Differential expression analysis of RNA-Seq data

Garrett Dancik, PhD

Data processing for between sample comparison

- In this example,
  - The sequencing depth (library size) of sample 2 is 1.5x that of sample 1
  - The expression of gene 3 in sample 2 is 2x as high as expression in sample 1
  - There is no difference in expression in gene 1 or gene 2

	Sample 1	Sample 2
Gene 1	10	15
Gene 2	20	30
Gene 3	10	30
N (library size)	40	75

What happens if we adjust for library size, as is the case for RPKM/FPKM?

- Divide sample 1 read counts by 40
- Divide sample 2 read counts by 75

(In RPKM/FPKM, we also multiply by 1 million, and scale each row by the gene length in kilobases; however neither of these impact the relative values of a gene across samples)

Data processing for between sample comparison

- In this example,
  - The sequencing depth (library size) of sample 2 is 1.5x that of sample 1
  - The expression of gene 3 in sample 2 is 2x as high as expression in sample 1
  - There is no difference in expression in gene 1 or gene 2

	Sample 1	Sample 2
Gene 1	.25	0.2
Gene 2	.50	0.4
Gene 3	.25	0.4
Total	1	1

Because only gene 3 is differentially expressed, an appropriate method would show a difference *only* in gene 3 across samples. However, with RPKM/FPKM we see that:

- All gene values are different across samples
- This is because "if a large number of genes are unique to, or highly expressed in, one experimental condition, the sequencing 'real estate' available for the remaining genes in that sample is decrease" (https://genomebiology.biomedcentral.com/articles/10.1186/gb-2010-11-3-r25)

Trimmed Mean of M values (TMM)*

- The fold change (FC) of a gene is the ratio of values across samples
  - e.g., FC of gene 1 is s2_value / s1_value = 15 / 10 = 1.5
  - If a gene's expression is consistent across samples, then FC ~ 1
- The goal is to make the FC between most genes as close to 1 as possible
  - (assumption is that most genes are not differentially expressed)

	Sample 1	Sample 2
Gene 1	10	15
Gene 2	20	30
Gene 3	10	30

Trimmed Mean of M values (TMM)*

- The fold change (FC) of a gene is the ratio of values across samples
  - e.g., FC of gene 1 is s2_value / s1_value = 15 / 10 = 1.5
  - If a gene's expression is consistent across samples, then FC ~ 1
- The goal is to make the FC between most genes as close to 1 as possible
  - (assumption is that most genes are not differentially expressed)

	Sample 1	Sample 2	M = FC*
Gene 1	10	15	15 / 10 = 1.5
Gene 2	20	30	30 / 20 = 1.5
Gene 3	10	30	30 / 10 = 3.0

The TMM normalization factors (library sizes) are calculated by taking a weighted average of the M values, but

- The M values are "trimmed" by 30% (we remove the largest 30% and lowest 30% of M values – we remove outlier genes or those that are differentially expressed)
- Genes with very high expression (top 5%) are also removed
- In this example, we ignore the M value of 3 and take the average of the others, which is 1.5. This is the normalization factor for sample 2.

*This is a simplification, but is enough to demonstrate the idea. For details, see the publication: https://genomebiology.biomedcentral.com/articles/10.1186/gb-2010-11-3-r25

Trimmed Mean of M values (TMM)

- We find the TMM values by dividing each column by its normalization factor

	Sample 1	Sample 2
Gene 1	10 / 1 = 10	15 / 1.5 = 10
Gene 2	20 / 1 = 20	30 / 1.5 = 20
Gene 3	10 / 1 = 10	30 / 1.5 = 20
Normalization factor	1	1.5

The findings indicate that:

- The expression of gene 1 is the same in both samples
- The expression of gene 2 is the same in both samples
- The expression of gene 3 is twice as high in sample 2 than it is in sample 1

With TMM normalization, we can accurately compare values across samples (even though the library size for sample 2 was 1.5x the library size of sample 1)

Identification of Differentially Expressed Genes

- Concerns:
  - 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 genes, 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
  - Reliable and robust estimates of standard deviation
    - Repeating the analysis using just one more or one less sample could produce very different results.
- We will use the *limma* package in *R* which addresses both of these concerns

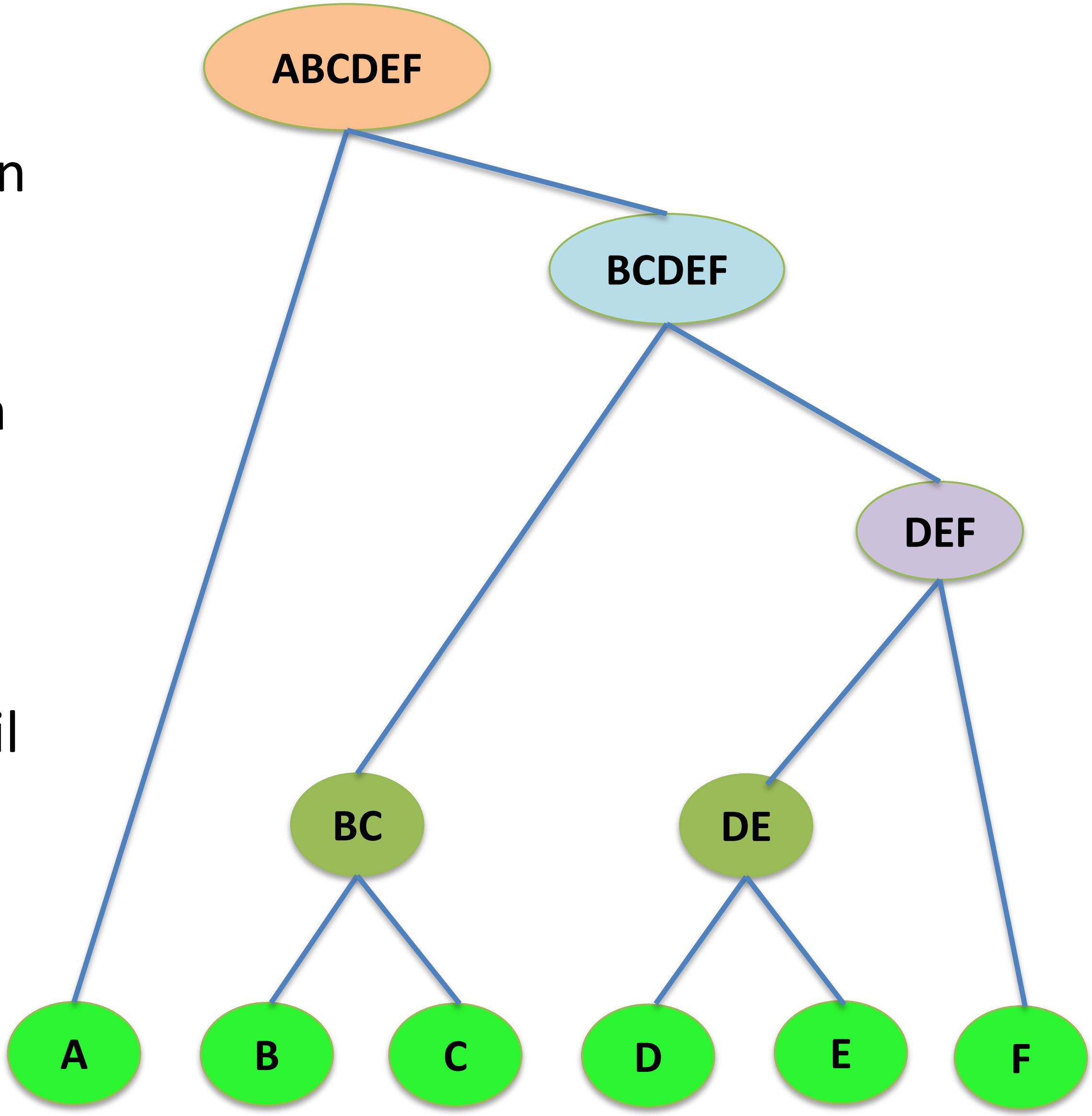
Limma: Linear Models for Microarray and RNA-Seq Data

- Limma uses a linear model to model expression data and tests for statistical significance using a *moderated t-test:*
  - $log2\ cpm = \beta_0 + \beta_1 x_1 + \ \ldots..$
  - TMM normalization is recommended, prior to calculating log2 cpm values
  - x_1 is a coded variable denoting group membership (e.g., tumor vs normal)
    - But to understand this, let's look at *contrasts.R*

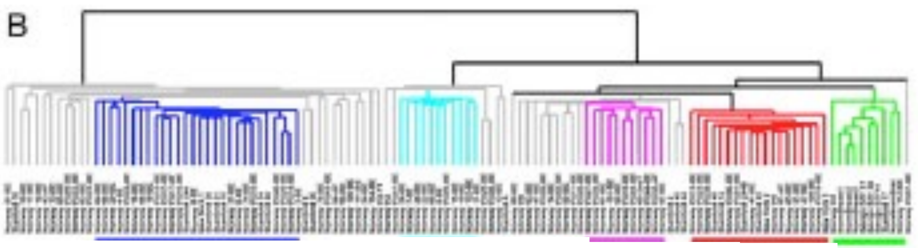
- User guide:
  - https://www.bioconductor.org/packages/devel/bioc/vignettes/limma/inst/doc/usersguide.pdf (we will follow Chapter 15: RNA-Seq data)
- Publications:
  - Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015 Apr 20;43(7):e47. doi: 10.1093/nar/gkv007. Epub 2015 Jan 20. PMID: 25605792; PMCID: PMC4402510. https://pubmed.ncbi.nlm.nih.gov/25605792/
  - Law, C.W., Chen, Y., Shi, W. *et al.* voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol* **15,** R29 (2014). https://doi.org/10.1186/gb-2014-15-2-r29

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



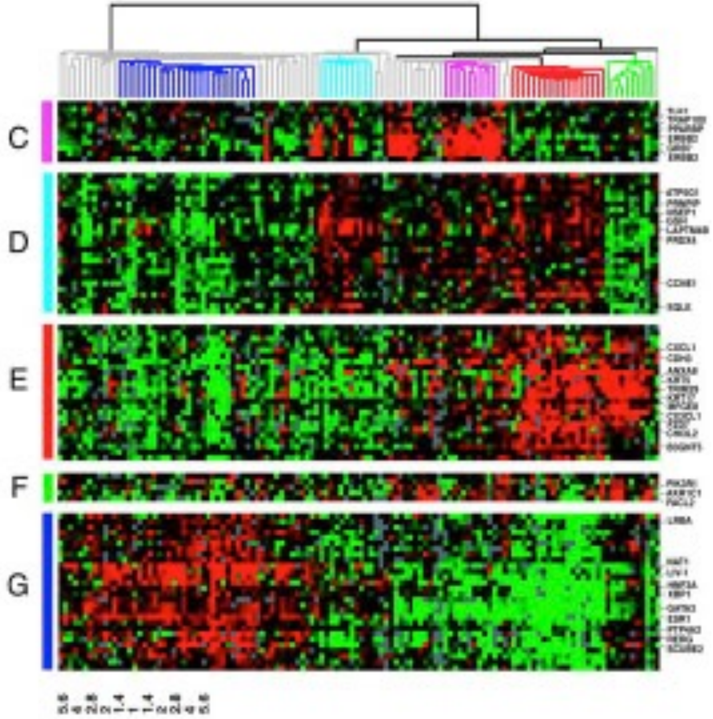
Luminal Subtype A

Luminal Subtype B

ERBB2

Basal

Normal-like



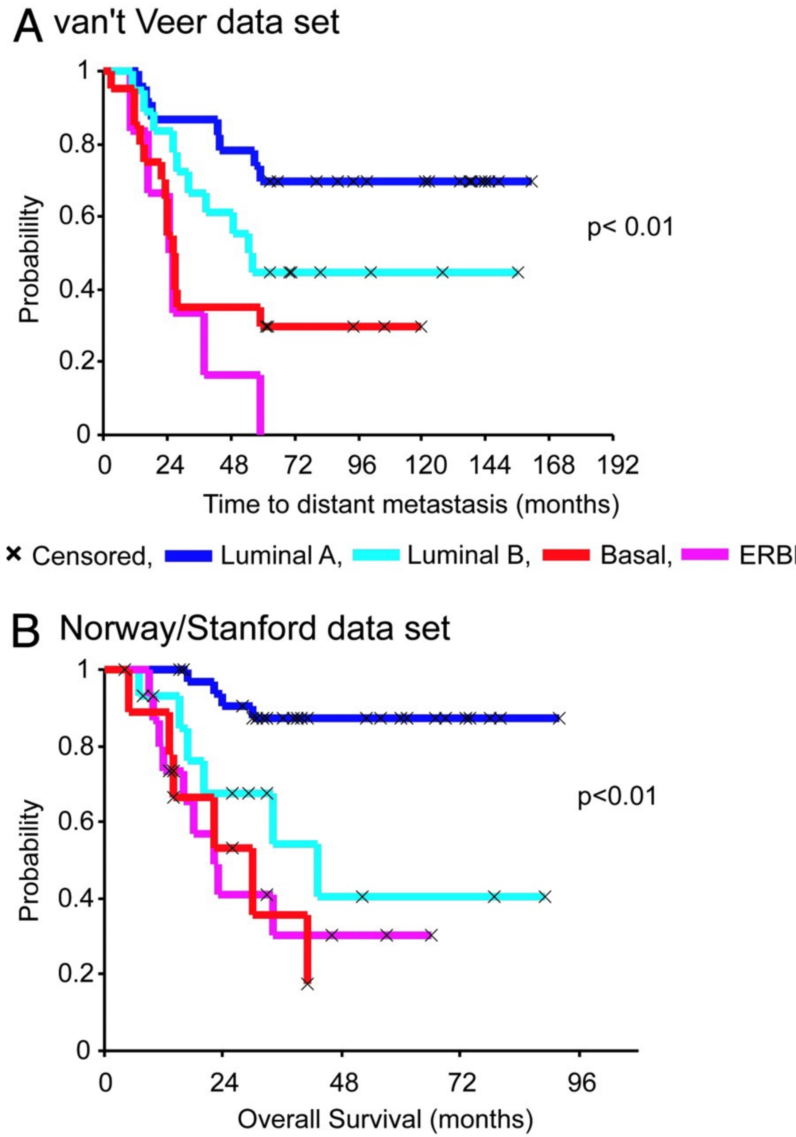
ERBB2 cluster

Luminal genes

Basal genes

Normal-like

Luminal subtype A



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

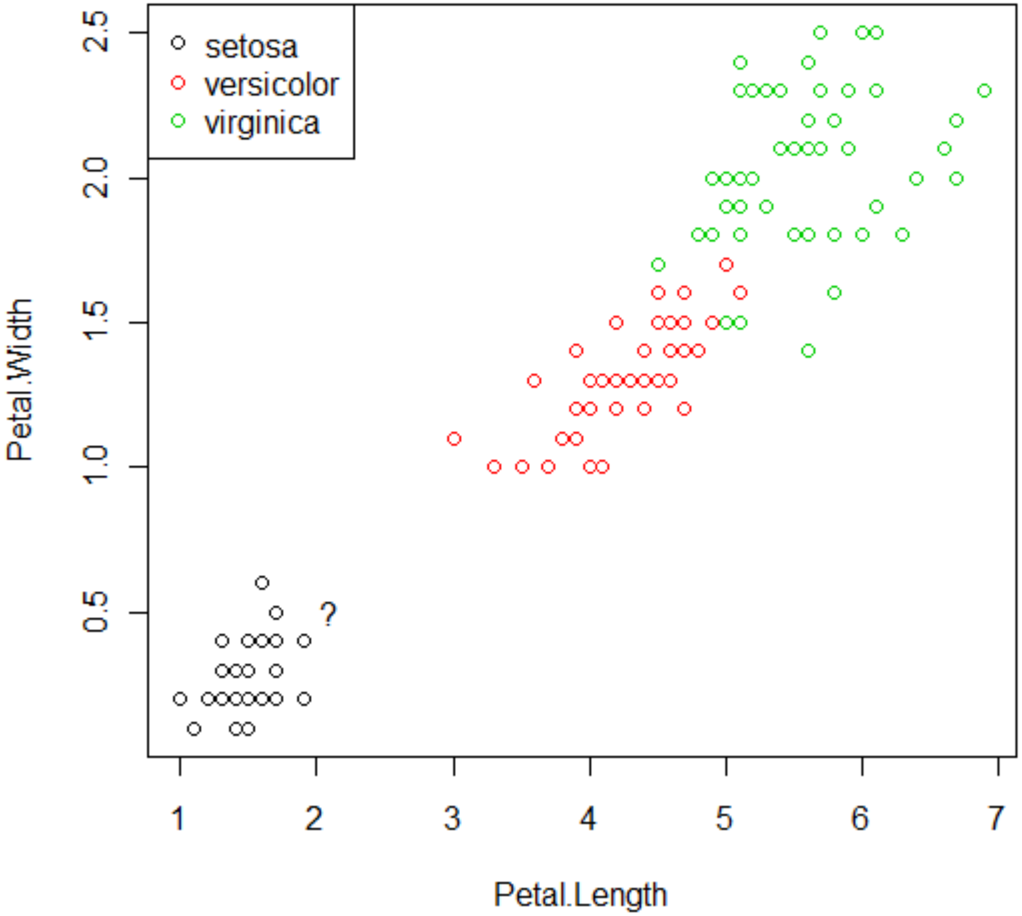
PNAS

Classification Methods

- Objective: Identify the class of an individual (e.g., male or female) based on observed features (e.g., gene expression levels)
- Classes: $c_1, c_2, \ldots, c_\mathrm{m}$ 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
    - Independent validation
    - Leave one out cross validation
    - K-fold cross validation
  - Classification / prediction of target data set

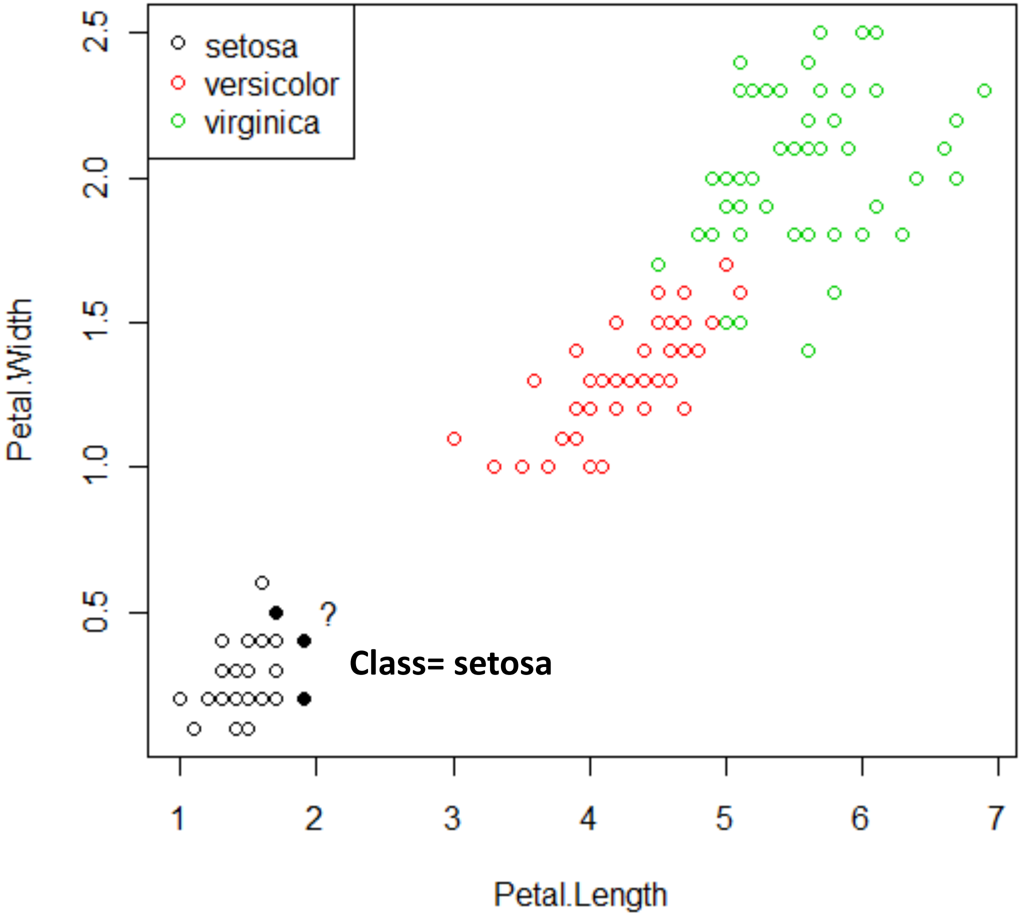
Classification Methods: K-Nearest Neighbors (KNN)

- 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



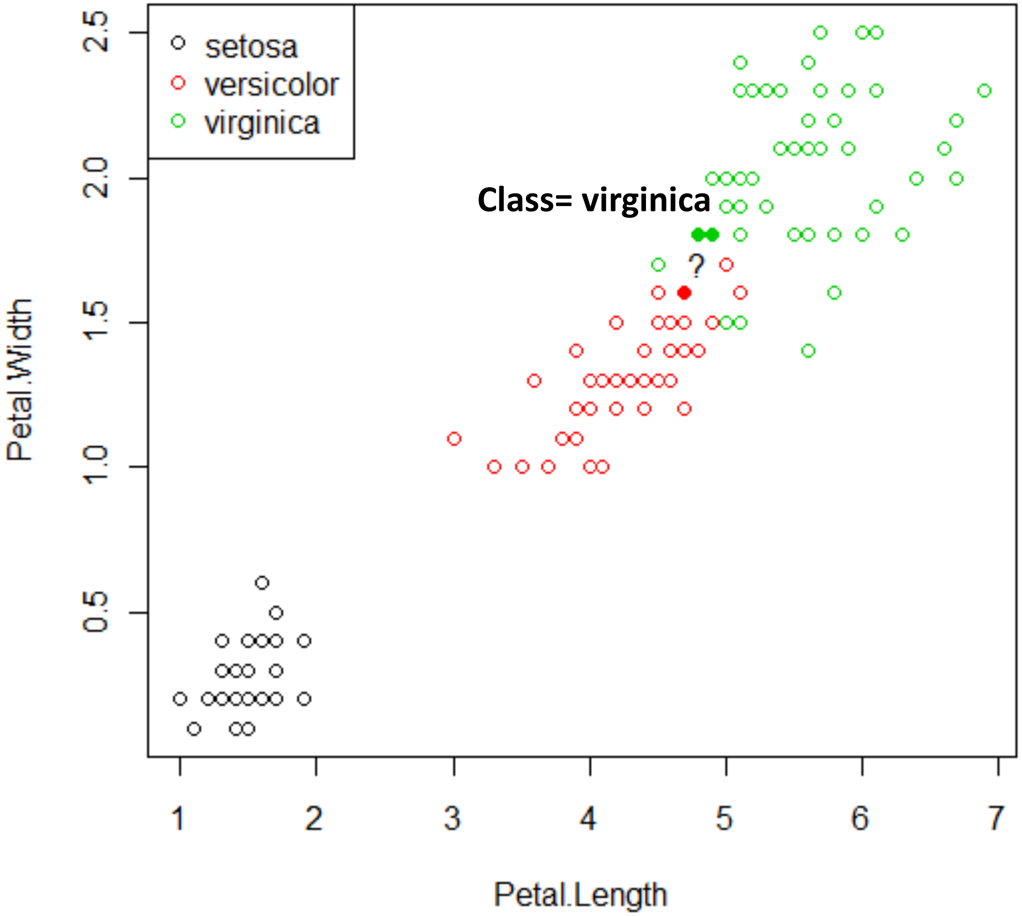
Classification Methods: K-Nearest Neighbors (KNN)

- 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



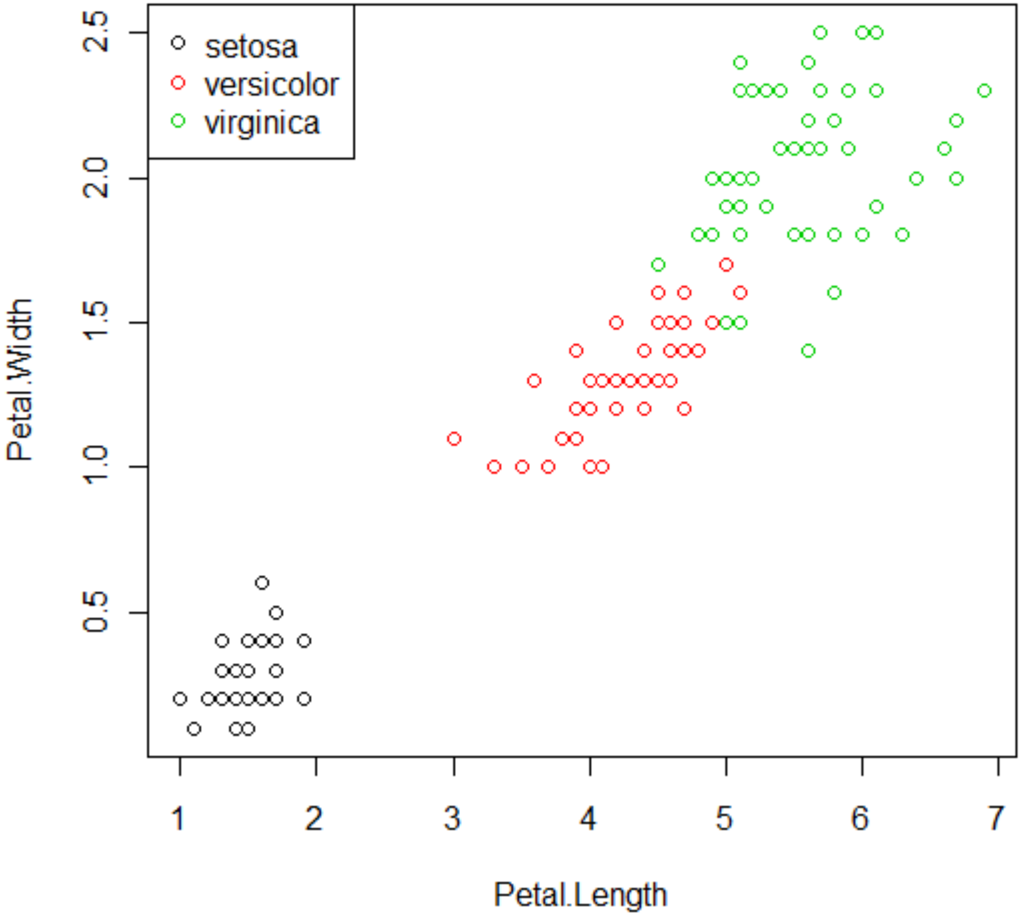
Classification Methods: K-Nearest Neighbors (KNN)

- 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



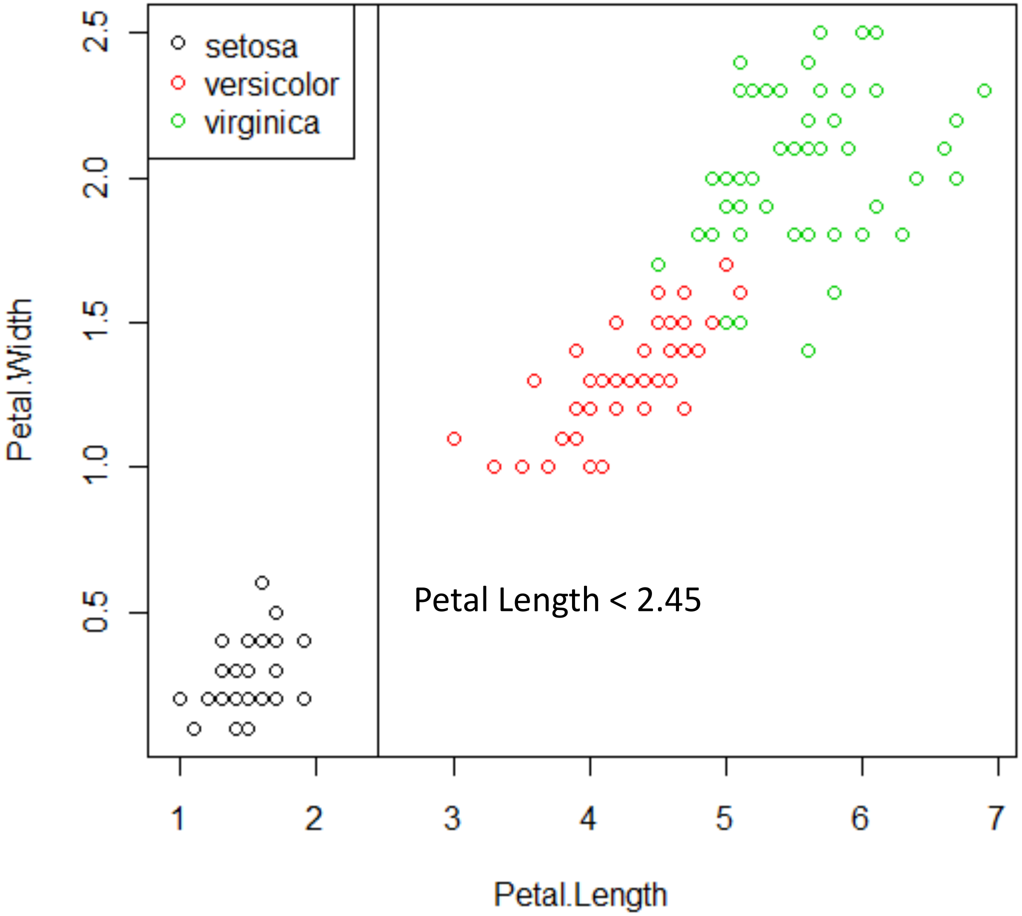
Classification Methods: Decision Trees (DT)

- 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



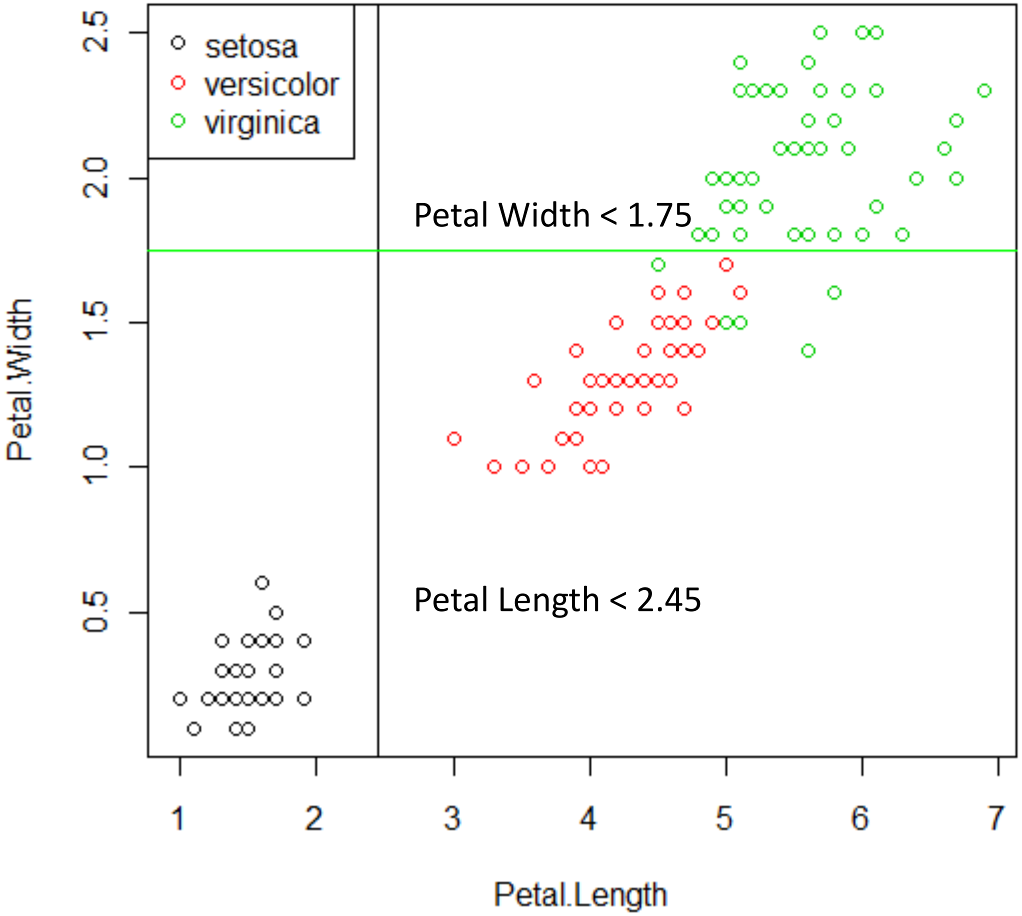
Classification Methods: Decision Trees (DT)

- 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

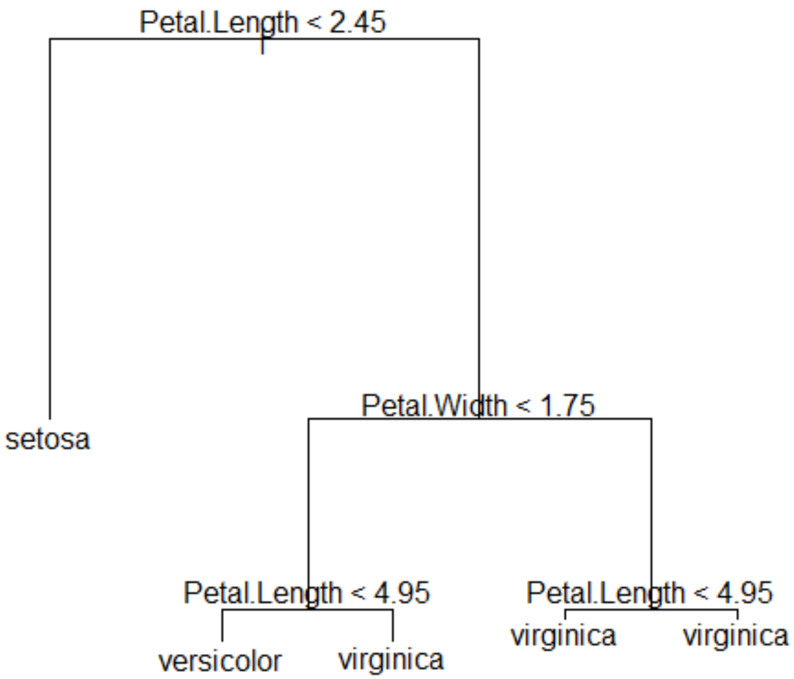


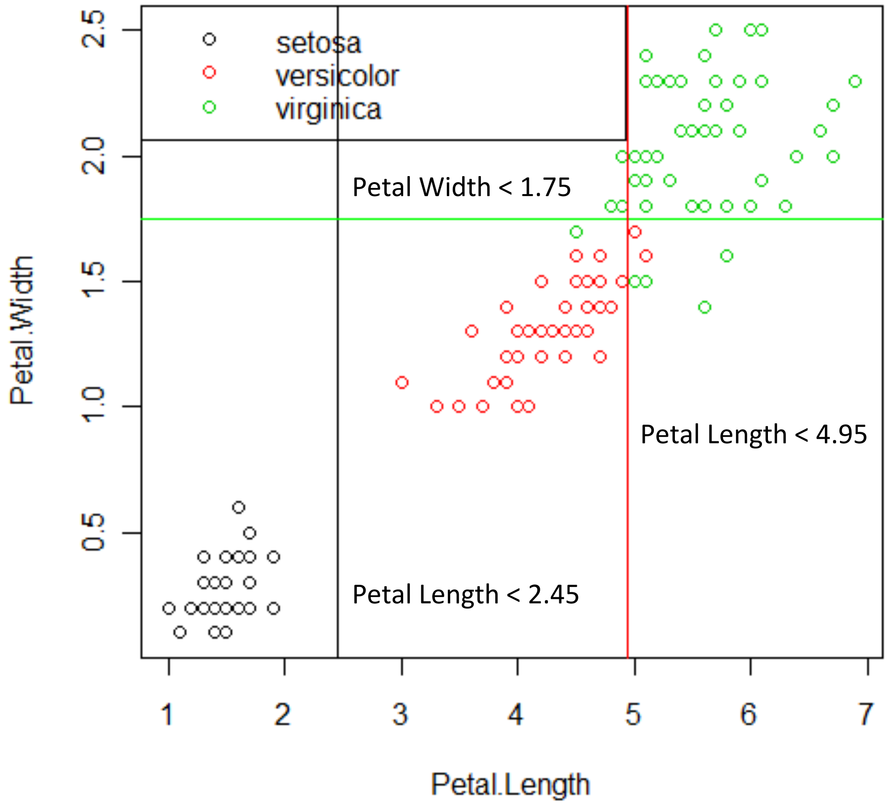
Classification Methods: Decision Trees (DT)

- 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

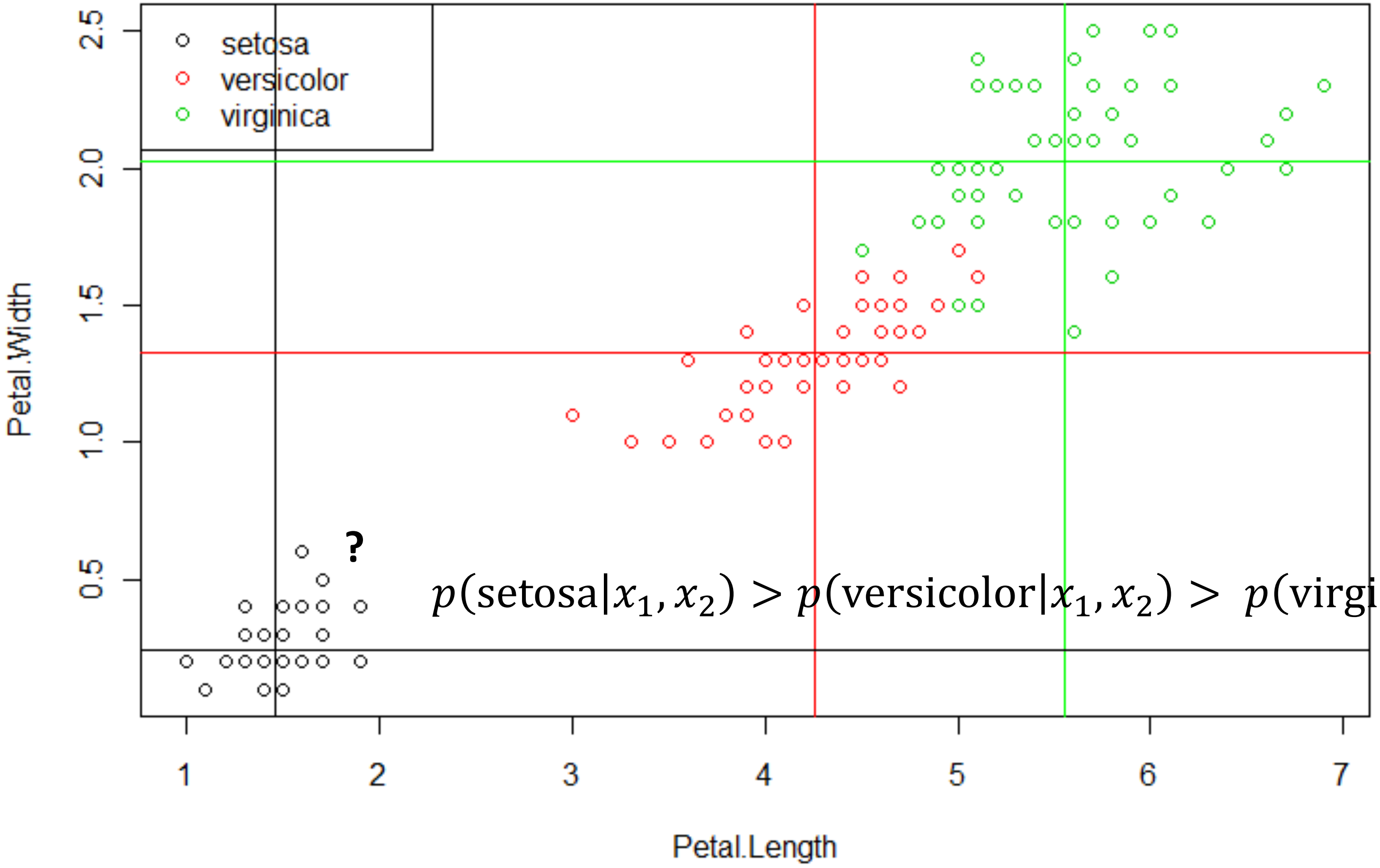
Classification Methods: Naïve Bayes (NB)

- Based on Bayes' theorem that relates conditional probabilities
$$p(C|x_1, \dots, x_n) \propto p(x_1, \dots, x_n|C)p(C)$$
- Naïve Bayes assumes independence of features, so that
$$p(x_1, \dots, x_n|C) \propto p(x_1|C)\times \cdots \times p(x_n|C)p(C)$$
- 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

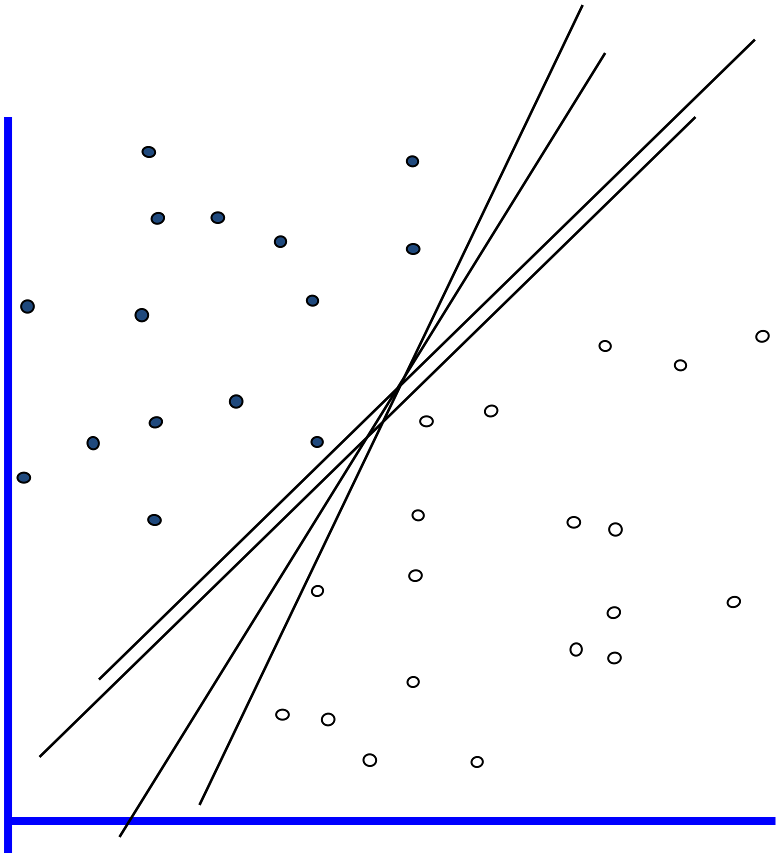
$$p(C|x_1,\dots,x_n) \propto p(x_1|C)\times\cdots\times p(x_n|C)p(C)$$



$$p(\text{setosa}|x_1,x_2) > p(\text{versicolor}|x_1,x_2) > \ p(\text{virginica}|x_1,x_2)$$

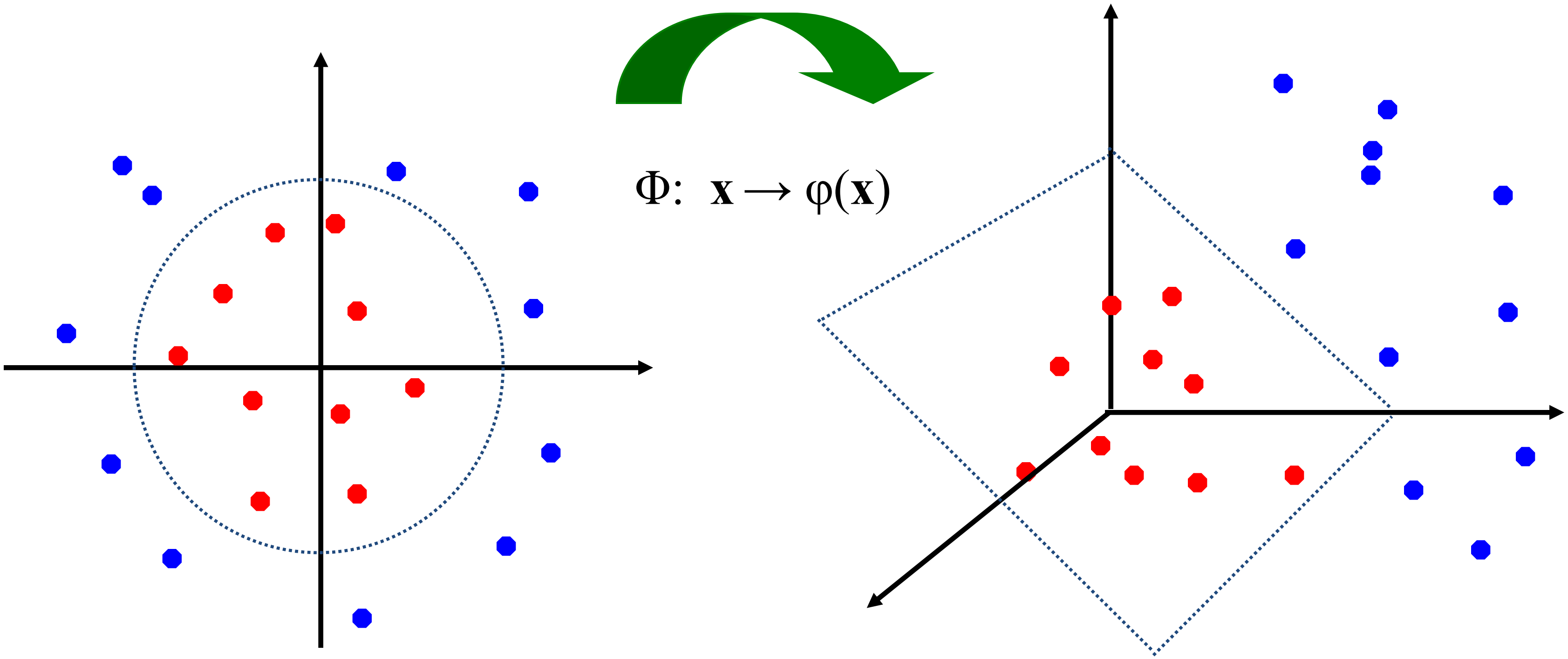
Classification Methods: Support Vector Machines (SVM)

- 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)



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

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)