Big Data Analytics

WEEK1 INTRODUCTION:

**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[advantage] if large amount of observation is given [disadvantage – large number of observations]

**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 commands:**

y ← seq(1,10) # creates vector with numbers 1 to 10

x ← matrix(data = c(1,2,3,4), nrow=2, ncol=2, byrow=TRUE)

x[2,4] # select row 2 and column 4

x = rnorm (n=10, mean=2, sd=1) # generates a vector of random variables where n is sample size with mean and sd as stated

cor(x), mean(x), var(x), sd(x) # correlation, mean, variance and standard deviation of x

plot(data), pairs(data) # scatter plot matrix

WEEK2 **Basic Statistics**:

For variance N-1 if unbiased sample, N for population and biased sample.

set.seed(number) – makes the results repeatable

WEEK3 **Linear Regression**:

Y – response, B0, B1 – parameters, X – predictor

Properties of least squares line: mean residual = 0 and residual sum of squares is minimized

R2 – proportion of variance that can be explained by predictor(s)

Confidence interval – estimate of an interval for future observations for a mean

Prediction interval – estimate of an interval for future observations for an individual

predict(model, data.frame(new values), interval = “prediction or confidence”)

WEEK4 **Logistic Regression**:

e = 2.71828

Maximum likelihood – used to estimate coefficients in logistic regression

Z-test - used to check null hypothesis instead of t test

model ← glm(results ~ predictor, data, family = “binomial”)

WEEK5 **Cross Validation**:

Bias is error introduced by modeling a real life problem by a much simpler problem

High bias = over simplified model

Variance – how much estimate for f would change by if you had a different training set

More complex method: bias decrease and variance increase

More simpler method: bias increase and variance decrease

Regression uses MSE, classification error rate

ROC curve is a plot of TPR (y axis ) vs FPR (x axis). The best is position 0,1 (top left).

To compare model accuracy we can compare area under the roc curve (AUROC).

Validation set approach has problem with results – they are not consistent (depending on the selected test/train data split).

LOOCV – less bias, less variable MSE but computationally expensive

LOOCV has higher variance than k-Fold

WEEK6 **Bayesian Networks and Decision Trees**:

Naive Bayes Classifier – use independence assumptions

Bayesian Network Classifier - use dependencies between features and class Y

Decision Tree – type of white box classifier, easy to explain, easy to show graphically, work well on classification and regression problems, disadvantage is poor prediction accuracy

WEEK7 **Random Forests**:

**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