Final Review

Max Turgeon

DATA 2010–Tools and Techniques in Data Science

Probability and Statistics

- We started with a review of probability and statistics.
- We focused on *conditional* probabilities.
    - If I have some information about a random variable, what does it tell me about another random variable?
    - Bayes Theorem
- Discrete vs Continuous distributions, etc.

Data Wrangling

- We spent several lectures (and labs!) on data manipulation.
  - Summarize (by groups or not), create new variables, filter data.
  - Joining multiple datasets
  - Tidy data
  - Dates, regular expressions
  - Pandas
- Analysts spend a lot of time cleaning data.
- **Where to go from here?**
  - Databases
  - Data engineering
  - Big data

Data Visualization

- Visualizations in R and Python
    - Histograms, bar plots, box plots, scatter plots, etc.
- We also talk about principles of effective data visualization.
    - What visual cues are you using? Are they effective?
    - How can I best highlight important comparisons?
- **Where to go from here?**
    - Dynamic data visualization
    - Dashboards

Modelling

- The next few modules focused on data analysis and modelling.
  - We started by discussing correlation, distributions and significance.
  - We then discussed scores and rankings.
  - We discussed how to build models and evaluate them
- We discussed linear regression in some detail, focusing on many aspects:
  - Evaluating regression models
  - Training and test data
  - Flexible modelling with splines
  - Regularized regression
- **Where to go from here?**
  - Generalized Linear Models
  - Bayesian statistics

Machine Learning

- We spent the last few weeks on machine learning.
  - Logistic regression
  - Nearest Neighbour
  - Decision trees and random forests
  - Clustering
- **Where to go from here?**
  - Optimization
  - Hyper-parameter tuning
  - Deep Learning