MONASH University

ETC3250/5250: Introduction to Machine Learning

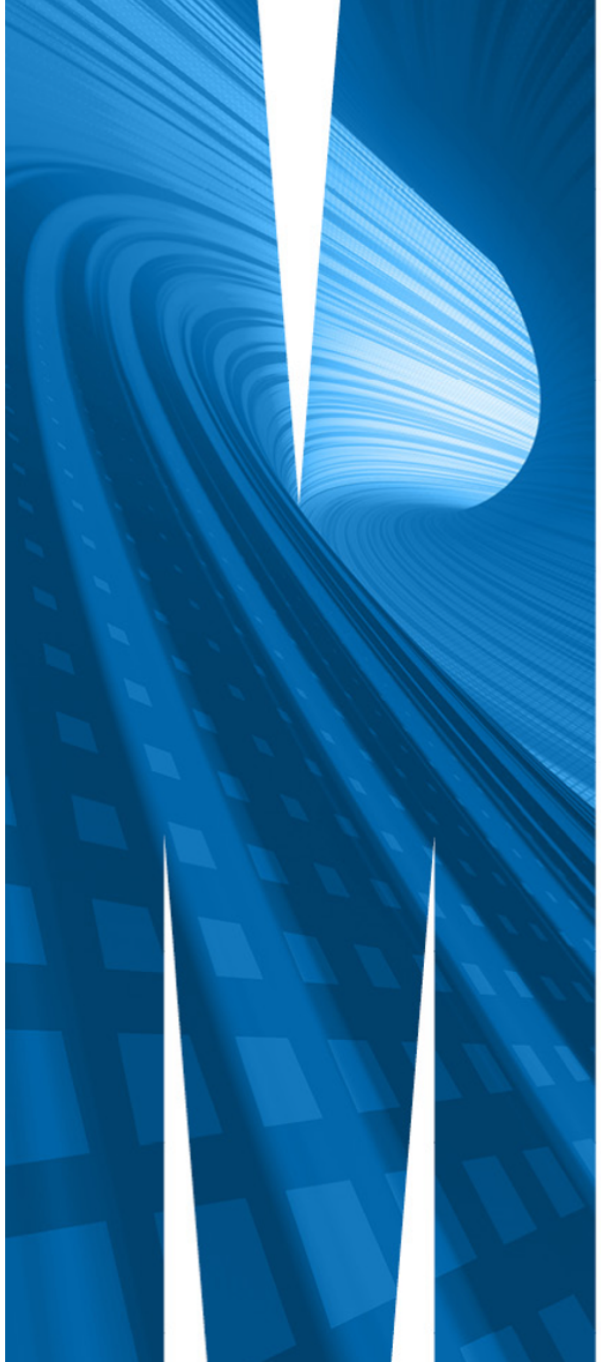
Unit wrap-up

Lecturer: *Professor Di Cook*

Department of Econometrics and Business Statistics

ETC3250.Clayton-x@monash.edu

Week 11b



Learning objectives for this class

- Select and develop appropriate models for regression, classification or clustering
- Estimate and simulate from a variety of statistical models, and measure the uncertainty of a prediction using resampling methods
- Manage large data sets in a modern software environment, and explain and interpret the analyses undertaken clearly and effectively
- Apply business analytic tools to produce innovative solutions in finance, marketing, economics and related areas

Outline

- What have we covered
  - Key concepts

- What type of problem do you have?
  - supervised (regression, classification),
  - unsupervised (PCA, cluster analysis)
- Should you use a flexible or less flexible model?
  - Parametric (more assumptions, easier estimation, strong inference)
  - Non-parametric (more flexible, fewer assumptions, more observations needed, less interpretable)
    - e.g. kNN, smoothers
- Measuring fit
  - MSE, R^2, BIC
  - accuracy, misclassification
- Bias vs variance trade-off
  - bias: error that is introduced by modeling a complicated problem by a simpler model

Outline

- What have we covered
  - Key concepts
  - Re-sampling
- Training/test/validation sets
- LOOCV, k-fold cross-validation
  - shortcut for computing MSE_i
- Bootstrap
  - out-of-bag error

Outline

- What have we covered
  - Key concepts
  - Re-sampling
  - Models
- Linear (and polynomial) regression
- Logistic regression
  - Logit function
  - Parameter interpretation
  - relationship to neural networks
- Linear discriminant analysis
  - assumptions
  - relationship to normal model
  - dimension reduction
- Quadratic discriminant analysis
  - heteroskedastic group variance-covariance

Outline

- What have we covered
  - Key concepts
  - Re-sampling
  - Models
- Decision trees
  - Regression: SST - (SSL+SSR)
  - Classification: Gini, entropy
- Random forests
  - Bagging
  - Sampling variables
  - Permutation
  - Diagnostics: Variable importance, Vote matrix, Proximity
- Support vector machines
  - maximal margin
  - relationship to LDA
  - kernels
- Neural networks
  - deconstructing the model fit
  - instability

Outline

- What have we covered
  - Key concepts
  - Re-sampling
  - Models
  - Visualisation

- Importance
  - Understanding structure in data, inform the model choice
  - Diagnose
  - Check assumptions
- Model in the data space
- Inference using randomization, bootstrap and permutation
- Tours
  - relationship to a biplot
  - Matching structure to variable contribution
  - Types: grand, guided,
- Parallel coordinate plots
  - Ordering variables
  - Scaling of axes
- Pairs plots

Outline

- What have we covered
  - Key concepts
  - Re-sampling
  - Models
  - Visualisation
  - Dimension reduction
- Principal component analysis
  - Eigendecomposition of variance-covariance
  - Scaling of variables
  - Choosing k
  - Total variance, and proportion
  - biplot
  - relationship to projection pursuit
- Regularization
  - Ridge regression
  - Lasso

Outline

- What have we covered
  - Key concepts
  - Re-sampling
  - Models
  - Visualisation
  - Dimension reduction
  - Cluster analysis
- Interpoint distance, similarity and dissimilarity
  - Rules for distance
- k-means
  - algorithm
  - random starts
- Hierarchical
  - Linkage
  - Dendrogram
- Choosing k, comparing solutions and summarising
  - Cluster fit stats: WBRatio
  - Summary statistics by cluster
  - Dimension reduction to show clusters
  - Multivariate plots of clusters
- Model-based clustering

How do you do well in this class?

Turn up to class, summarise your notes after each, note what you understand, and what you don't 🌊

Participate actively in computer labs, work with team mates to solve problems, get best answers 🍏

Do exercises from the textbook related to material each week, check your answers with online solutions 🧗

After this course

ETC3555 - Statistical Machine Learning

This unit covers the methods and practice of statistical machine learning for modern data analysis problems. Topics covered will include recommender systems, social networks, text mining, matrix decomposition and completion, and sparse multivariate methods. All computing will be conducted using the R programming language.

Prerequisites: ETC3250 or FIT3154

ETC5550 - Advanced Statistical Modelling

This unit introduces extensions of linear regression models for handling a wide variety of data analysis problems. Three extensions will be considered: generalised linear models for handling counts and binary data; mixed-effect models for handling data with a grouped or hierarchical structure; and non-parametric regression for handling non-linear relationships. All computing will be conducted using R.

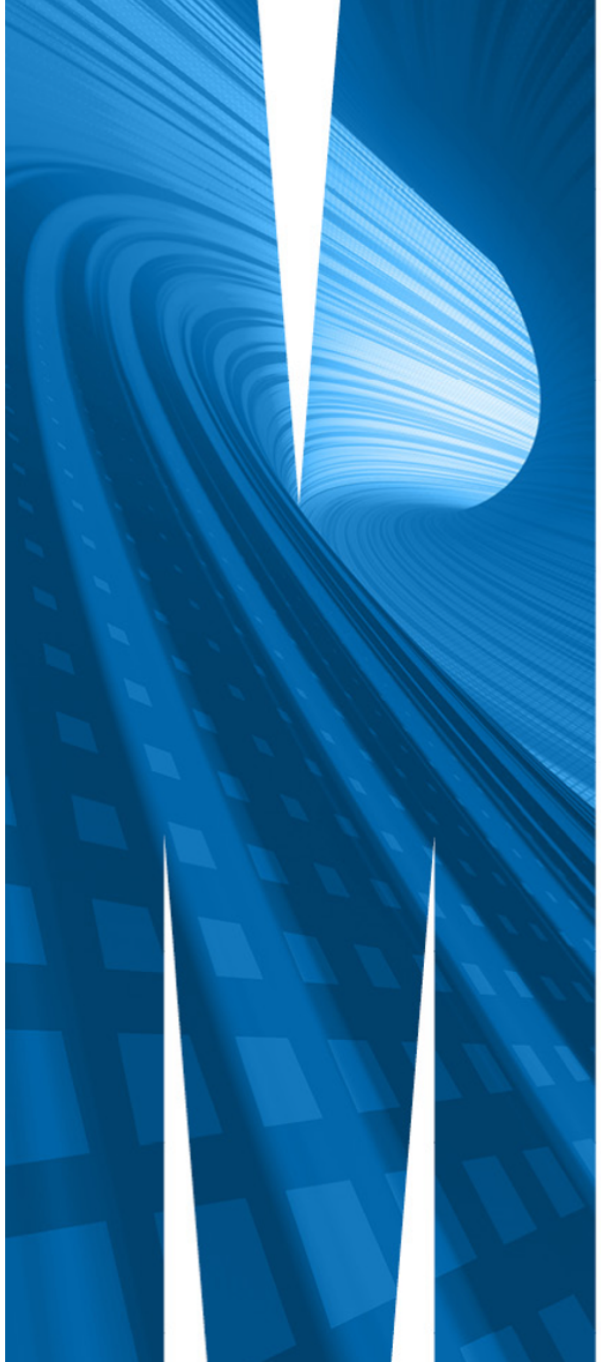
Prerequisites: ETC2410, ETC2420, ETC3440 or equivalent.

👩‍💻 All the best with the rest of the class, and the final exam.

For those of you wrapping up your studies, good luck in your future, and I hope that what you have learned in this class can be useful in your journey.

My data analyses, slides, papers, web site uses the R workflow. Very powerful framework for a business analyst/data scientist. There is no competition with python. Both extraordinary resources for our world. The people who develop and maintain these resources need our applause. We are extremely lucky to be living in a world where these resources are available free to us.

MONASH University



Lecturer: *Professor Di Cook*

Department of Econometrics and Business Statistics

ETC3250.Clayton-x@monash.edu

Week 11b