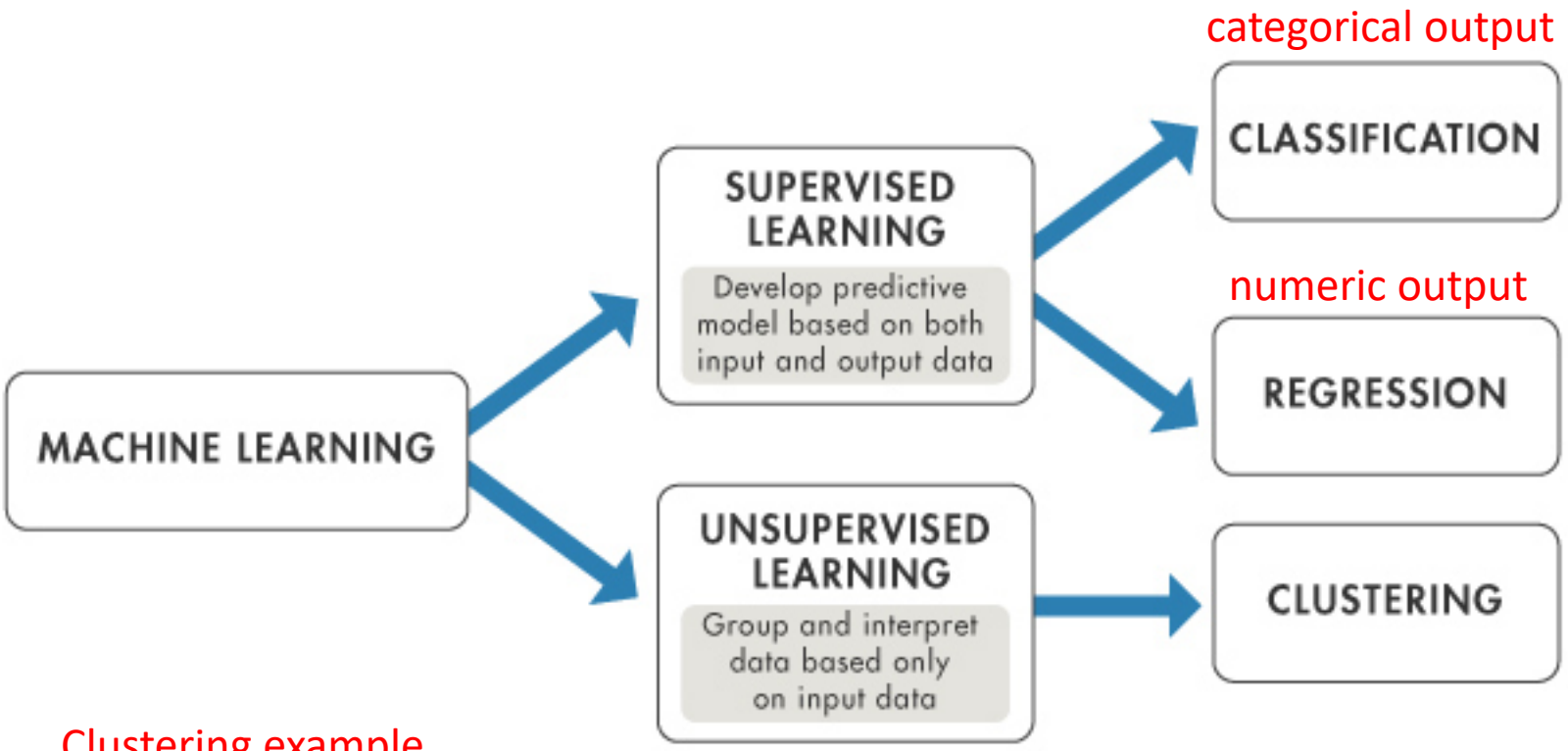
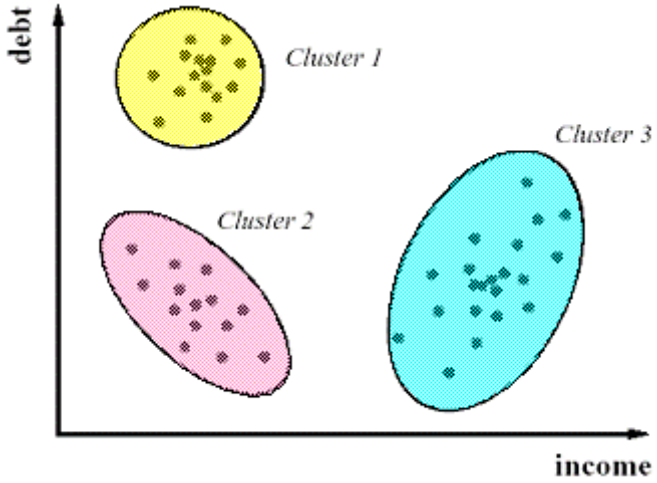
Classification Overview

Garrett Dancik, PhD



Source: https://www.mathworks.com/help/stats/machine-learning-in-matlab.html

Clustering example



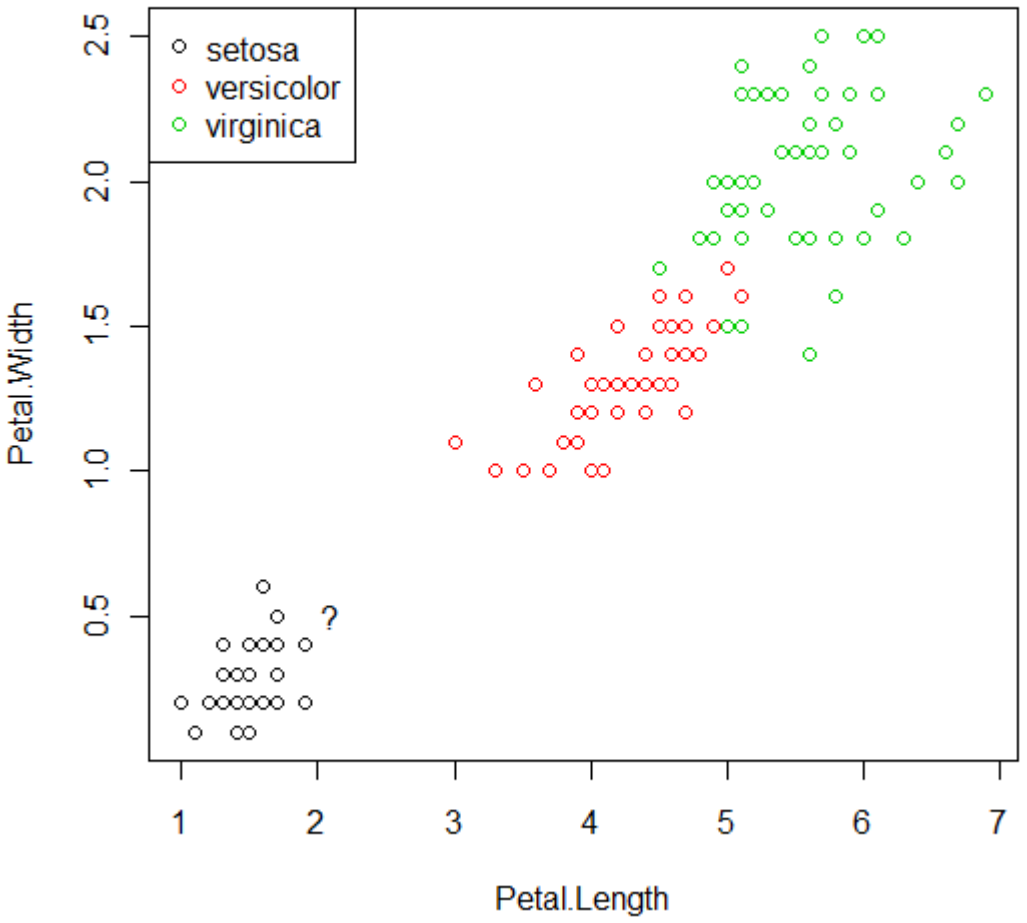
https://www.analyticsvidhya.com/blog/2013/11/getting-clustering-right/

Classification Methods

- Objective: Identify the class (category/label) of an individual (e.g., male or female) based on observed features (e.g., height, weight, etc)
- Classes: $c_1, c_2, \ldots, c_m$ Features: $x_1, \ldots, x_k$
- General Procedure
  - Train the classifier: Using a *training* data set, determine the mapping function $f(x) \rightarrow c$
  - Validation: assess the accuracy of the classifier by applying it to a *test* data set with known classes

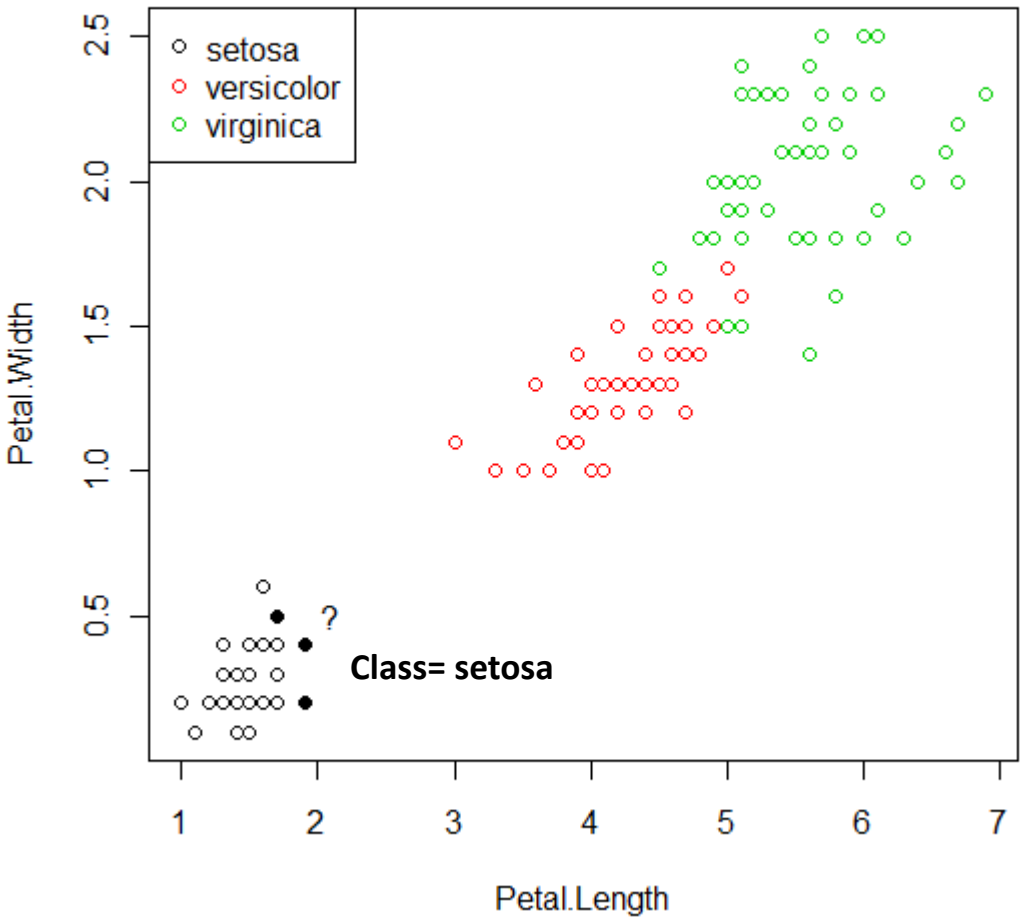
Classification Methods: K-Nearest Neighbors (KNN)

- For a test observation *A*, find the distance between *A* and every other observation in the feature space
- Classify the test observation based on the votes of its *K* nearest neighbors



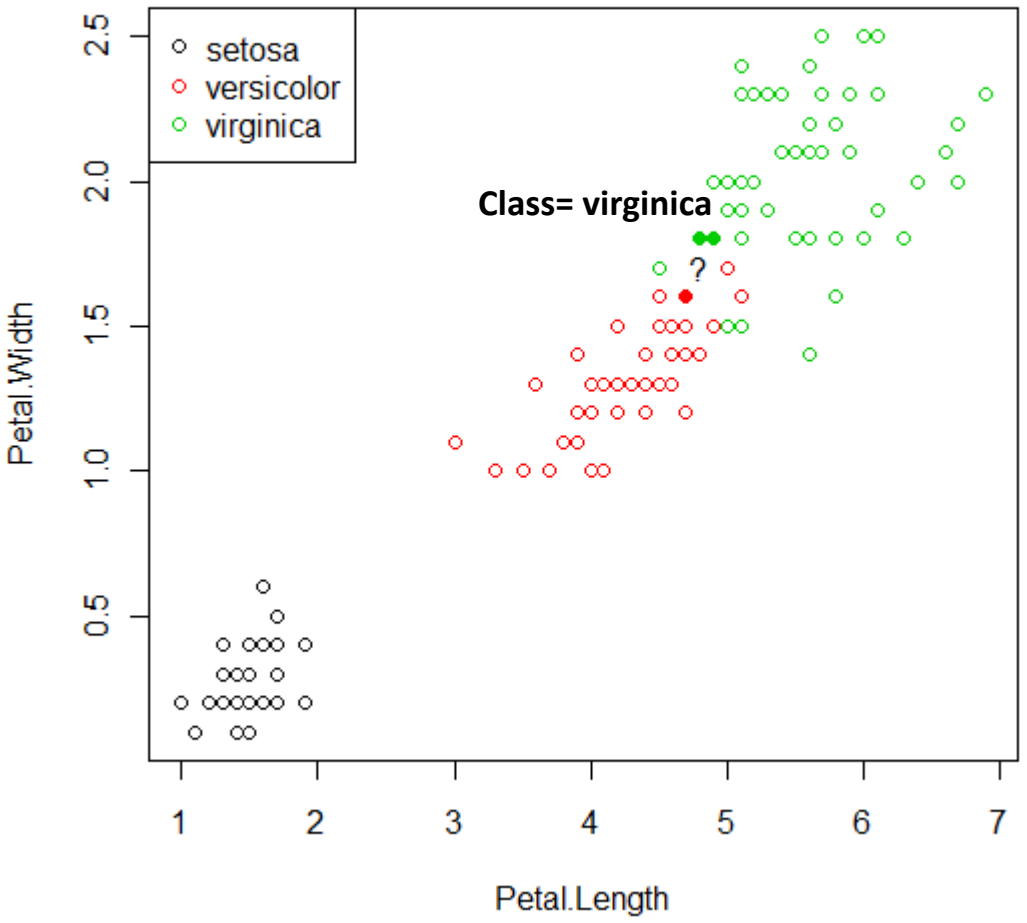
Classification Methods: K-Nearest Neighbors (KNN)

- For a test observation *A*, find the distance between *A* and every other observation in the feature space
- Classify the test observation based on the votes of its *K* nearest neighbors



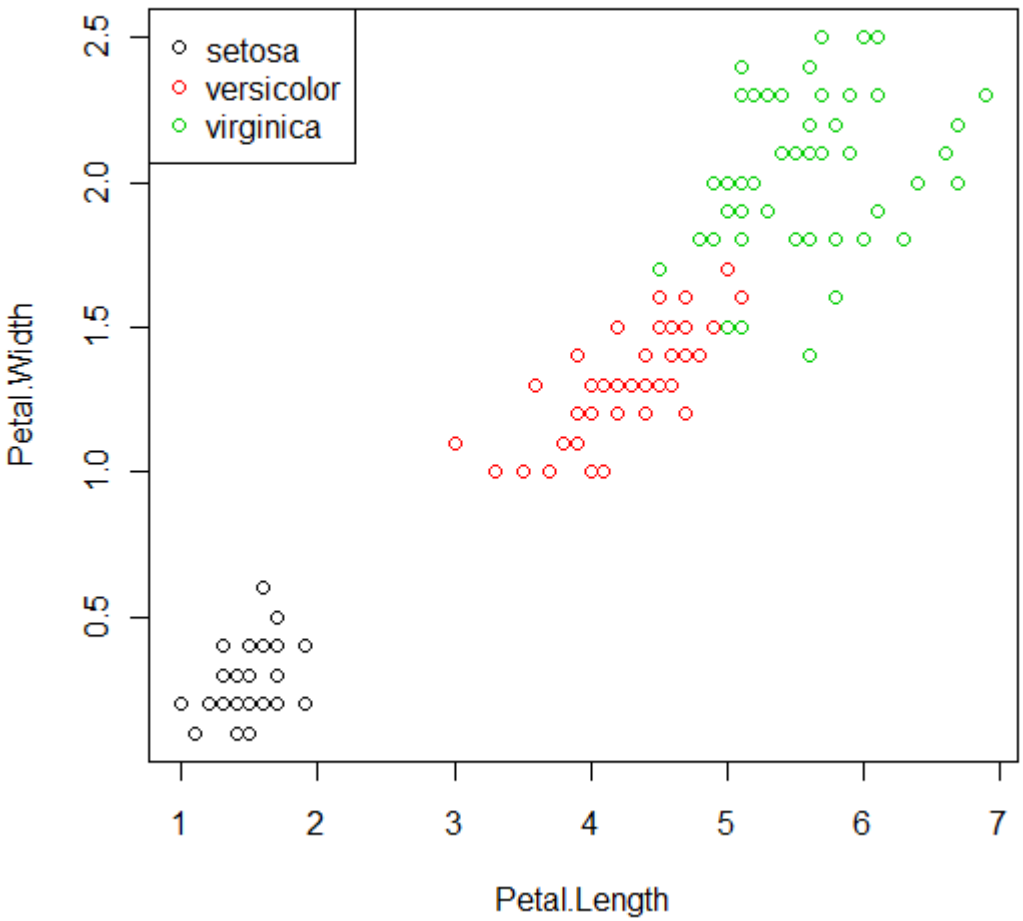
Classification Methods: K-Nearest Neighbors (KNN)

- For a test observation *A*, find the distance between *A* and every other observation in the feature space
- Classify the test observation based on the votes of its *K* nearest neighbors



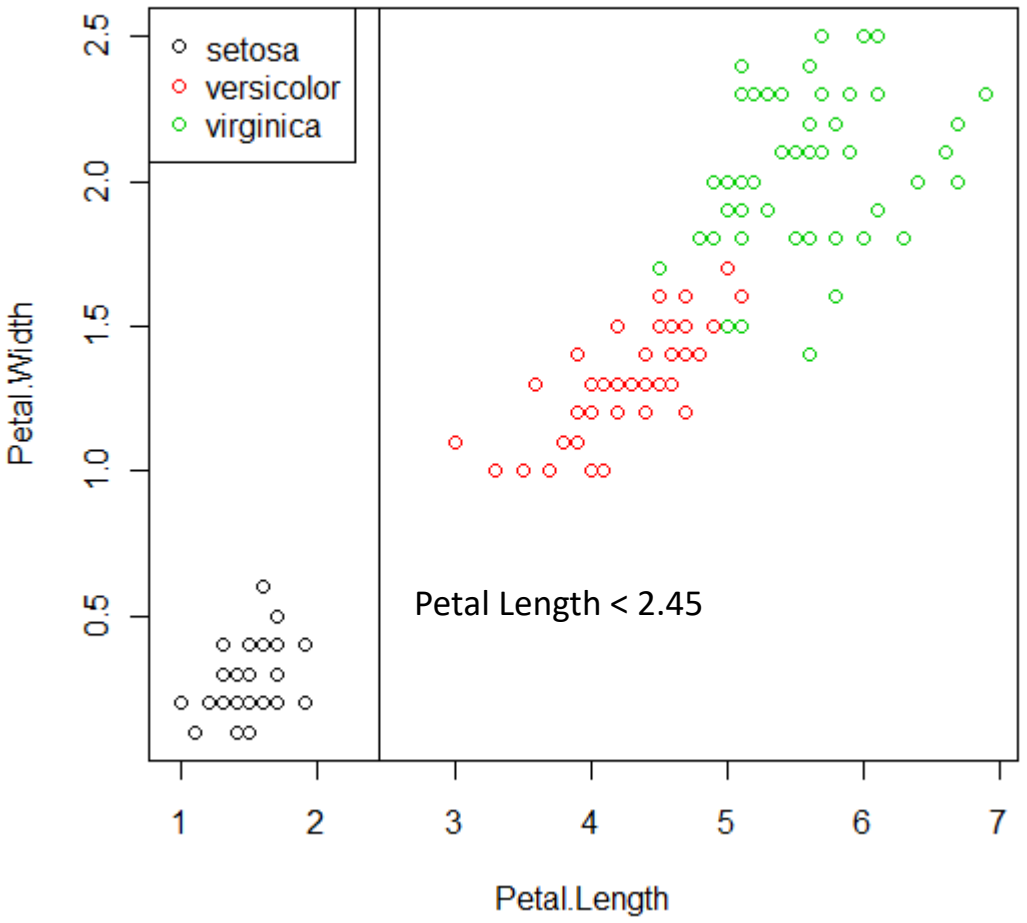
Classification Methods: Decision Trees (DT)

- Create a node by splitting the data according to a feature that optimally splits the data
- Repeat on data subsets until a stopping criterion is met
- Each leaf corresponds to a class



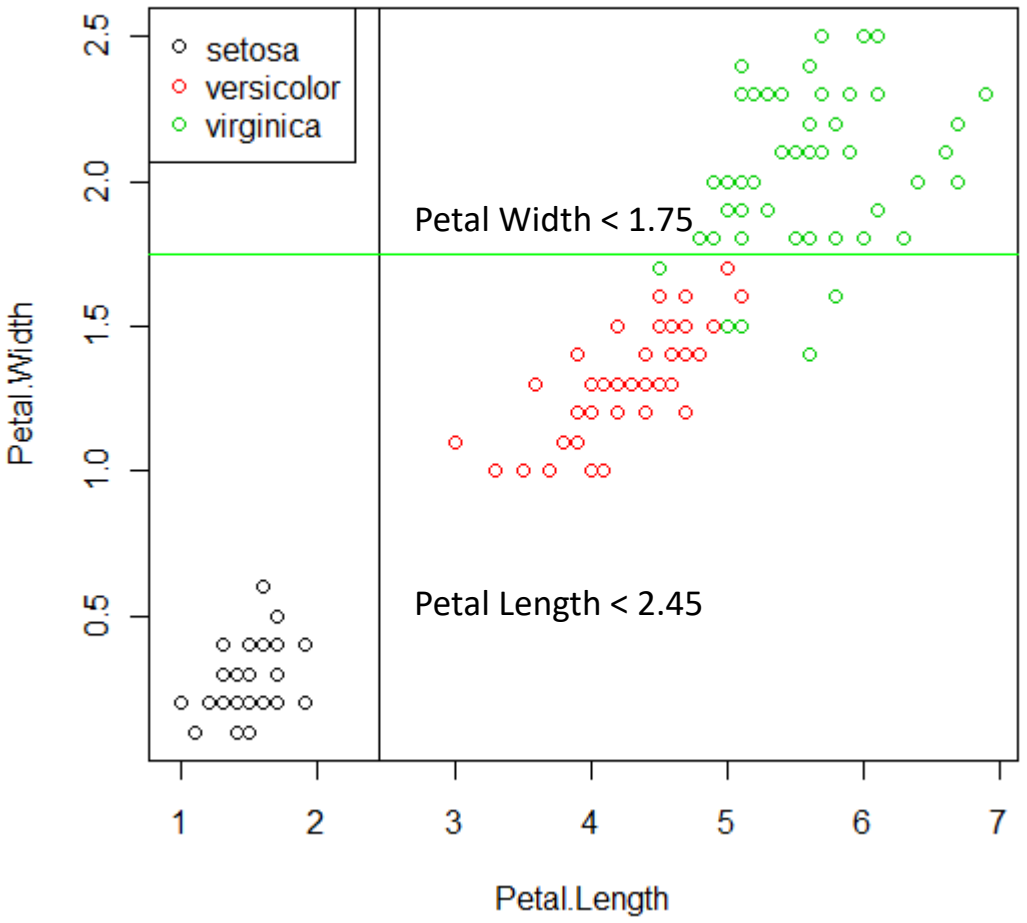
Classification Methods: Decision Trees (DT)

- Create a node by splitting the data according to a feature that optimally splits the data
- Repeat on data subsets until a stopping criterion is met
- Each leaf corresponds to a class

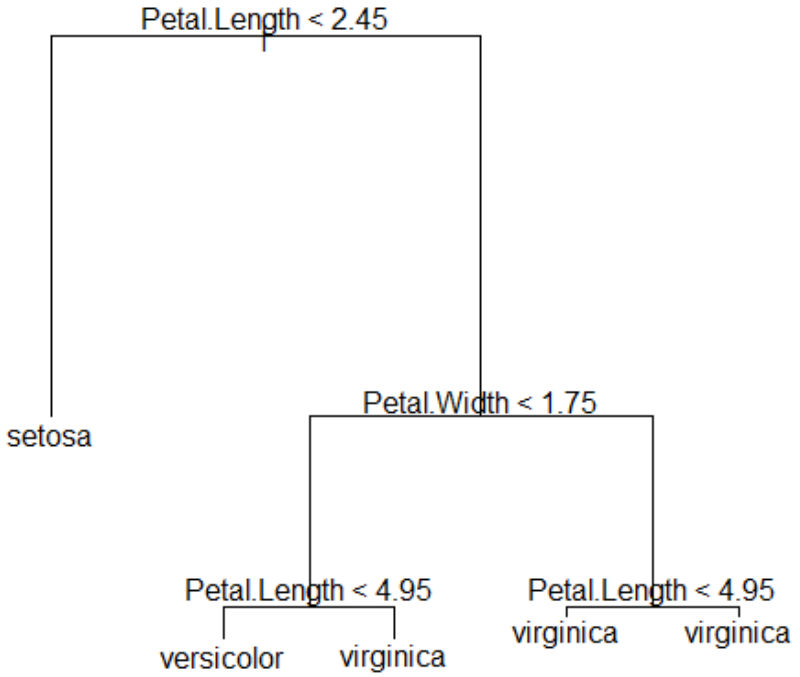


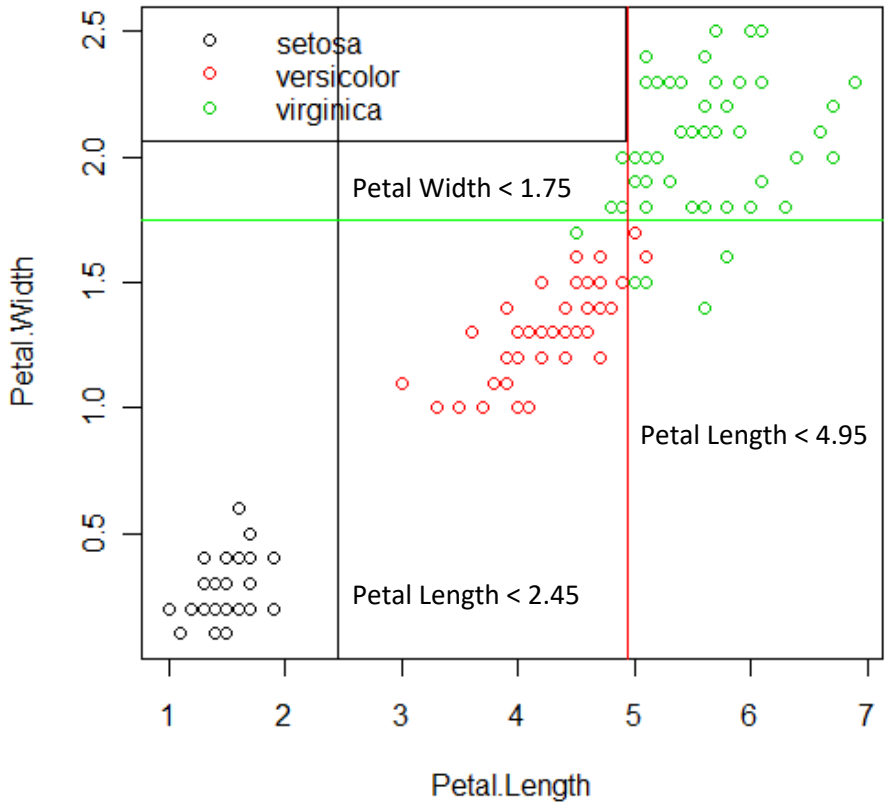
Classification Methods: Decision Trees (DT)

- Create a node by splitting the data according to a feature that optimally splits the data
- Repeat on data subsets until a stopping criterion is met
- Each leaf corresponds to a class



Classification Methods: Decision Trees (DT)





Note: DT are known to overfit data. However more rubust methods such as Random Forests can be used

Classification Methods: Naïve Bayes (NB)

- Based on Bayes' theorem that relates conditional probabilities

$$p(C|x_1, \dots, x_n) \propto p(x_1, \dots, x_n|C)$$

- Naïve Bayes assumes independence of features, so that

$$p(x_1, \dots, x_n|C) = p(x_1|C) \times \cdots \times p(x_n|C) p(C)$$

- For quantitative features, calculate by treating

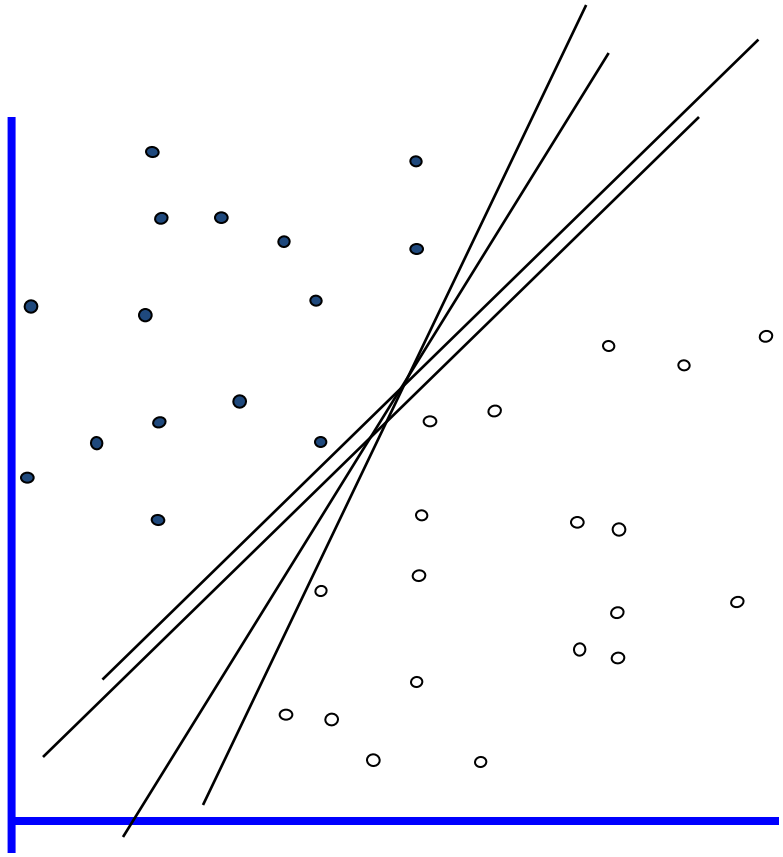
$$p(x|C) \sim N(\mu_x, \sigma_x)$$

- Select the class *C* that maximizes

$$p(C|x_1, \dots, x_n) \propto p(x_1|C) \times \cdots \times p(x_n|C) p(C)$$

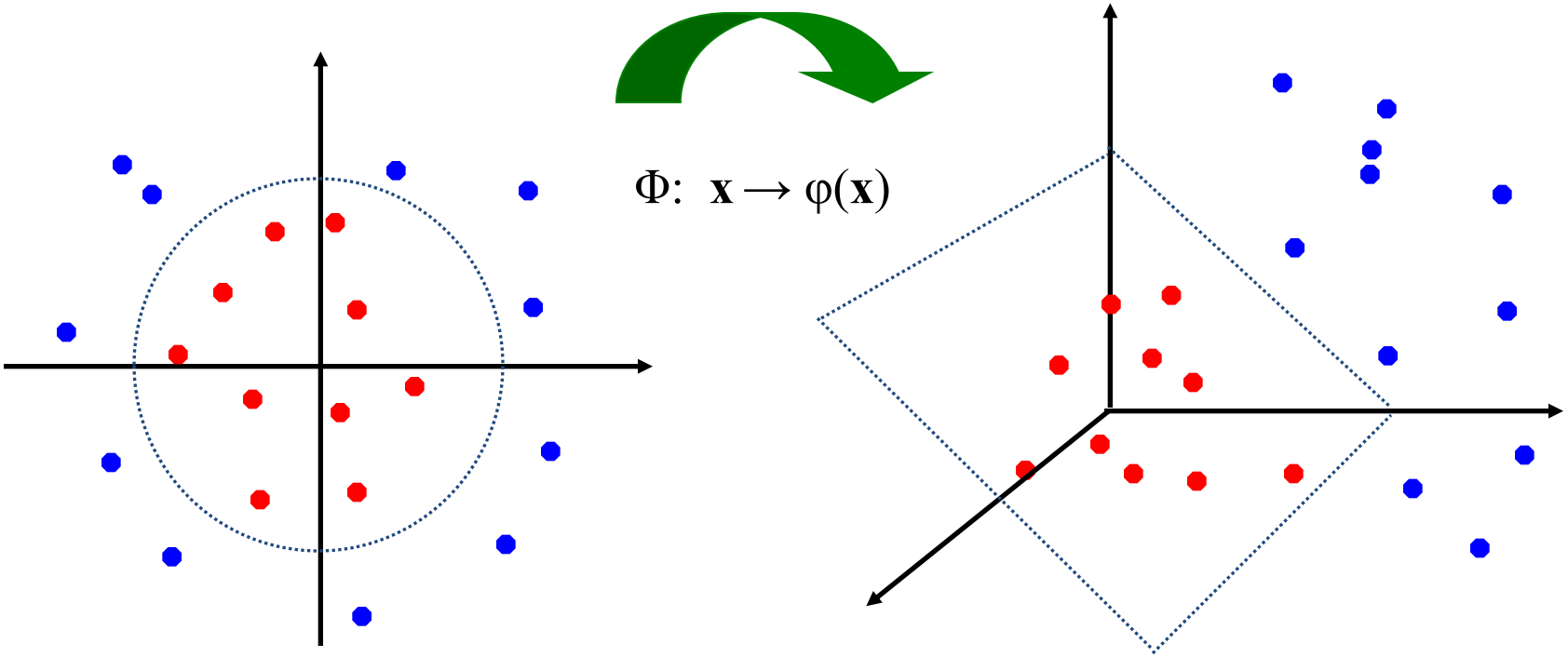
Classification Methods: Support Vector Machines (SVM)

- Find the optimum hyperplane that linearly separates the classes
- If classes are not linearly separable, map the data into a higher dimensional space through the use of a kernel function



Images modified from Mingyue Tan / Andrew Moore

Classification Methods: Support Vector Machines (SVM)



Images modified from Mingyue Tan / Andrew Moore