Classification

Course: Analytics, Machine Learning, and the Digital Economy

- **Lecturer Radjabova Dilnora**

The Classification Setting

- For a classification problem, we can use the misclassification error rate to assess the accuracy of the machine learning method.

$$Error\ Rate = \sum_{i=1}^{n} I(y_i \neq \hat{y}_i)/n$$

which represents the fraction of misclassifications.

- $I(y_i \neq \hat{y}_i)$ is an indicator function, which will give 1 if the condition $(y_i \neq \hat{y}_i)$ is correct, otherwise it gives a 0.

Bayes Error Rate

- The Bayes error rate refers to the lowest possible error rate that could be achieved if somehow we knew exactly what the “true” probability distribution of the data looked like.

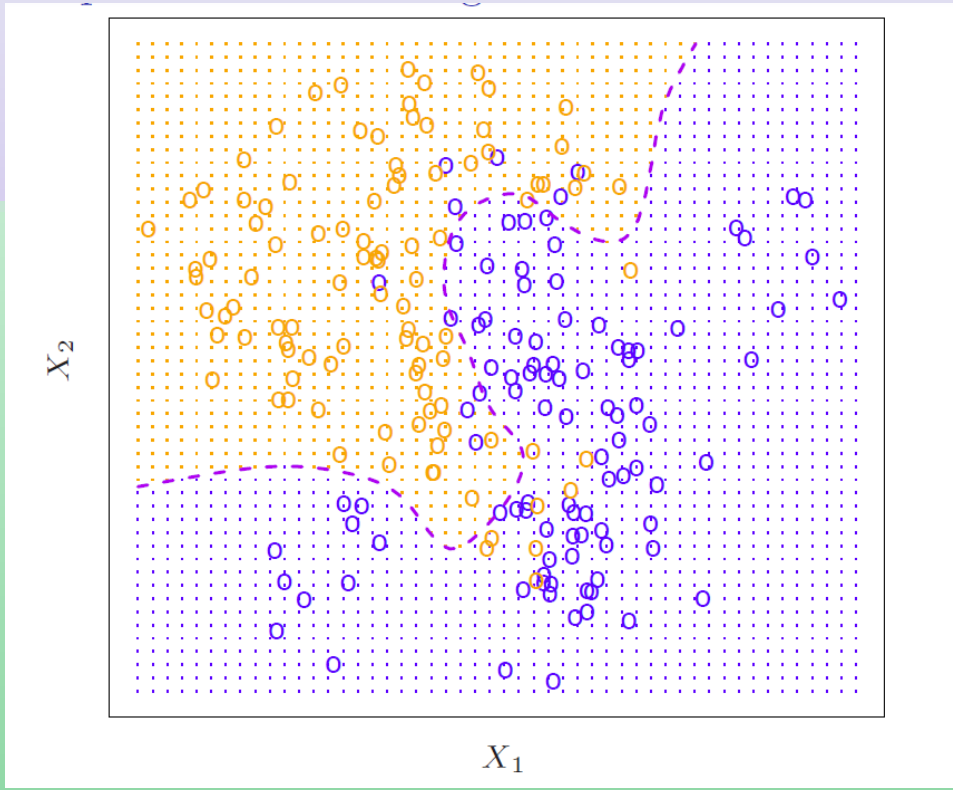
Bayes Error Rate

- On test data, no classifier can get lower error rates than the Bayes error rate.

Bayes Error Rate

- In real-life problems, the Bayes error rate can’t be calculated exactly.

Bayes Decision Boundary

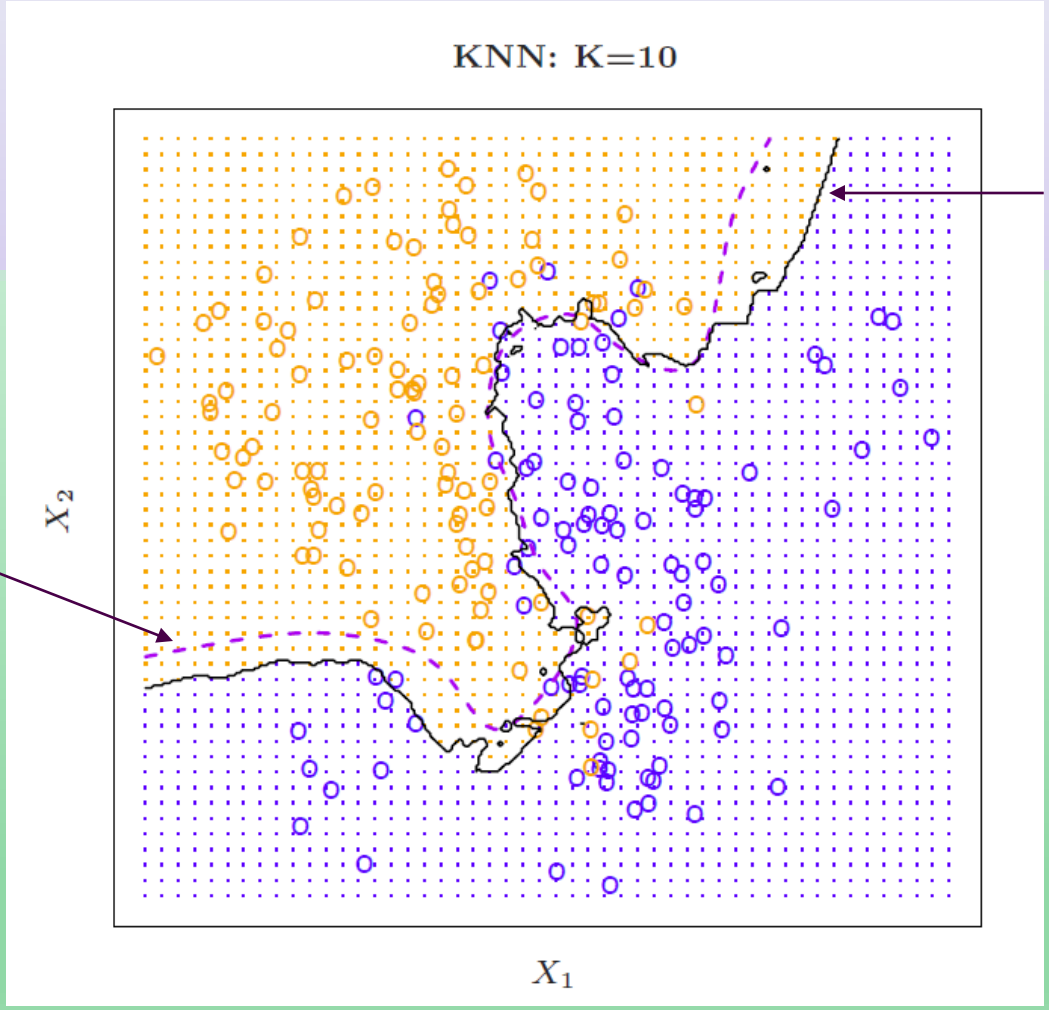


- The purple dashed line represents the points where the probability is exactly 50%.
- The Bayes classifier's prediction is determined by the Bayes decision boundary

K-Nearest Neighbors (KNN)

- KNN is a flexible approach to estimate the Bayes classifier.
- For any given X, we find the k closest neighbors to X in the training data and average their corresponding responses Y.
- If the majority of the Y's are orange, then we predict orange otherwise guess blue.
- The smaller that k is, the more flexible the method will be.

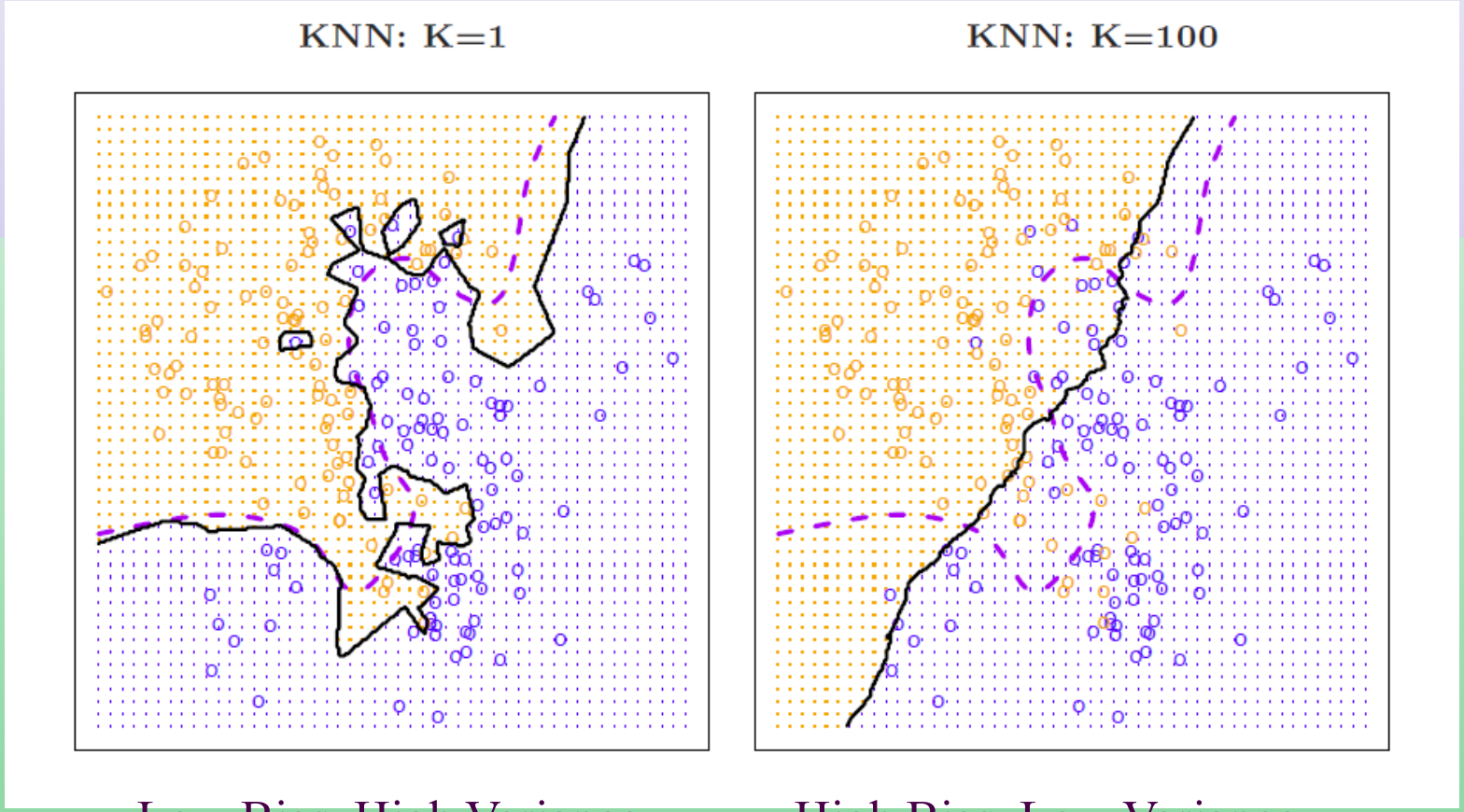
KNN: K=10



Bayes decision boundary

KNN decision boundary

KNN: K=1 and K=100



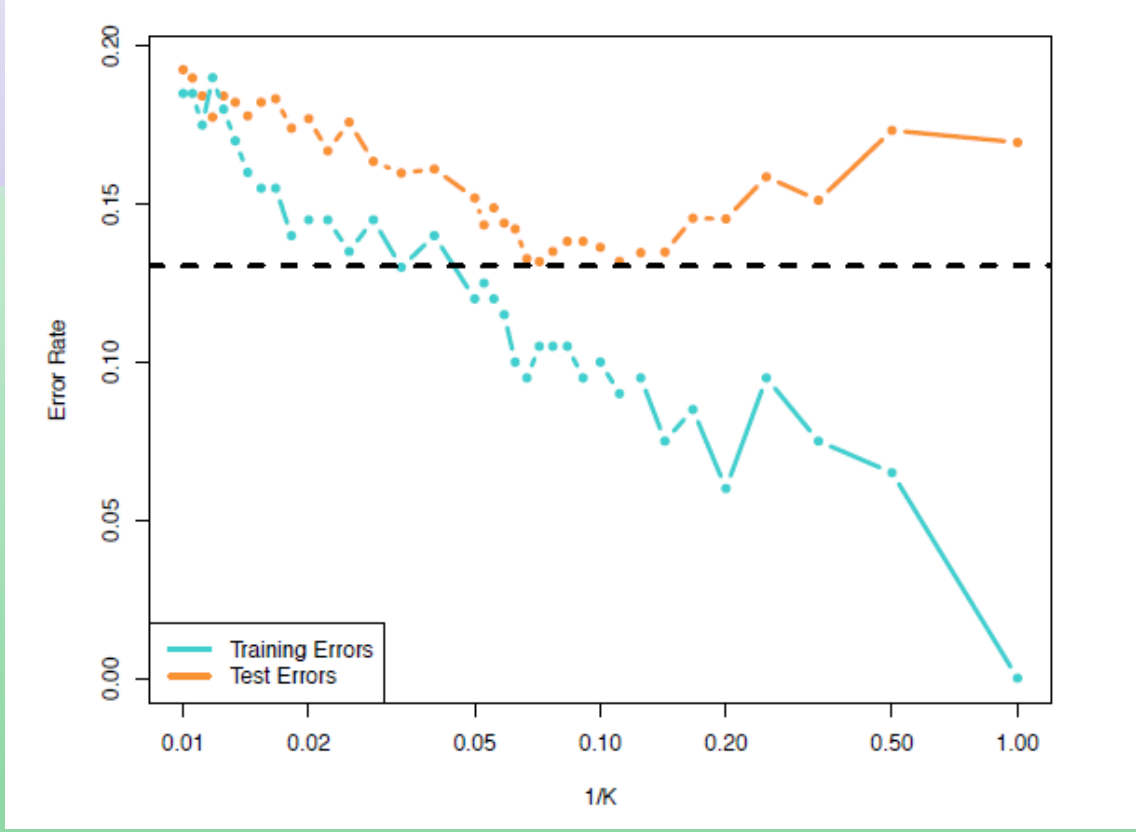
Low Bias, High Variance
Overly Flexible

High Bias, Low Variance
Less Flexible

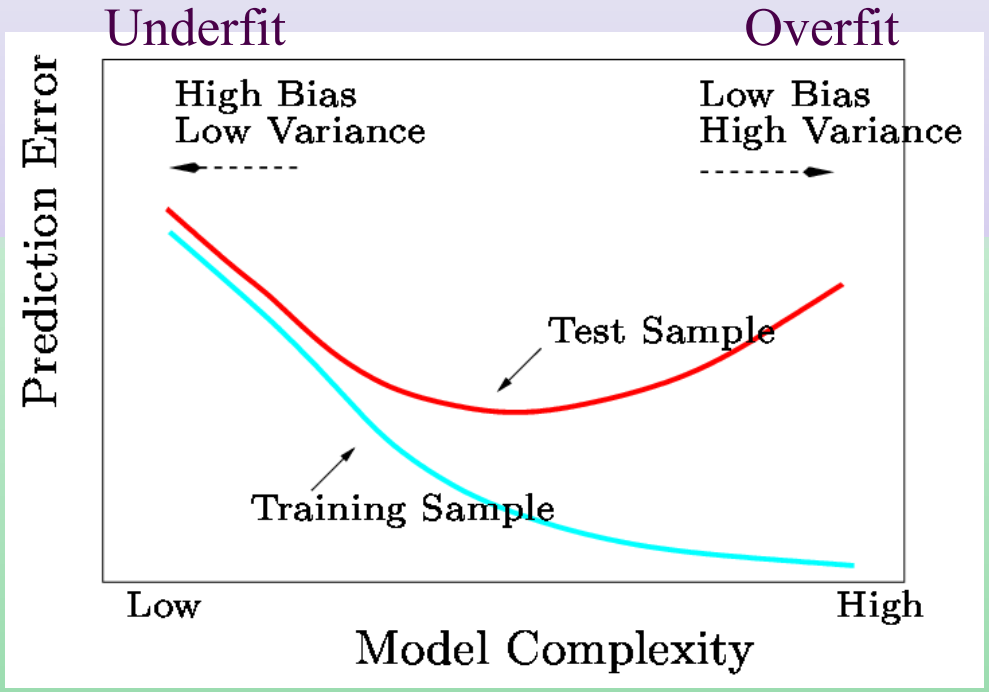
KNN Training vs. Test Error Rates

- Notice that the KNN training error rates (blue) keep going down as *k* decreases (i.e. as the flexibility increases).

- However, note that the KNN test error rate at first decreases but then starts to increase again.



Key Note: Bias-Variance Trade-Off



When selecting a machine learning method, remember that more flexible/complex is not necessarily better!!

- In general, training errors will always decline.
- However, test errors will decline at first (as reductions in bias dominate) but will then start to increase again (as increases in variance dominate).

What is R?

- Open-source, free software environment for statistical computing and graphics.
- Recommend using RStudio (GUI interface).
- 4,000+ packages available, with many used for machine/statistical learning and data mining.

R – Exploratory Data Analysis

- We use the “iris” data set to demonstrate exploratory data analysis in R.

R – Exploratory Data Analysis

- We inspect the dimensionality, structure and data of an R object.

R – Exploratory Data Analysis

- We view basic statistics and explore multiple variables.

R – Exploratory Data Analysis (cont.)

- We use the "iris" data set to demonstrate exploratory data analysis in R

```
> # We first check the size and structure of data
> dim(iris)
[1] 150   5
> names(iris)
[1] "Sepal.Length" "Sepal.Width"  "Petal.Length" "Petal.Width"  "Species"
> str(iris)
'data.frame': 150 obs. of  5 variables:
 $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
 $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
 $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...
> attributes(iris)
$names
[1] "Sepal.Length" "Sepal.Width"  "Petal.Length" "Petal.Width"  "Species"

$row.names
  [1]   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26  27  28  29  30  31
 [32]  32  33  34  35  36  37  38  39  40  41  42  43  44  45  46  47  48  49  50  51  52  53  54  55  56  57  58  59  60  61  62
 [63]  63  64  65  66  67  68  69  70  71  72  73  74  75  76  77  78  79  80  81  82  83  84  85  86  87  88  89  90  91  92  93
 [94]  94  95  96  97  98  99 100 101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124
[125] 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141 142 143 144 145 146 147 148 149 150

$class
[1] "data.frame"
```

R – Exploratory Data Analysis (cont.)

- We next look at the first five rows of data.

```
> # We next look at the first five rows of data
> iris[1:5,]
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
> head(iris)
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
6          5.4         3.9          1.7         0.4  setosa
> tail(iris)
    Sepal.Length Sepal.Width Petal.Length Petal.Width   Species
145          6.7         3.3          5.7         2.5 virginica
146          6.7         3.0          5.2         2.3 virginica
147          6.3         2.5          5.0         1.9 virginica
148          6.5         3.0          5.2         2.0 virginica
149          6.2         3.4          5.4         2.3 virginica
150          5.9         3.0          5.1         1.8 virginica
```

R – Exploratory Data Analysis (cont.)

- We can also retrieve the values of a single column.

```
> # We can also retrieve the values of a single column
> iris[1:10, "Sepal.Length"]
 [1] 5.1 4.9 4.7 4.6 5.0 5.4 4.6 5.0 4.4 4.9
> iris$Sepal.Length[1:10]
 [1] 5.1 4.9 4.7 4.6 5.0 5.4 4.6 5.0 4.4 4.9
```

- We can also get the summary statistics.

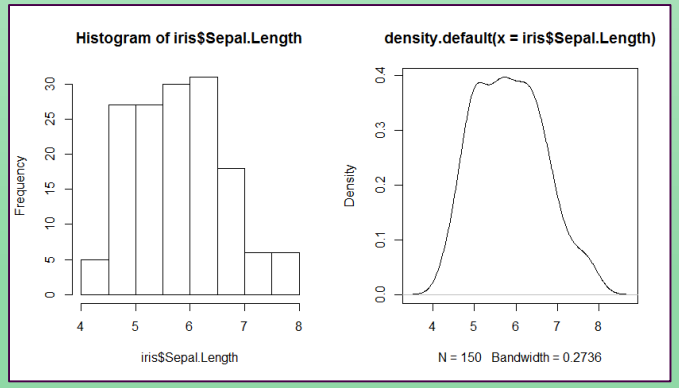
```
> summary(iris)
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width          Species
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   setosa     :50
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   versicolor:50
 Median :5.800   Median :3.000   Median :4.350   Median :1.300   virginica :50
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
>
```

R – Exploratory Data Analysis (cont.)

- We can also get quartiles and percentiles.

```
> # We can also get quartiles and percentiles
> quantile(iris$Sepal.Length)
  0%  25%  50%  75% 100%
 4.3  5.1  5.8  6.4  7.9
> quantile(iris$Sepal.Length, c(.1, .3, .65))
 10%  30%  65%
4.80 5.27 6.20
```

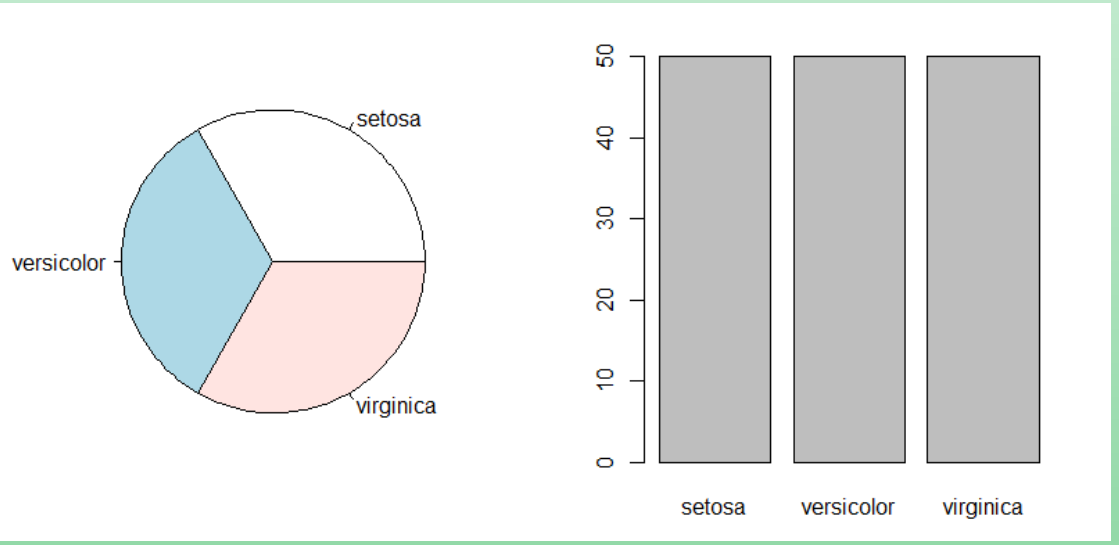
- We can check variance and get its distribution via histogram and density functions.




```
> # We can check variance and get its distribution via histogram and density functions
> var(iris$Sepal.Length)
[1] 0.6856935
> par(mfrow = c(1, 2))
> hist(iris$Sepal.Length)
> plot(density(iris$Sepal.Length))
```

R – Exploratory Data Analysis (cont.)

- We can get a frequency of the factors and plot a pie chart or bar chart.




```
> # We can get the frequency of facto
> table(iris$Species)

    setosa versicolor  virginica
        50         50         50
> par(mfrow = c(1, 2))
> pie(table(iris$Species))
> barplot(table(iris$Species))
```

R – Exploratory Data Analysis (cont.)

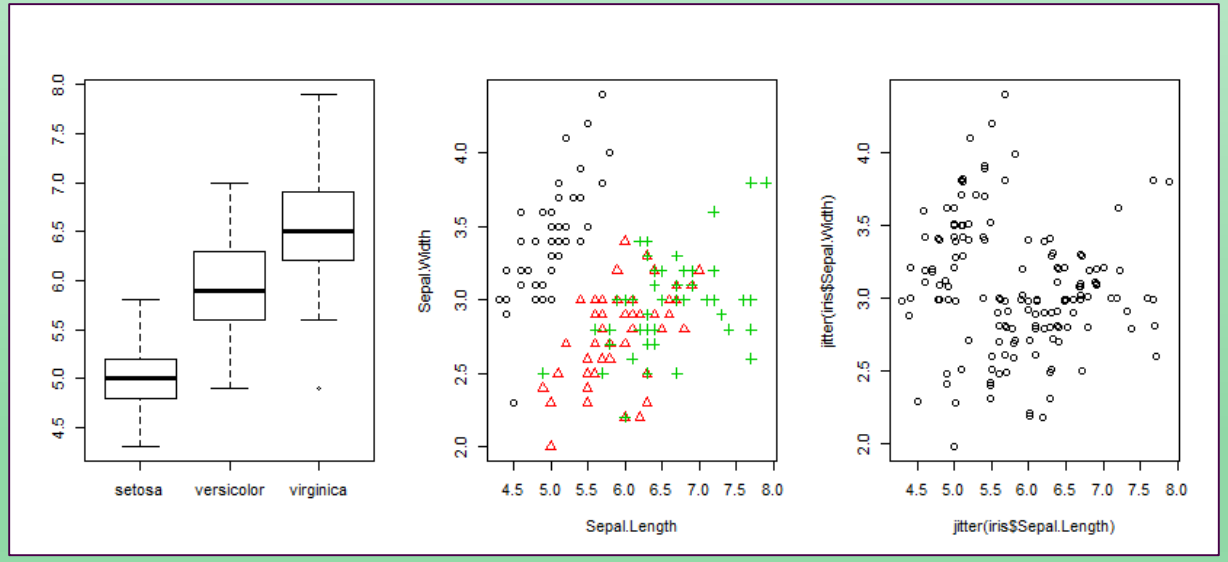
- We can explore multiple variables.

```
> # We can explore multiple variables
> cov(iris$Sepal.Length, iris$Petal.Length)
[1] 1.274315
> cov(iris[,1:4])
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length    0.6856935  -0.0424340    1.2743154   0.5162707
Sepal.Width    -0.0424340   0.1899794   -0.3296564  -0.1216394
Petal.Length    1.2743154  -0.3296564    3.1162779   1.2956094
Petal.Width     0.5162707  -0.1216394    1.2956094   0.5810063
> cor(iris$Sepal.Length, iris$Petal.Length)
[1] 0.8717538
> cor(iris[,1:4])
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length    1.0000000  -0.1175698    0.8717538   0.8179411
Sepal.Width    -0.1175698   1.0000000   -0.4284401  -0.3661259
Petal.Length    0.8717538  -0.4284401    1.0000000   0.9628654
Petal.Width     0.8179411  -0.3661259    0.9628654   1.0000000
> aggregate(Sepal.Length ~ Species, summary, data=iris)
     Species Sepal.Length.Min. Sepal.Length.1st Qu. Sepal.Length.Median Sepal.Length.Mean Sepal.Length.3rd Qu.
1     setosa             4.300                4.800                5.000              5.006                5.200
2 versicolor             4.900                5.600                5.900              5.936                6.300
3  virginica             4.900                6.225                6.500              6.588                6.900
  Sepal.Length.Max.
1             5.800
2             7.000
3             7.900
```

R – Exploratory Data Analysis (cont.)

- Boxplots, scatterplots and scatterplots with jitter (small amount of noise).

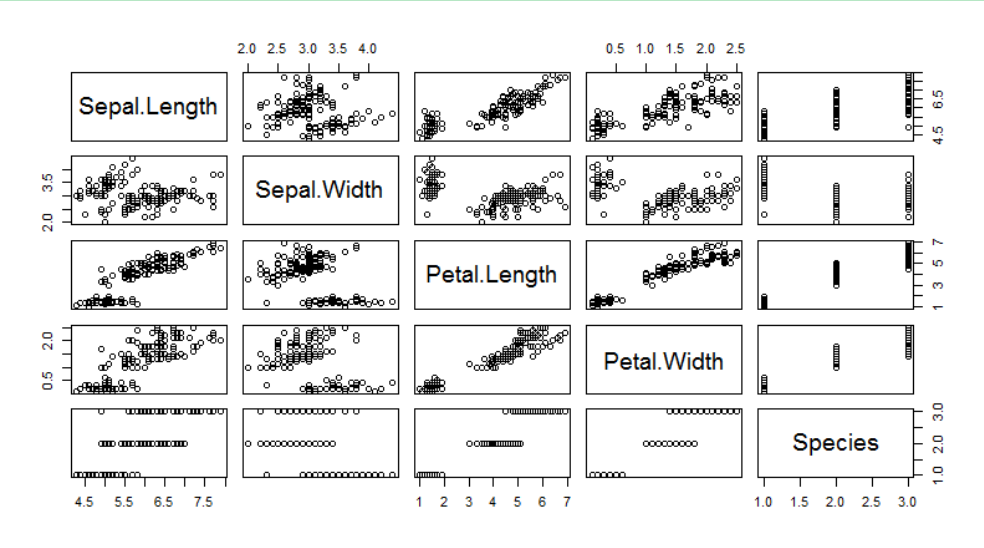
```
> # Boxplots, scatterplots and scatterplots with jitter (small amount of noise)
> par(mfrow = c(1, 3))
> boxplot(Sepal.Length~Species, data=iris)
> with(iris, plot(Sepal.Length, Sepal.Width, col=Species, pch=as.numeric(Species)))
> plot(jitter(iris$Sepal.Length), jitter(iris$Sepal.Width))
```

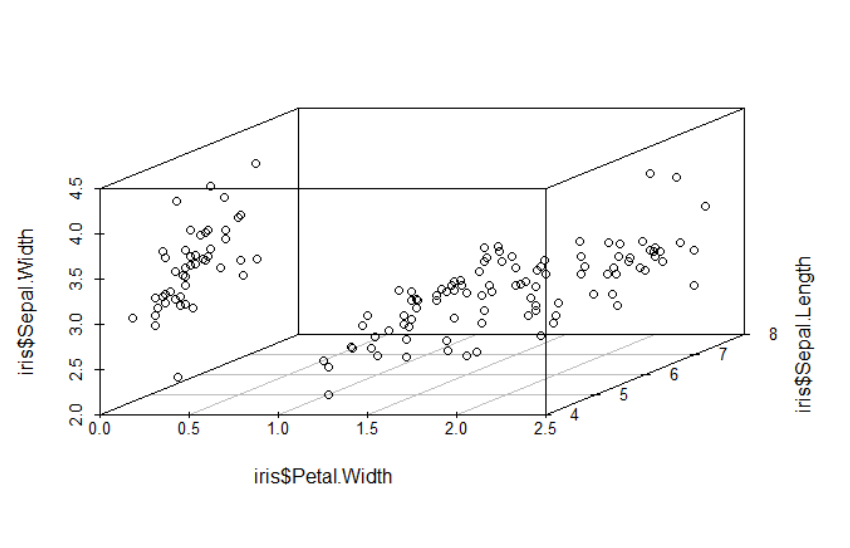


R – Exploratory Data Analysis (cont.)

- Produce a matrix of scatterplots or a 3D scatterplot.

```
> # Produce a matrix of scatterplots or a 3D scatterplot
> par(mfrow = c(1, 1))
> pairs(iris)
> library(scatterplot3d)
> scatterplot3d(iris$Petal.Width, iris$Sepal.Length, iris$Sepal.Width)
```

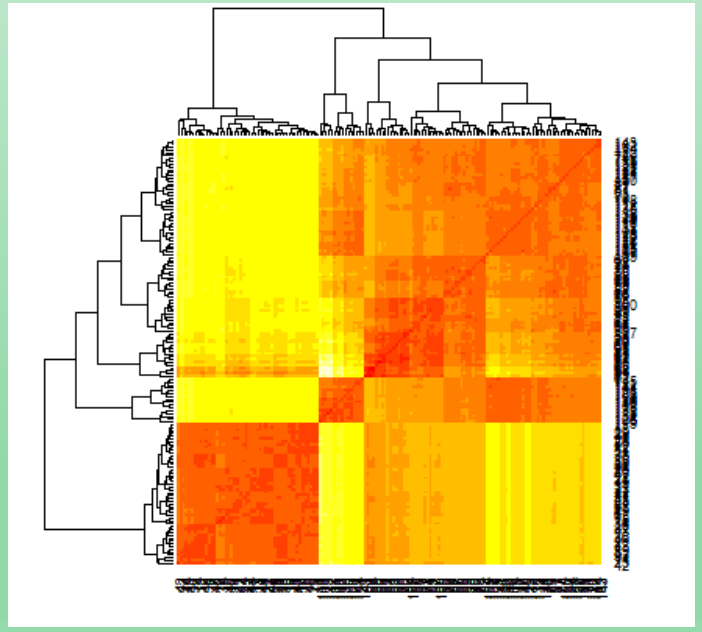


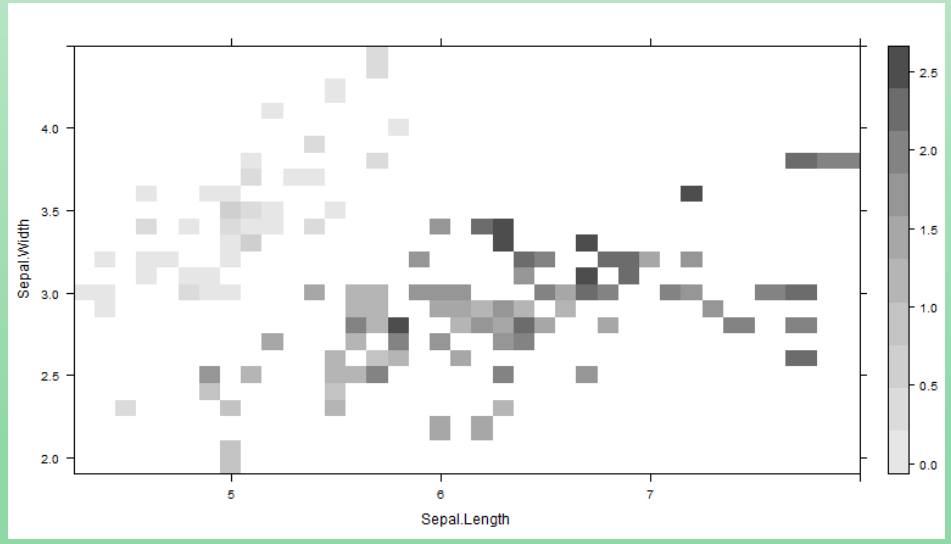


R – Exploratory Data Analysis (cont.)

- Produce a heat map or a level plot.

```
> # Produce a heat map or a level plot
> distMatrix <- as.matrix(dist(iris[,1:4]))
> heatmap(distMatrix)
> library(lattice)
> levelplot(Petal.Width~Sepal.Length*Sepal.Width, iris, cuts=9,
+           col.regions=grey.colors(10)[10:1])
```

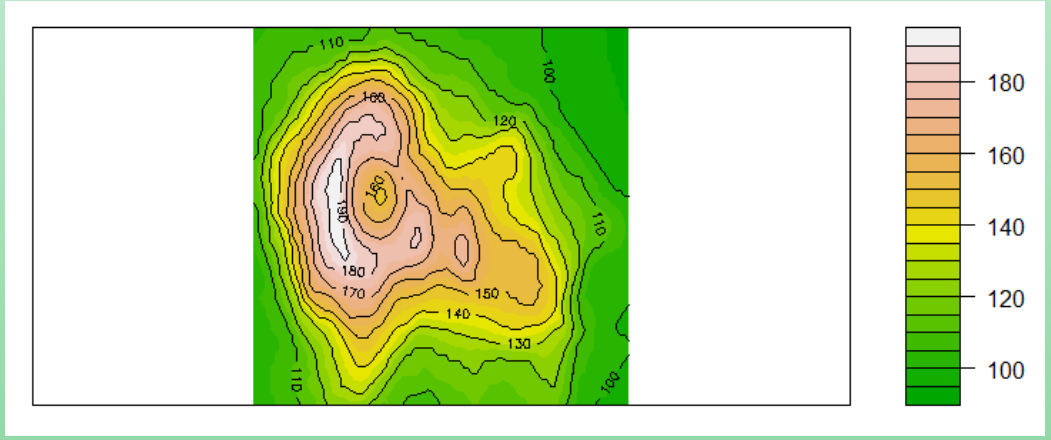


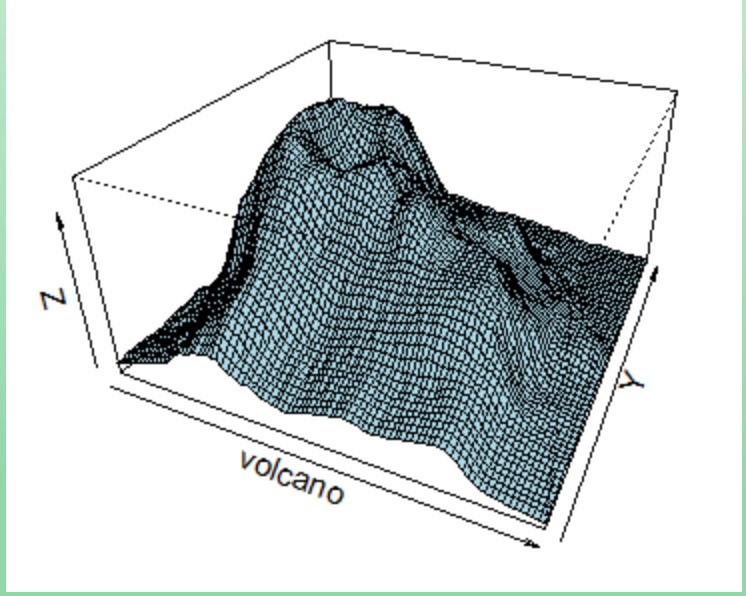


R – Exploratory Data Analysis (cont.)

- Produce a contour plot or 3D surface plot.

```
> # Produce a contour plot or 3D surface plot
> par(mfrow = c(1, 1))
> filled.contour(volcano, color=terrain.colors, asp=1,
+                plot.axes=contour(volcano, add=T))
> persp(volcano, theta=25, phi=30, expand=0.5, col="lightblue")
```

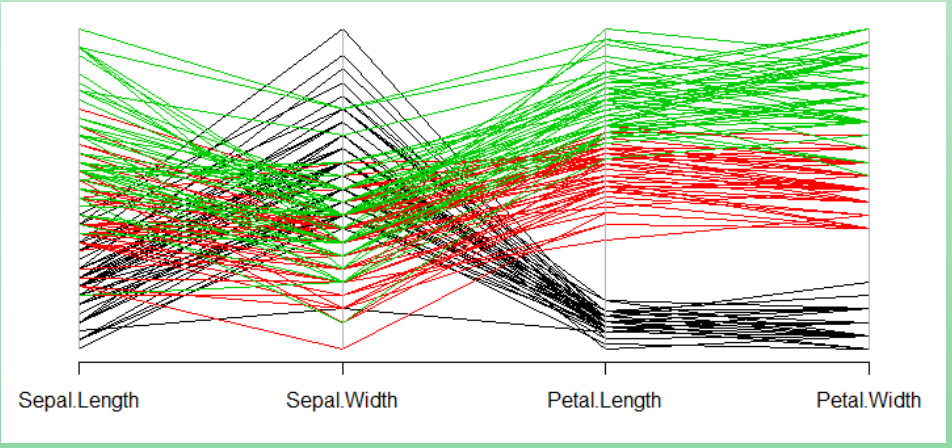


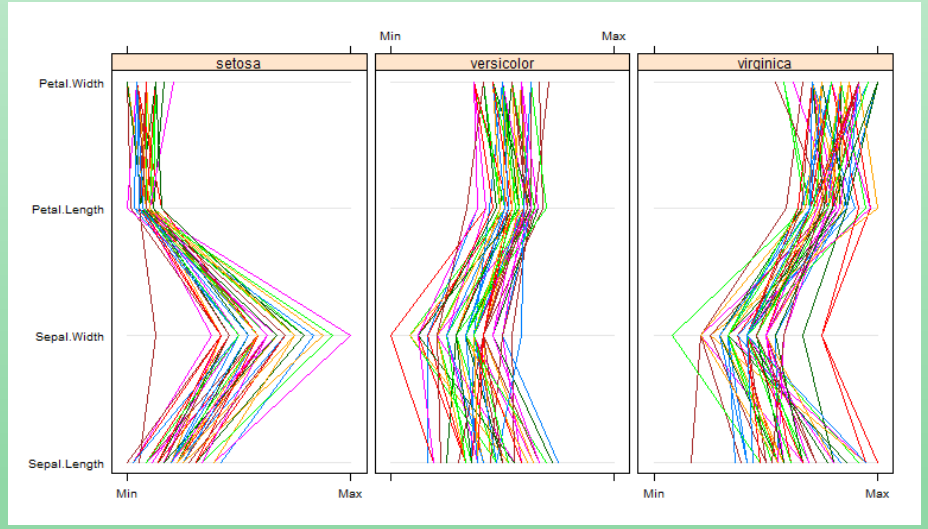


R – Exploratory Data Analysis (cont.)

- Plot parallel coordinates.

```
> # Plot parallel coordiantes
> library(MASS)
> parcoord(iris[1:4], col=iris$Species)
> library(lattice)
> parallelplot(~iris[1:4] | Species, data=iris)
```





More about R….

- Review the R tutorials posted on Canvas.
- Check out the CRAN website.
- Use help command = ?
- Practice, practice, practice…

Outline of the Course Concepts

- Ordinary Least Squares Linear Regression (Simple, Multiple)
- Resampling Methods in Machine Learning (Cross-Validation, The Bootstrap)
- Linear Model Selection and Regularization (Subset Selection, Shrinkage Methods, Dimension Reduction Methods)
- Non-Linear Models (Polynomial Regression, Regression Splines, Smoothing Splines, Local Regression, Generalized Additive Models)

Outline of the Course Concepts

- Classification Models (Logistic Regression, Discriminant Analysis, K-Nearest Neighbors)
- Artificial Neural Networks
- Tree-Based Methods (Decision Trees, CART, Bagging, Random Forests, Boosting)
- Vector Machines and Kernel Methods
- Unsupervised Learning (Principal Components Analysis, K-Means Clustering, Hierarchical Clustering)

Summary

- Overview of machine learning
- Key concepts of the learning problem
- Learning algorithm trade-offs
- Supervised versus unsupervised learning
- Regression versus classification methods
- Assessing model accuracy
- Bias-Variance trade-offs
- Introduction to R statistical programming

Reference

- 1. Shalev-Shwartz and Ben-David. Understanding Machine Learning: From Theory to Algorithms (Cambridge University Press, 2014)
- 2. Daum´e. A Course in Machine Learning.
- 3. The Art of Statistics: How to Learn from Data by David Shpigelter
- 4. Learning From Data – January 1, 2012 by Yaser S. Abu-Mostafa (Author), Malik Magdon-Ismail (Author), Hsuan-Tien Lin (Author)
- 5. Statistics: The Art and Science of Learning from Data by Alan Agresti
- 6. Learning From Data: An Introduction To Statistical Reasoning by M.Glenber.
- 7. Statistics: Learning from Data (with JMP Printed Access Card) by Rocky Pek
- 8. The Elements of Statistical Learning by Gerim Garold
- 9. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems 2nd Edition
- by Aurélien Géron (Author)