Microarray and Gene Expression Analysis

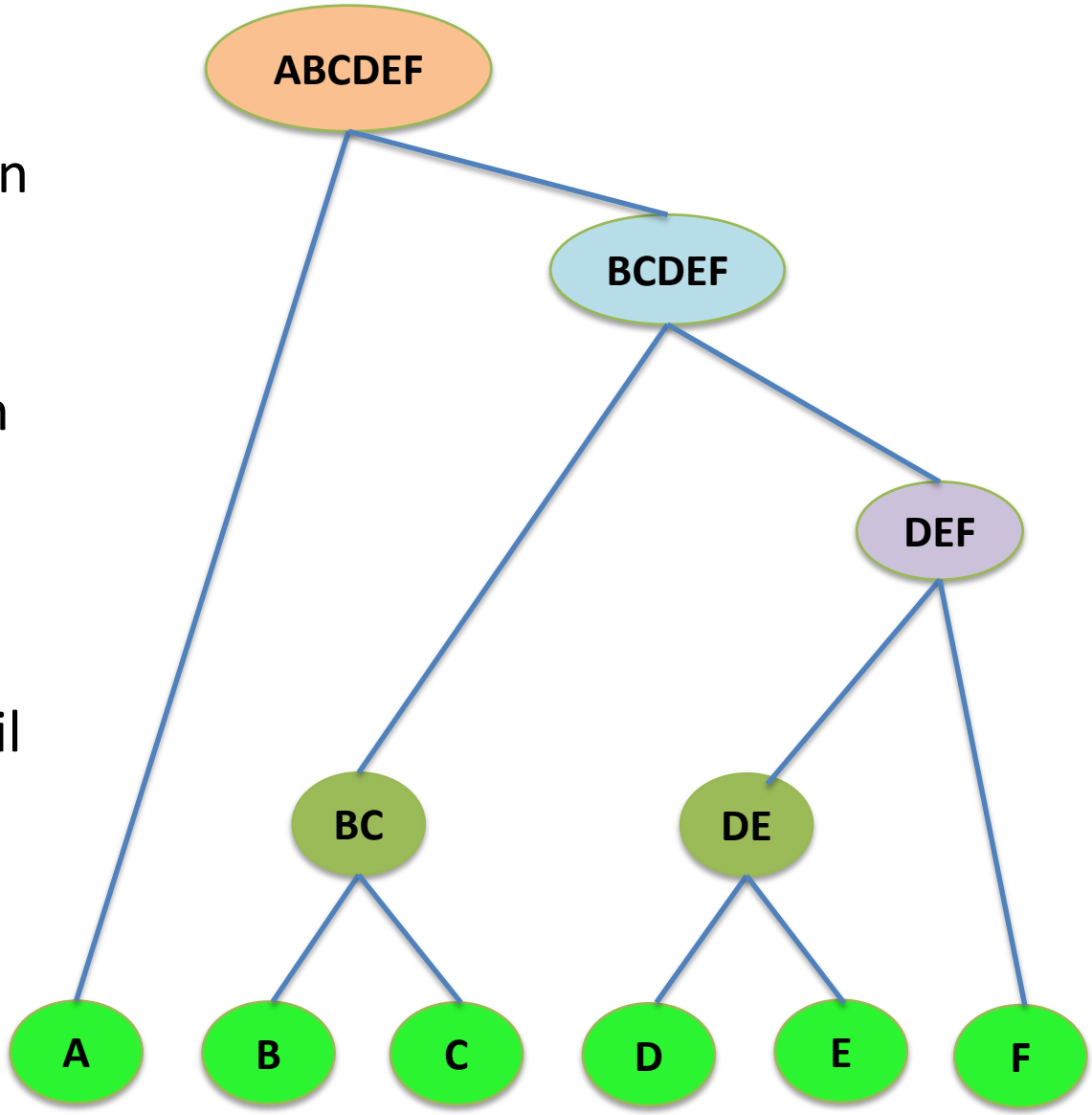
Garrett Dancik, PhD

Identification of Differentially Expressed Genes

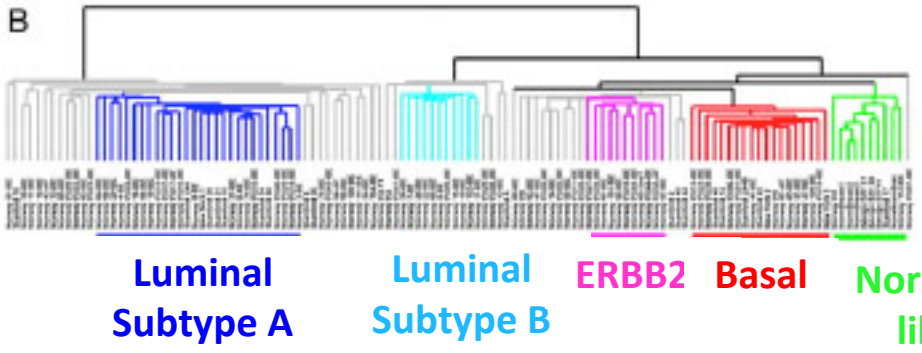
- Concern #1: Multiple comparison problem
  - Type I error probability (typically 5%) does not hold when you have multiple comparisons
  - If **no** genes are differentially expressed, and we analyzed 20,000 probes, there would be 1,000 false positives at significance level of 0.05!
  - In practice, p-values are adjusted to a false discovery rate (FDR, also called a q-value), which is the expected proportion of false positives in list of genes with adjusted p-values <= FDR
- Concern #2 (minor): Estimates of standard deviation are not stable (i.e., they are highly variable). Repeating the analysis using 1 more or less sample could produce different results.
- We will use the *limma* package in R which addresses both of these concerns

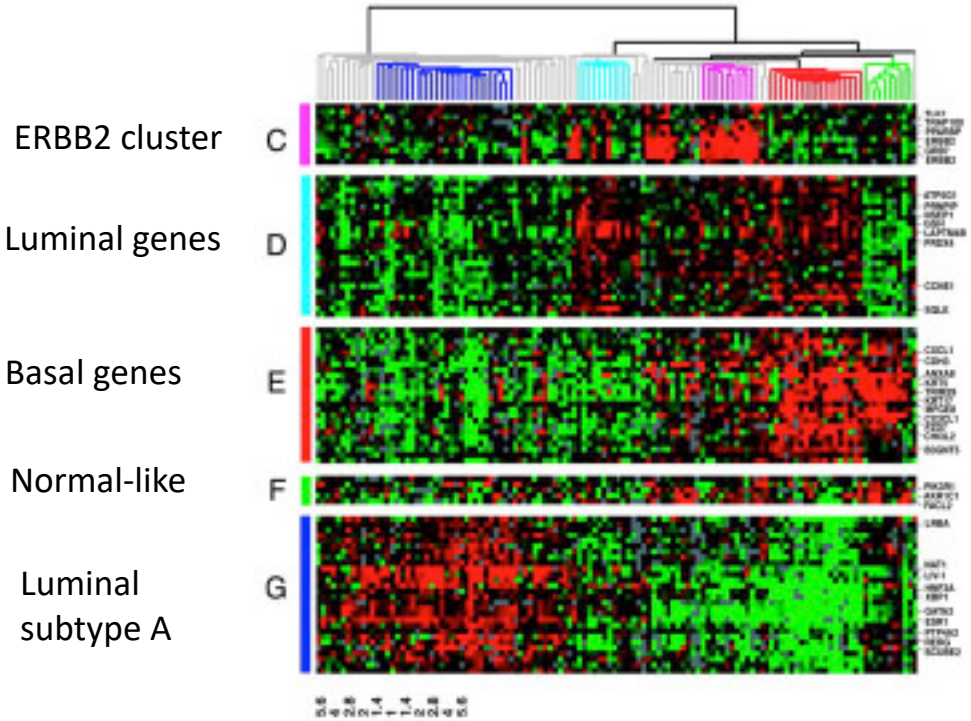
Hierarchical agglomerative ("bottom up") clustering groups samples by similarity

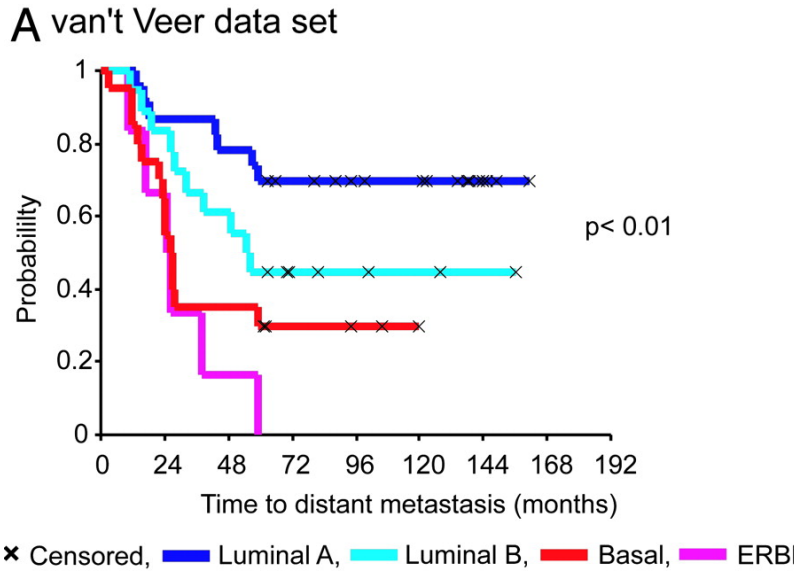
- Each observation starts in its own cluster
- Pairwise distances are calculated between each cluster
- The two most similar clusters are merged
- This process repeats until there is only one cluster

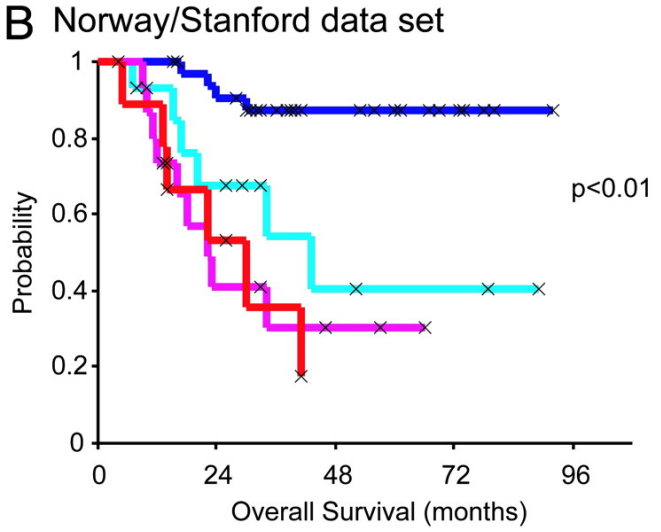


Hierarchical clustering of gene expression data identifies intrinsic breast cancer subtypes









Sørlie T et al. PNAS 2003;100:8418-8423

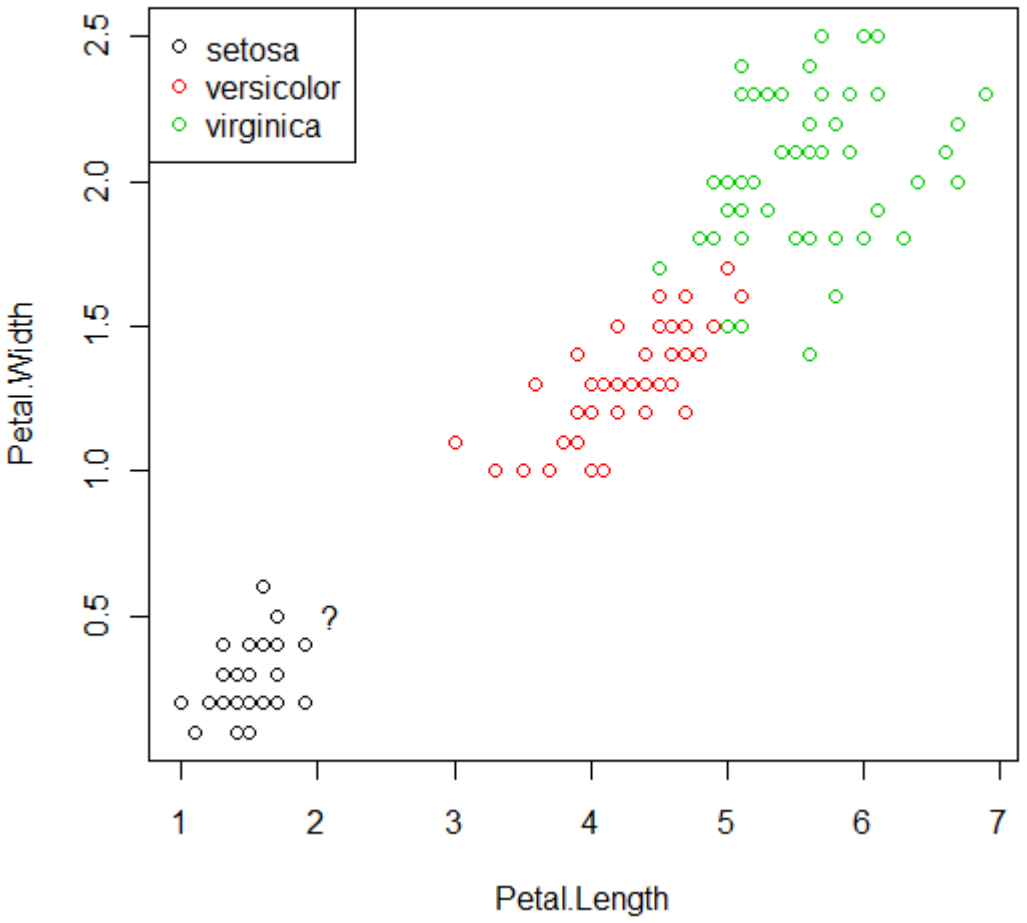
PNAS

Classification Methods

- Objective: Identify the class of an individual (e.g., male or female) based on observed features (e.g., gene expression levels)
- 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
    - Independent validation
    - Leave one out cross validation
    - K-fold cross validation
  - Classification / prediction of target data set

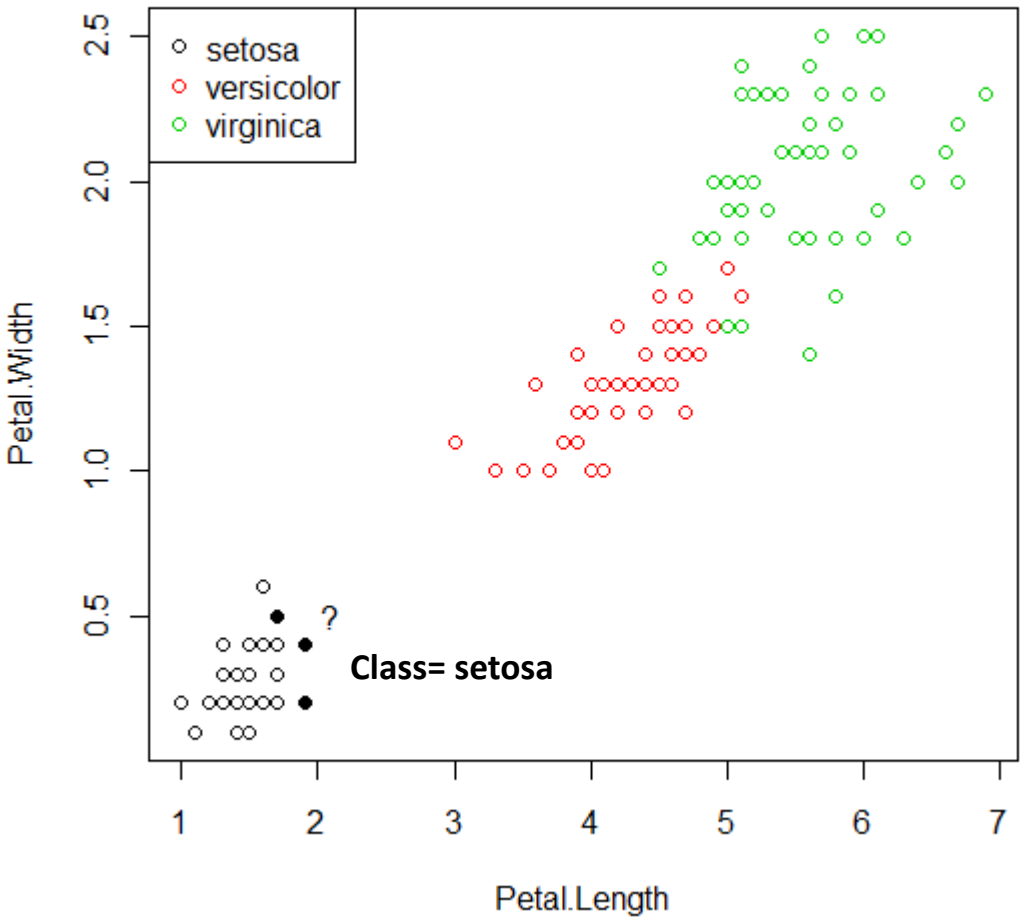
Classification Methods: K-Nearest Neighbors (KNN)

- 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



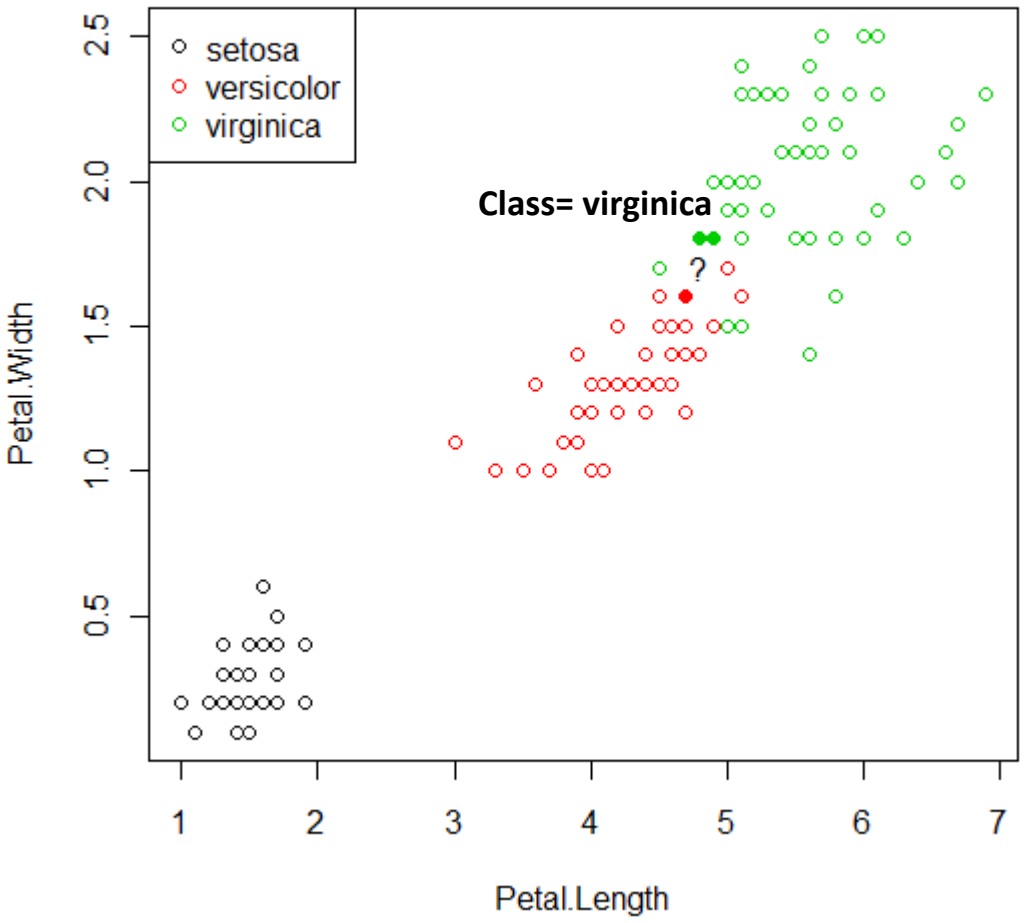
Classification Methods: K-Nearest Neighbors (KNN)

- 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



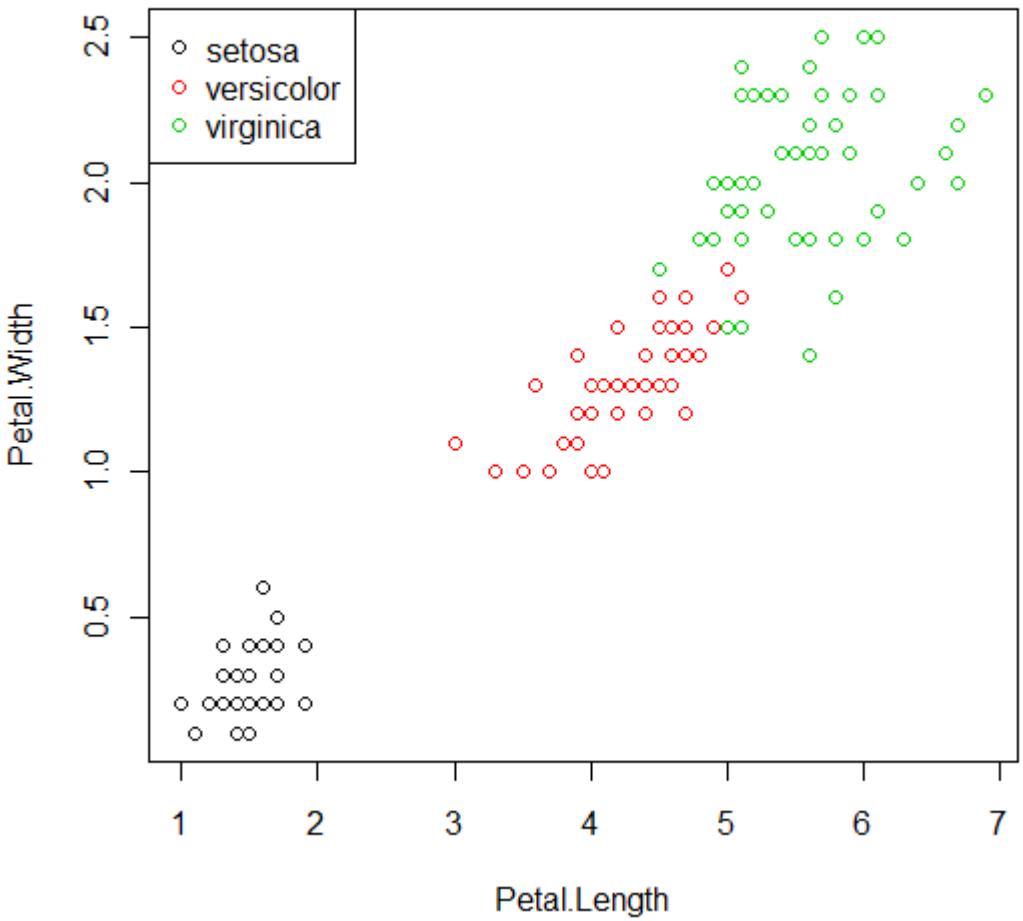
Classification Methods: K-Nearest Neighbors (KNN)

- 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



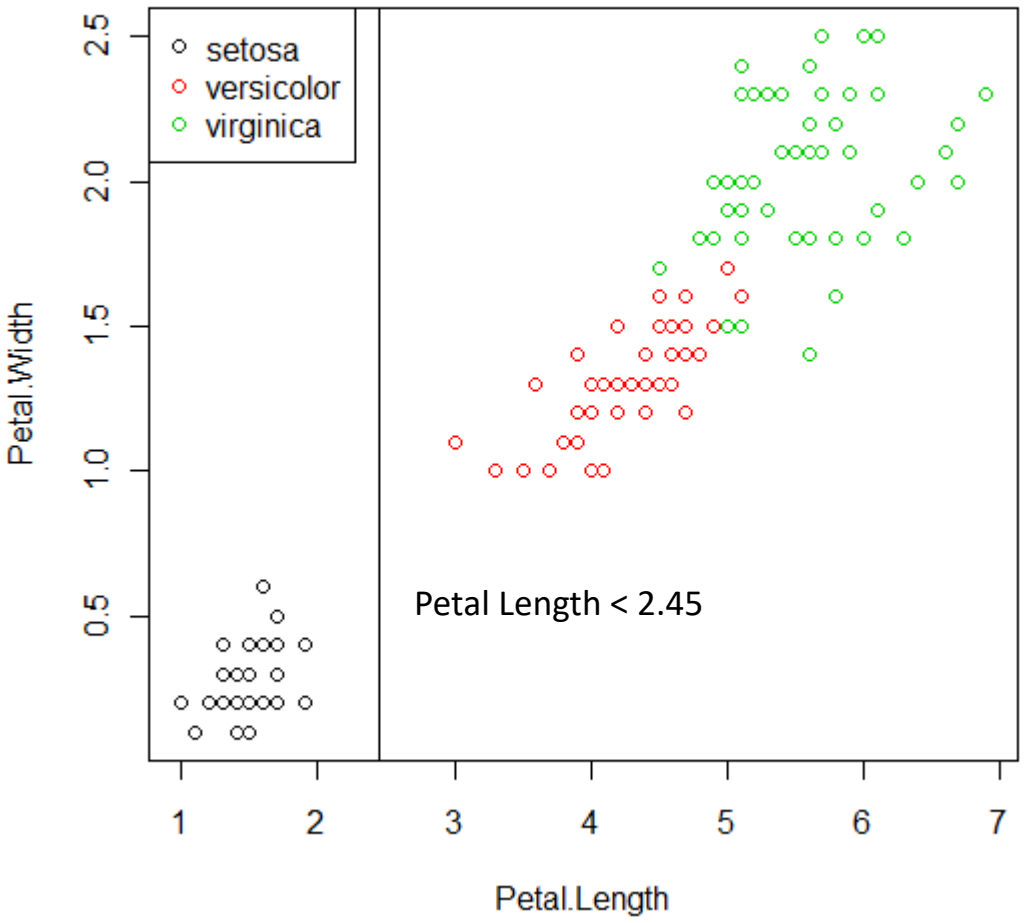
Classification Methods: Decision Trees (DT)

- 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



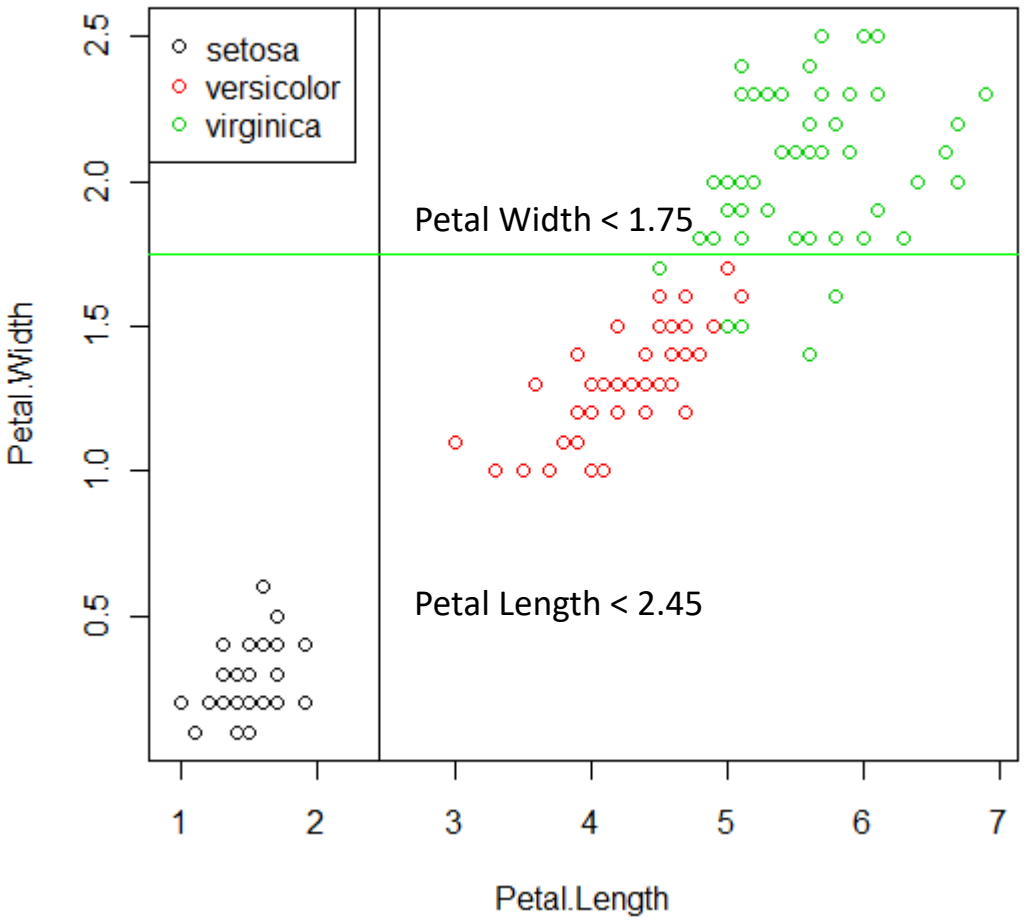
Classification Methods: Decision Trees (DT)

- 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

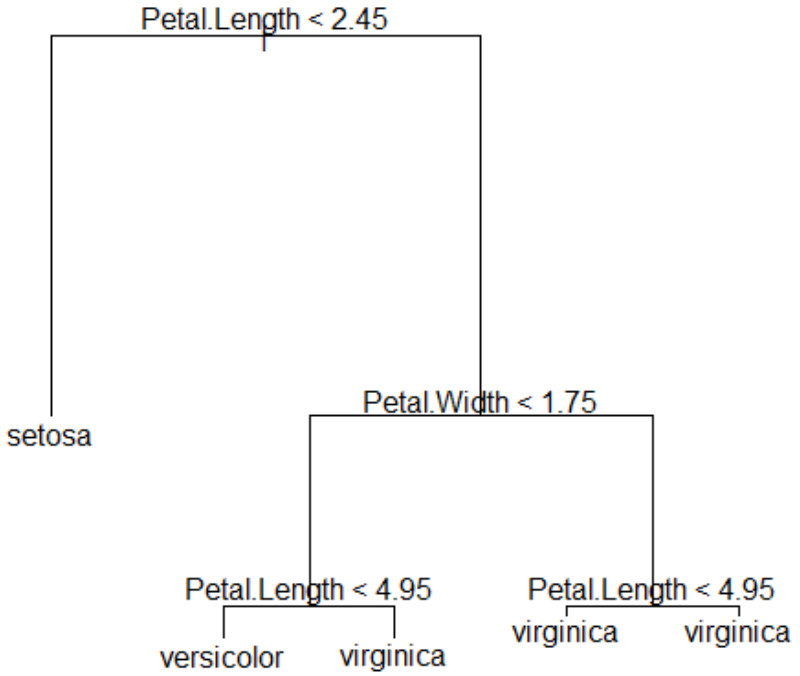


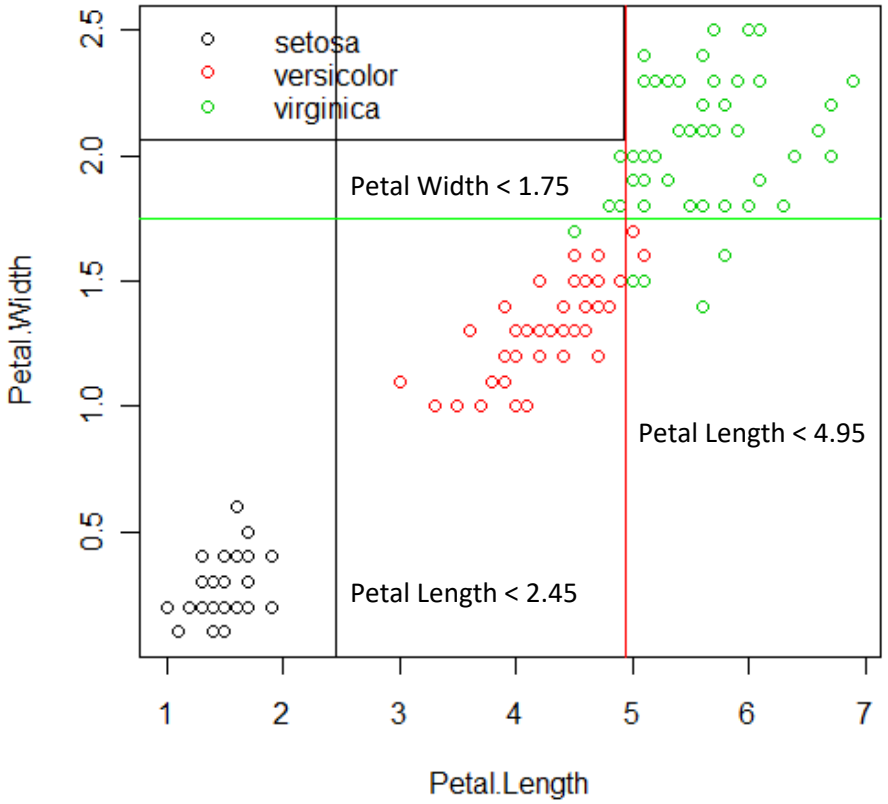
Classification Methods: Decision Trees (DT)

- 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

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

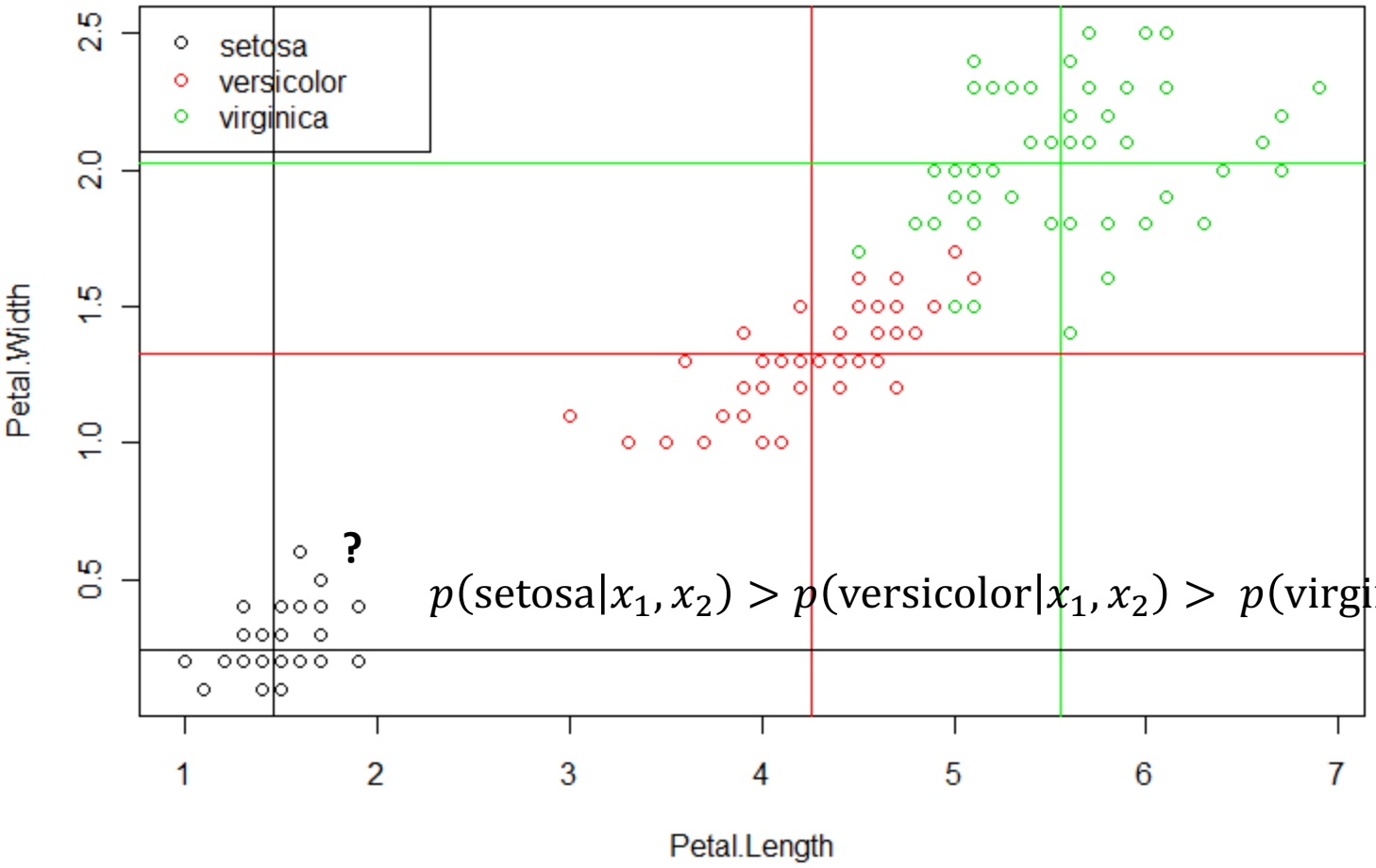
- Select the class *C* that maximizes

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

Classification Methods: Naïve Bayes (NB)

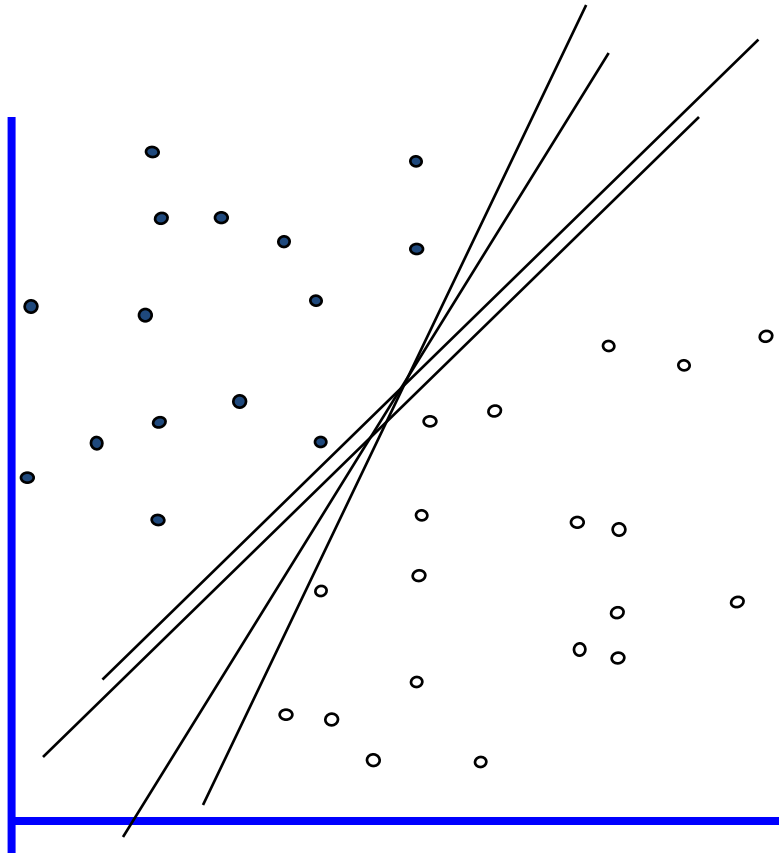
- Select the class *C* that maximizes

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



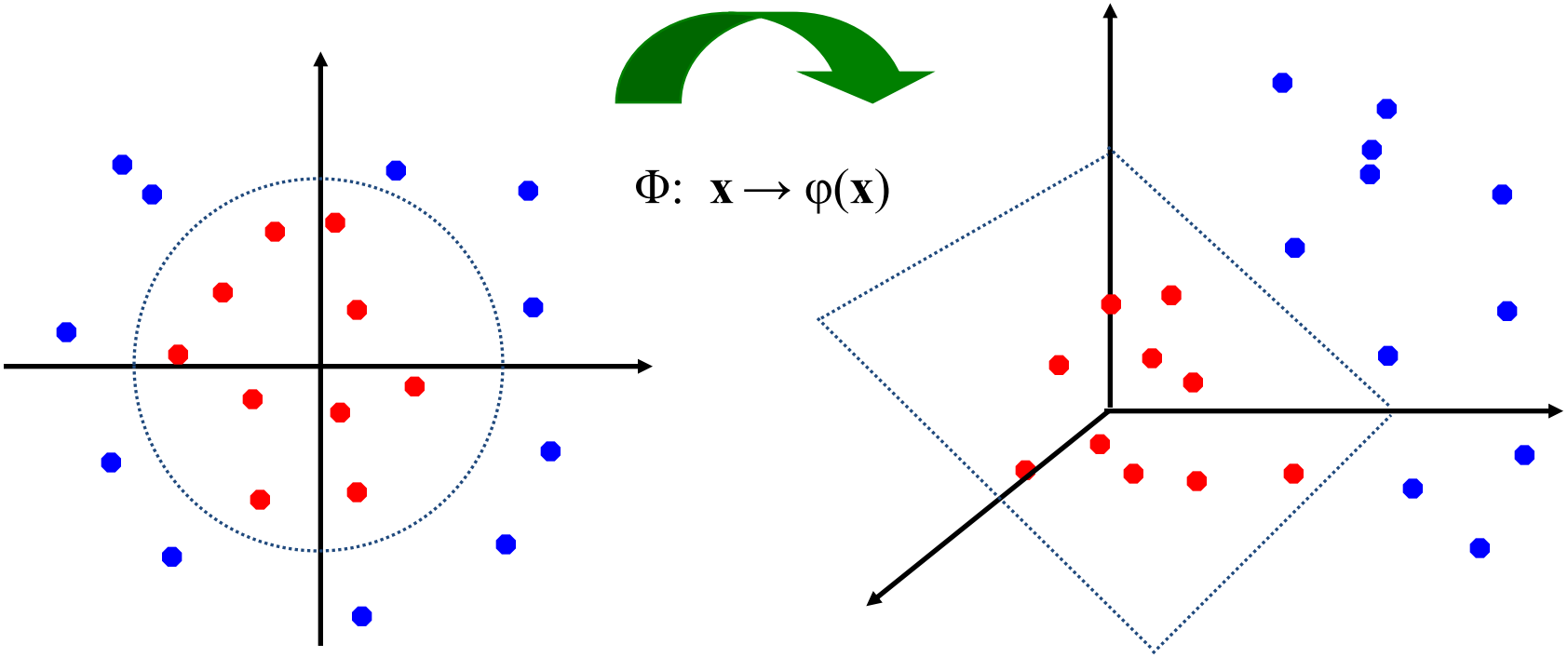
Classification Methods: Support Vector Machines (SVM)

- 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

Caveats and strategies

- Validation
  - Overfitting is often a problem: a classifier can perform very well on a training data set but may not generalize to additional data sets
  - Validation on independent data sets are ideal
  - Cross-validation is useful when data is limited
- Basic Strategy
  - Use cross-validation to select
    - The number of features (e.g., probes/genes)
    - Optimal parameters for classification model (e.g, value of *k* in knn)