Big Data Analytics

Life of data: collect, clean, integrate, analyse, visualise

Four “V” of Big Data: volume, velocity, variety, veracity

Statistical learning – finding predictive function f based on a data

Irreducible error – no matter how well we model our function this error will stay

Prediction – predicting new value of X

Inference – relationship between Y and X

Parametric methods – make some assumptions about function f

Non-parametric methods – can be more accurate if large amount of observation is given

Prediction Accuracy vs Model Interpretatbility: more accurate methods can be more difficult to interpret, additionally more accurate methods can over fit the data

Overfitting – model describes random error (noise) instead of underlying relationship

Supervised methods: predictor X and response Y are observed (regression, classification)

Unsupervised methods: only predictors X are observed (clustering, PCA)

Classification methods: logistic regression, decision trees, SVM, tree based methods

R - indexing: X[row, column]

R – matrix: X = matrix(data, nrow, ncol, byrow = True)

R – normal distribution: x = rnorm (sample size, mean, sd)

**Questions from past papers:**

2015:

Standard deviation

Principal Component Analysis:

dominant vector (first principal component)

how many eigenvectors for a given dataset

Interpretation of regression – if B0 = 20 and B1 = -20 – what does it mean

Test Error Rate vs Training Rate Rate

How pruning the tree change training and test rate

Size of decision tree and training/test rate

Probabilities in logistic regression

4V of big data

Examples how CV is used to tune parameters

k-means is sensitive to initialisation

similarity matrix to hierarchical clustering

what is bias and variance

bias/variance in LOOCV and k-fold / tree and pruned tree / tree and bagged tree

what is overfitting, relation to bias-variance

quadratic kernel SVM and cost function

two ways of randomisation in random forest

random forest vs bagging

what is MSE / equation

what is error rate

R code to build classification tree model to compute test error rate

R code for linear regression: model, new prediction, confidence interval plot

2016:

usupervised learning

bias and variance

what is PCA / why scaling is needed in PCA

variance + equation

covariance

logistic regression odds

how to use CV to get cost in SVM

fitted line (slope) of linear regression

null hypothesis, test statistics, p-value for linear regression

overfitting

R code for bagged model

R code for testing mse for that model

R code to find best value of ntree

R code for prediction

4 scales of measurment: ration, interval, ordinal, nominal

R code to create matrix

R code for indexing – positive and negative indices

Unbiased covarience, correlation

Overfitting – increasing cost in SVM, increasing data amount, removing vector instances

Good clustering, what is centroid

R code for logistic regression

2018

why prune the tree, how cv helps prune

why portion data into clusters in k-means clustering

linkages in hierarchical clustering

types of clustering e.g. single

support vector classifier vs support vector machines

what to do with unknown values in the dataset

2019

random sampling

4 main objectives of PCA

what is more accurate: SVM or linear regression and why