Session 1 - Introduction:

1. What do the 4Vs stand for?
2. Understand what the 4Vs mean?
3. What is statistical learning?
4. Why do estimate f?
5. How do estimate f?
6. What are supervised and unsupervised learning?
7. How supervised and unsupervised learning work?
8. Name supervised and unsupervised learning methods.

Session 2 – Basic Statistics:

1. What descriptive analysis is?
2. What are numerical measures?
3. When to use and how to calculate:
   1. Mean, median, mode
   2. Variance, standard deviation
   3. Range, quartiles, interquartile range
   4. Correlation, covariance
4. What are tabular and graphical representations?
5. When to use and how to draw:
   1. Histogram, box plot, scatter plot, bar graph, pie chart
   2. Frequency distribution table, cumulative distribution table, stem-and-leave table
6. What is a difference between a biased sample measure and an unbiased sample measure?
7. What is biased sample variance and unbiased sample variance?
8. Which is a better estimator for population variance?

Session 3 – Linear Regression:

1. How does a linear regression function look like + write down the formula.
2. What are the coefficients, what are the variables?
3. Properties of simple linear regression line? How it is chosen?
4. What does null hypothesis checks in linear regression?
5. What are t-value and p-value?
6. When can be the null hypothesis rejected?
7. Explain results from R program:
   1. Coefficients and useful statistics
   2. How to interpret them?
   3. How to determine whether there is a significant linear relationship between two variables?
   4. How to tell whether a model fits the data?
8. What are RSE (residual standard error) and R squared / R squared adjusted
9. What are prediction and confidence interval?
10. What is the difference between them and when to use which one?

Session 4 – Logistic Regression:

1. Why linear regression cannot be used in certain cases, like the juice?
2. Write down the formulas for: probabilities and odds
3. Which are parameters?
4. Given R results determine whether two variables are related and how they are related (positively or negatively).
5. How to make prediction, given the values of all the X variables.

Session 5 – Cross Validation:

1. What is the difference between training and testing dataset?
2. What are MSE and Error rate and how are they calculated?
3. What is relationship between training and test MSE/ error rate?
4. What are false positives and false negatives?
5. How to define true positive rate (TPR) and false positive rate (FPR)?
6. What is ROC curve and how to draw a ROC curve?
7. What is bias and variance?
8. What is relationship between bias and variance?
9. What is overfitting? When is a model more likely to be overfitted?
10. Examples of more and less flexible methods.
11. How validation set approach (split data in half), LOOCV and k-fold CV work?
12. What are their advantages and disadvantages?
13. How to compare them?
14. Other uses of cross validation apart from validation? Give example.

Session 6 – Bayes rule, Decision Trees:

1. What is Bayes rule?
2. How to create conditional probability tables for learning a Bayesian network classifier?
3. As above for naïve Bayes classifier?
4. How to use both classifiers to make predictions?
5. What is a decision tree?
6. Where to split?
7. How are the leaf nodes values determined in a regression and classification trees?
8. What is pruning and how to prune a tree? How far back to prune a tree?
9. What’s the difference between regression and classification trees?
10. Advantages and disadvantages of decision trees.
11. How to build a tree and prune in R. How to know the best number for a pruned tree? How to make prediction?

Session 7 – Random Forests:

1. How do bagging and random forests work?
2. How can both improve decision trees?
3. How random forests improve decision tree or bagging?
4. What is OOB and its purpose? What’s the advantage?
5. How to use random forests in R?

Session 8 - SVM:

1. How MMC, SVC and SVM classifiers work?
2. What is the meaning of hyper-parameter cost and how does it change the margin, support vectors and bias/variance?
3. As above for gamma?
4. How to tune cost and gamma using k-Fold cv?

Session 9 – Clustering:

1. What is clustering? What makes a good clustering?
2. How does K-means clustering work?
3. What is local optimum problem and how to solve it?
4. How does hierarchical clustering works?
5. What are four linkages?
6. How to build dendrogram using bottom up approach given a similarity matrix using different linkages?

Session 10 – PCA:

1. What can PCA be used for?
2. How to obtain the principal components?
3. How to convert the original dataset into the zero-mean dataset?
4. How to reduce the original dimensions of dataset by using the obtained PCs?