# Classification Overview

Garrett Dancik, PhD

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

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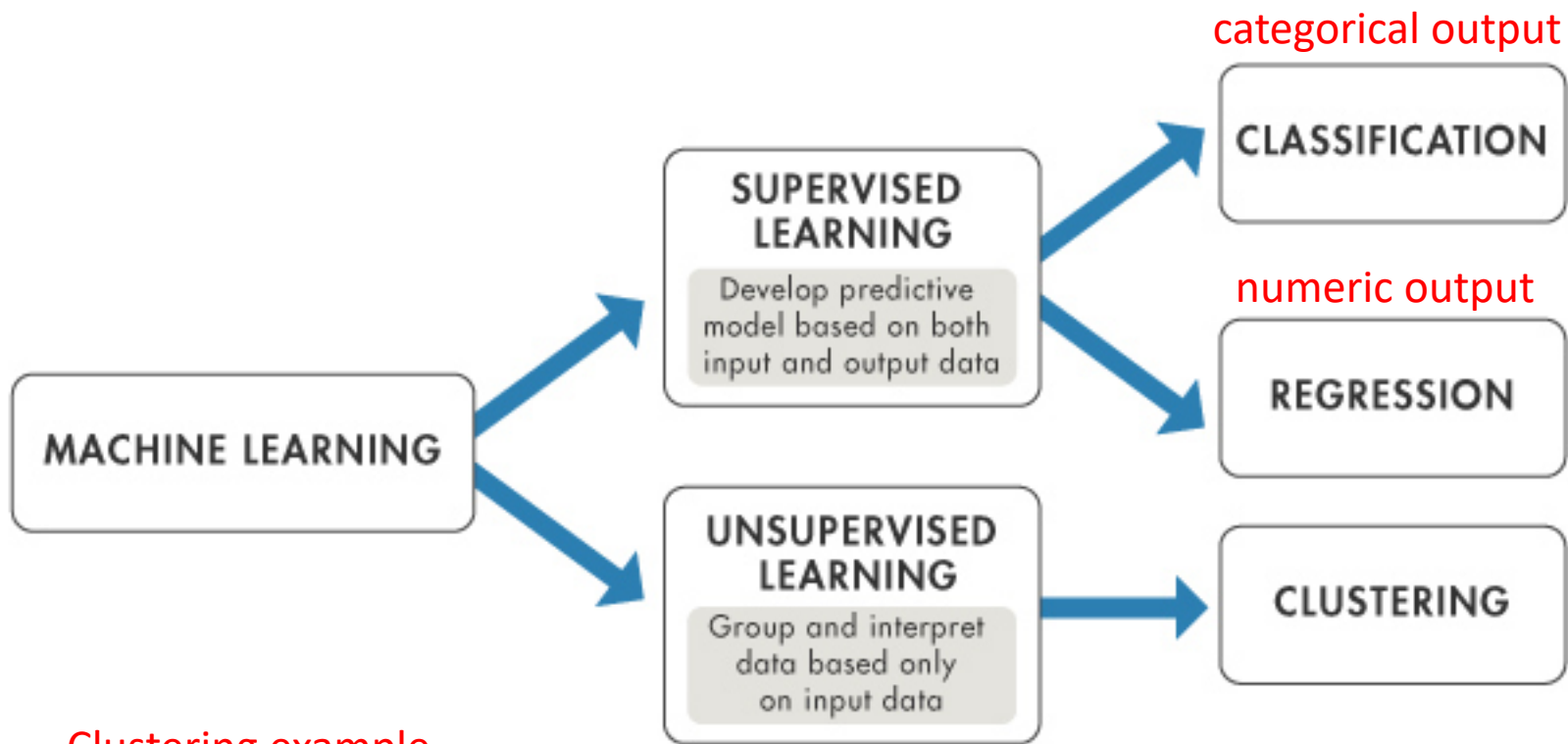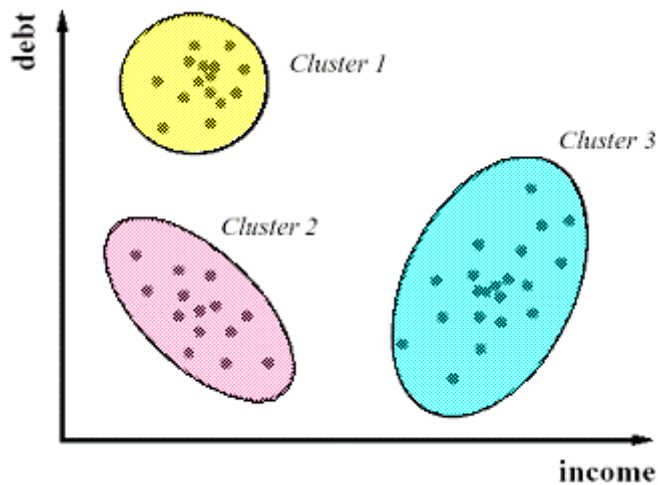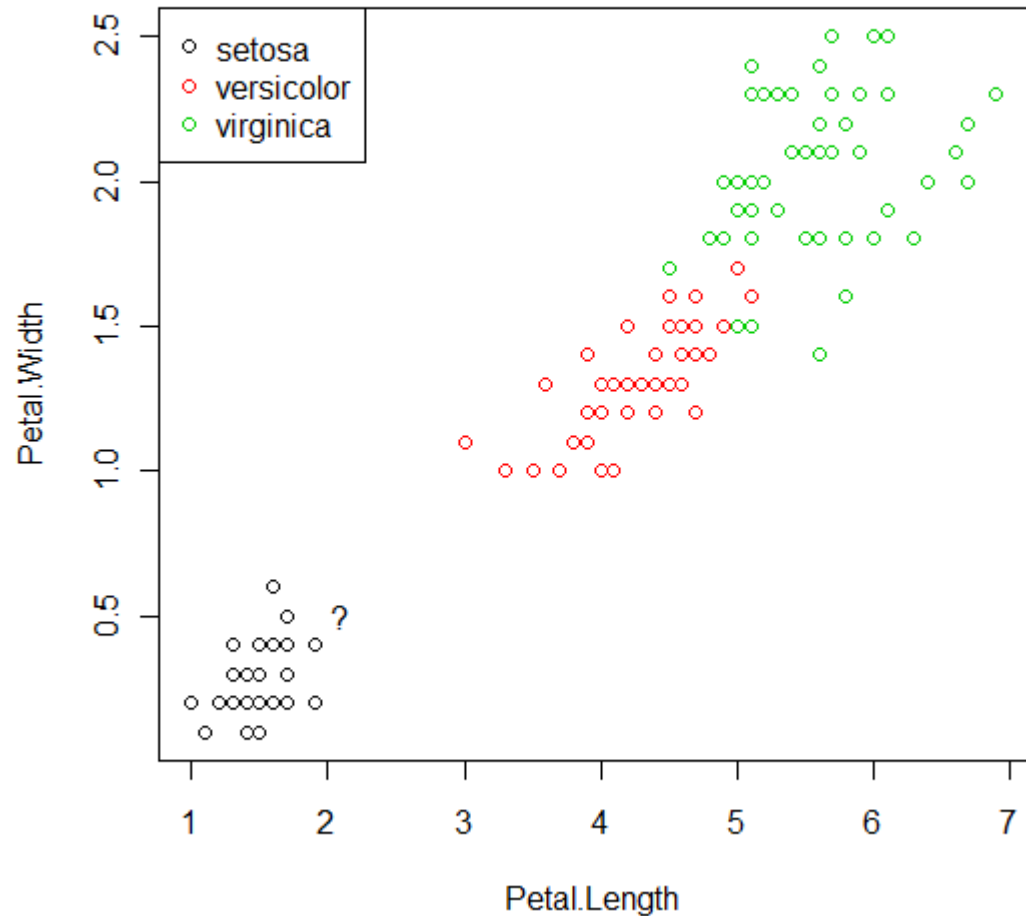
# Classification Methods

- Objective: Identify the class (category/label) of an individual (e.g., will purchase a specific product – Yes/No) based on observed features (e.g., gender, location, age, purchase history, etc)

- Classes: $c_1, c_2, ..., c_m$        Features: $x_1, ..., x_k$

- 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
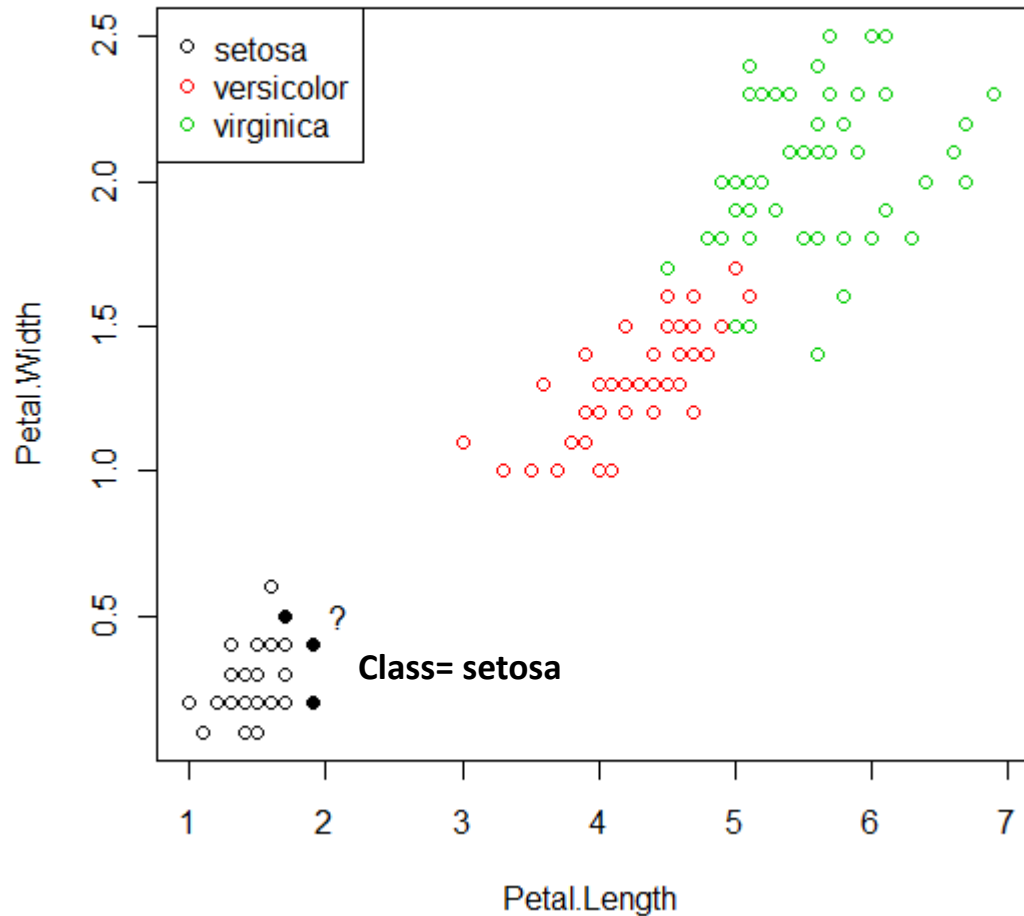
# Classification Methods: K-Nearest Neighbors (KNN)

- 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
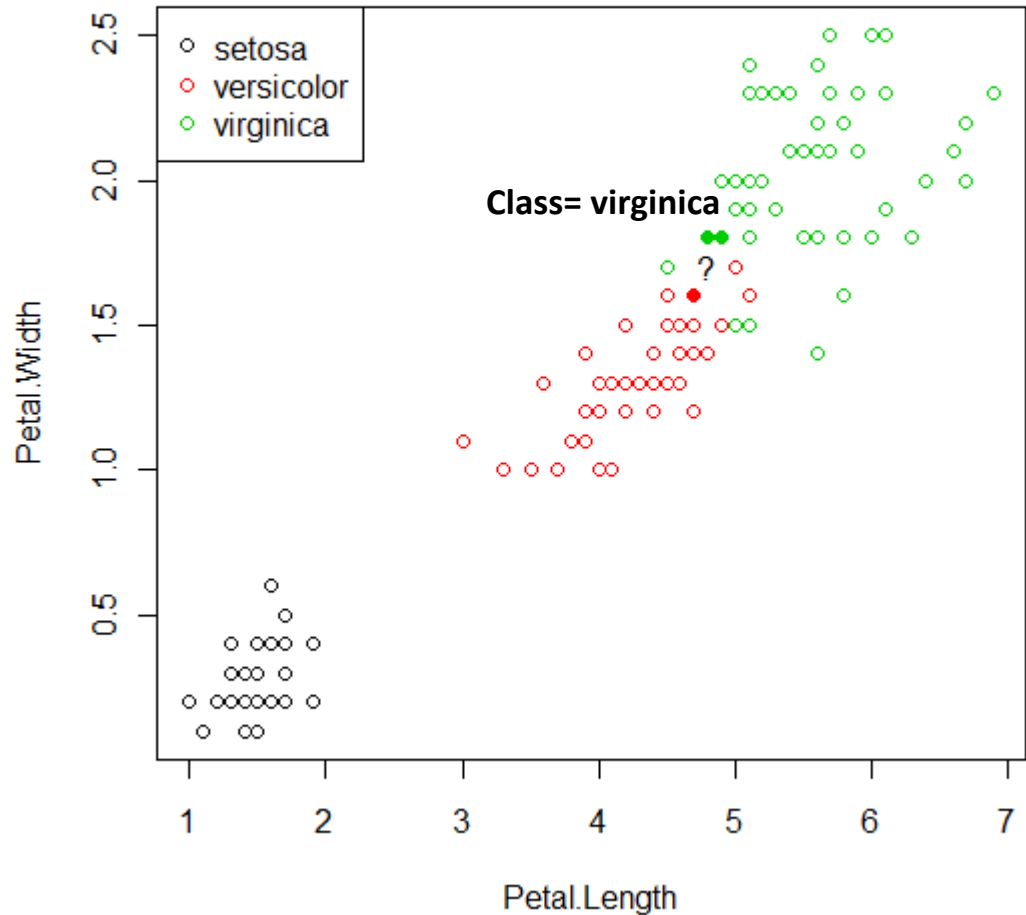
# Classification Methods: K-Nearest Neighbors (KNN)

- 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
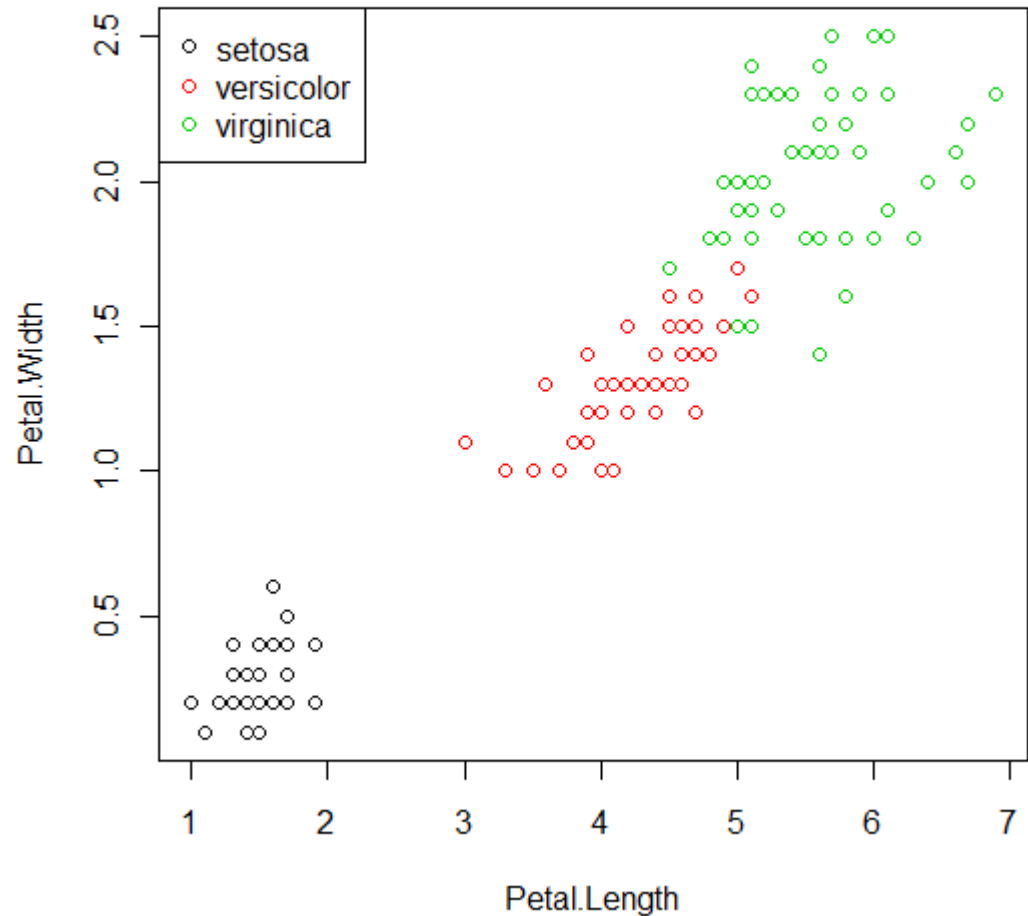
# Classification Methods: K-Nearest Neighbors (KNN)

- 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
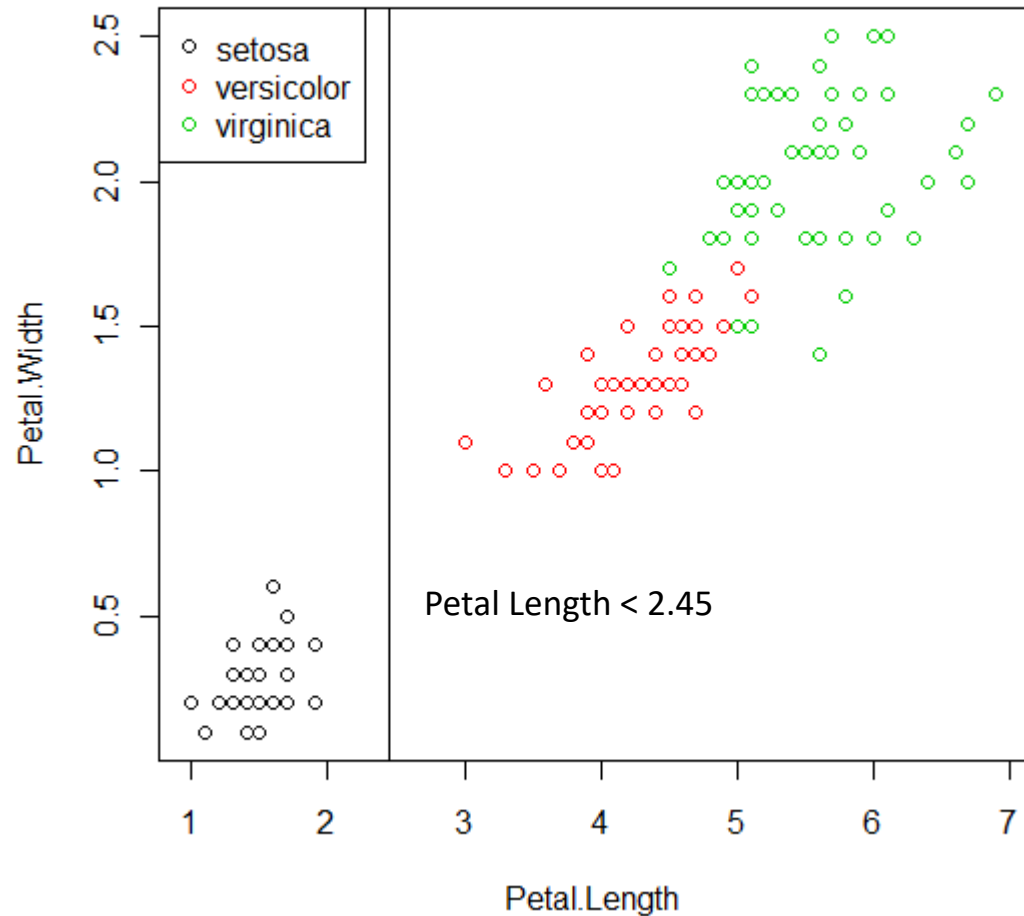
# Classification Methods: Decision Trees (DT)

- 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
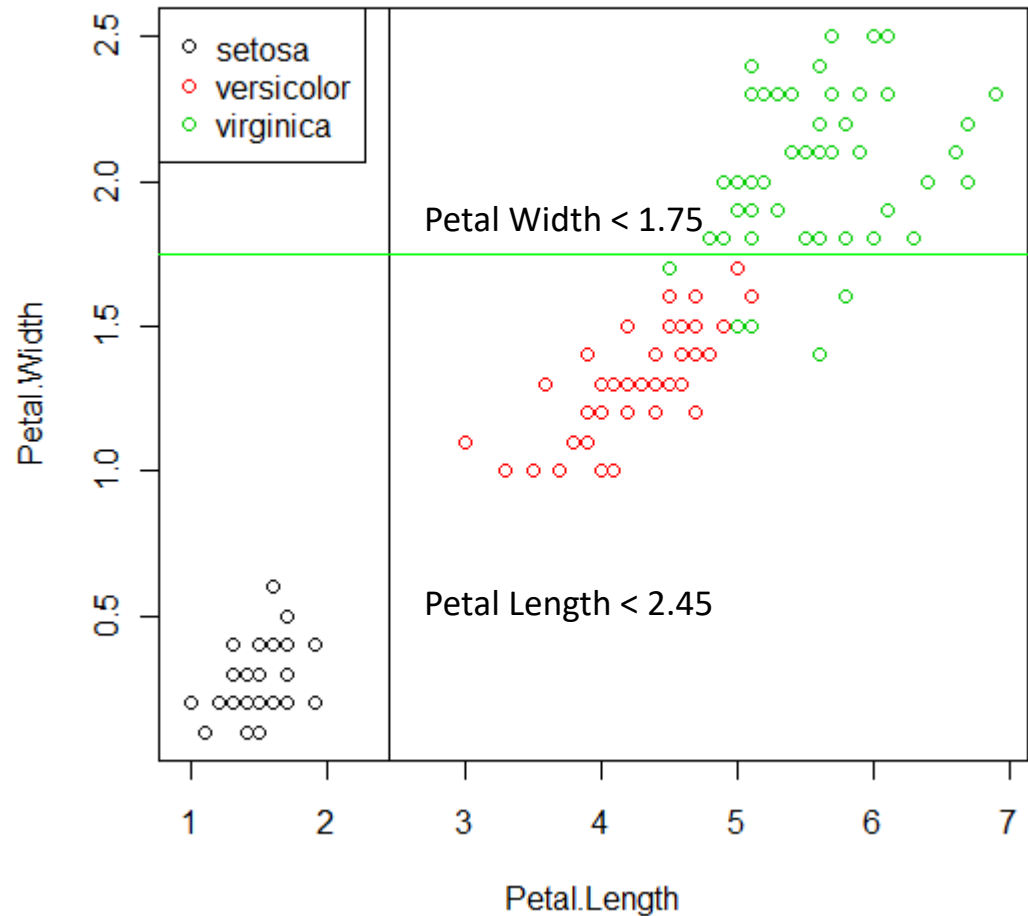- Each leaf corresponds to a class

# 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
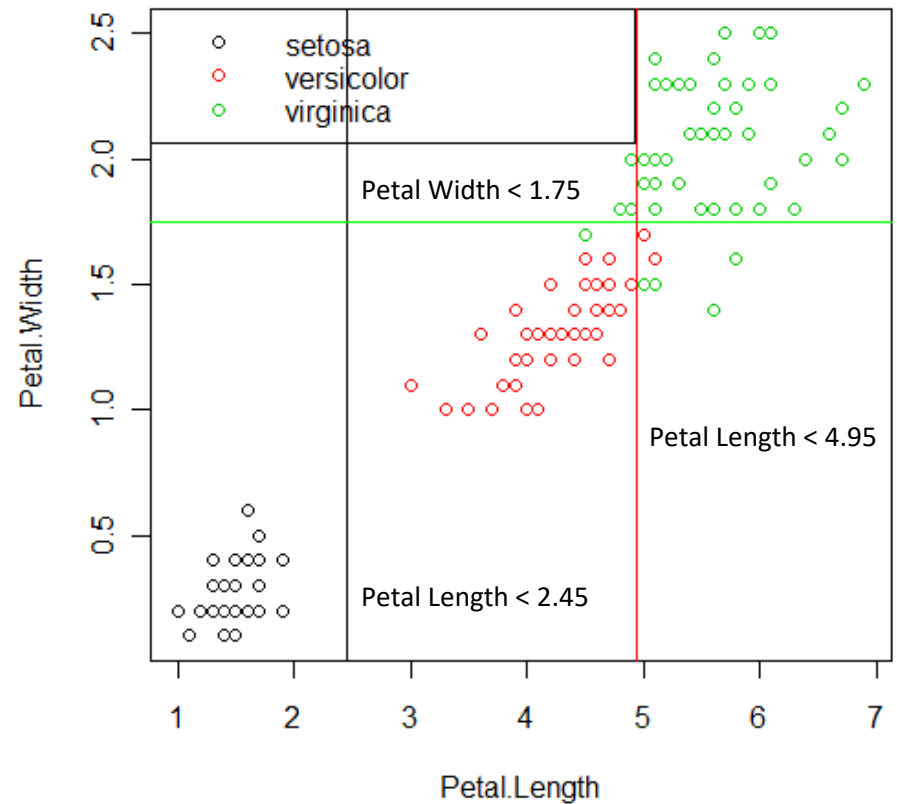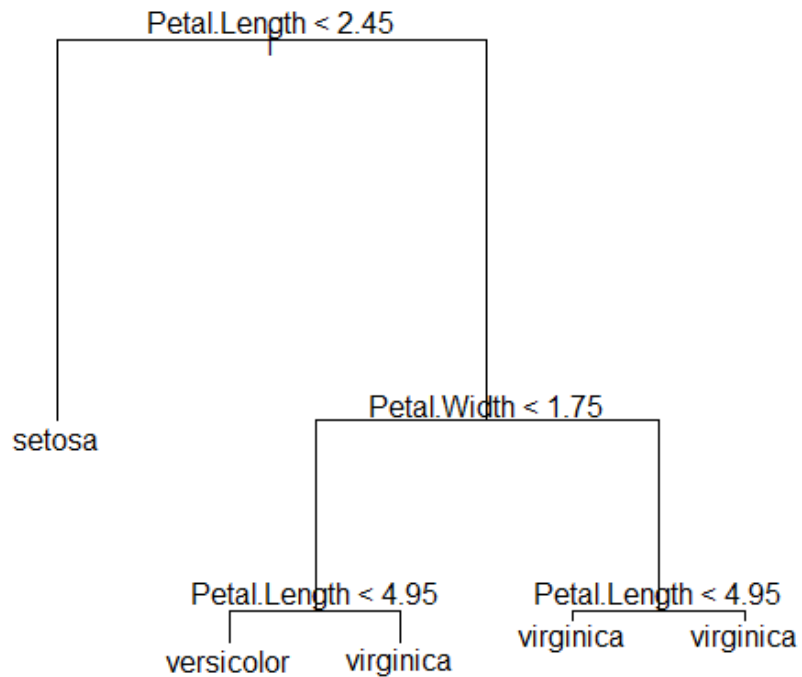- Each leaf corresponds to a class

# 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
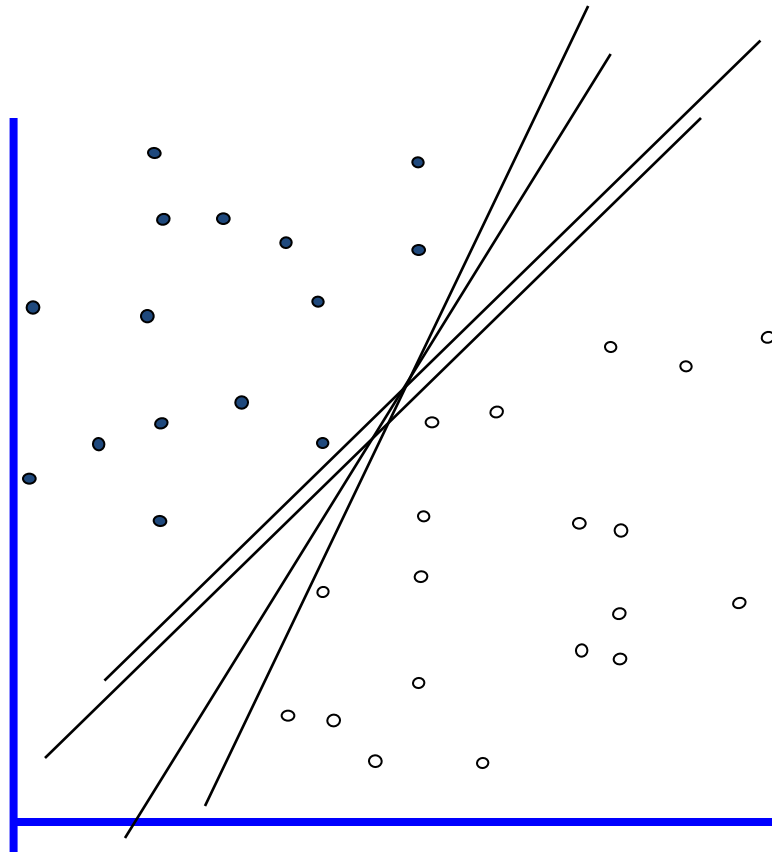- Each leaf corresponds to a class

# 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: 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

# Classification Methods: Support Vector Machines (SVM)

$$\Phi: \ \mathbf{x} \rightarrow \varphi(\mathbf{x})$$

Images modified from Mingyue Tan / Andrew Moore