00_Introduction

June 24, 2019

1 Workshop title: A "Revue" of Models for Statistical Inference and Machine Learning

2 Workshop description

- A high-level overview of common models used for inference (linear, generalized linear, generalized linear mixed, LASSO, ElasticNet) and prediction (random forests, gradient boosted trees, neural networks).
- Intended use case, deployment strategies, advantages and common pitfalls for each will be discussed.
- Example code for all models provided in both R and Python for quick adaptation to your project.
- From known parameters, we will create synthetic data with ever-more exotic variance structures (non Gaussian-distributed, non i.i.d., heteroscedastic data), visualize the data, and use appropriate models to back out the parameters we used to make the data.
- Considerations including preprocessing, interpretation, diagnostics, model selection, outliers, overdispersion, and corrections for multiple comparisons will be discussed.

3 Prerequisites

- Must have taken an Intro to Stats course at some time in your life.
- Must have run some R or Python code of your own accord at some time in your life.
- Must know what a data frame is and what it's used for.

4 Motivation: Poohsticks game

- Idea: input data with known properties into various models, and see what happens.

5 Workshop outline

1. Day 1. Introduction; exploratory data analysis and pre-processing; interpreting coefficients from LM & ANOVA
2. Day 2. Models for statistical inference: LASSO, GLM (Logistic & Poisson), GLMM
3. Day 3. Models for machine learning: Random Forests, XGBoost, Neural Networks
4. Day 4. Model assessment; ensemble learning; AutoML; troubleshooting

6 Day 1 outline

1. Personal introductions
2. Modelling caveats
3. Difference between statistical inference and machine learning
4. Types of data
5. Data mining pipeline in general
6. Modelling a Step function in R
7. Modelling a Linear function in R

7 About the Computational Biology Core (CBC)

Core facility housed in LGG
Room 10C222
Seminar or training every month
Two powerful Windows computers with lots of software and remote access available
Soon: BRC cloud computing (RAM, GPUs, virtual machines)

7.0.1 CBC Staff

- Supriyo De, Ph.D., Staff Scientist
- Krystina Mazan-Mamczarz, Ph.D., Senior Research Fellow
- Qiong (Joan) Meng, Ph.D., Post-doctoral fellow
- Christopher Coletta, M.S. expected 2019, Computer Scientist

8 Modelling Caveats

8.1 "All models are wrong, but some are useful."

- Sir David Cox, originator of the Cox proportional hazards model, said: The idea that complex physical, biological or sociological systems can be exactly described by a few formulae is patently absurd."
- Statistician George Box said: "Cunningly chosen parsimonious models often do provide remarkably useful approximations." He then cites the ideal gas law pv=nRT as an example. "For such a model there is no need to ask the question 'Is the model true?'. If 'truth' is to be the 'whole truth' the answer must be 'No'. The only question of interest is 'Is the model illuminating and useful?'."

8.2 Corellation does not imply causation

- Causal inference = correlation, plus causal reasonong, which involves the "ceteris paribus" assmption- "The only difference is what we changed."
- Otherwise reporting correlation is the best we can do.

8.3 Weapons of Math Destruction

- Evil : Measuring proxies rather than measuring the actual thing

  - Less evil: Credit score - it is a measure of how likely a person is to default on a loan. Have you defaulted before? There is redress to fix things if there are discrepancies.
  - More evil: US News and World Report college rankings
  - When a measure becomes a target, it ceases to be a good measure.
- Evil: Predictive models that use past outcomes as training set, TO PERPETUATE future outcomes
  - Algorithms that set bailbonds amounts
- Cognitive bias in Machine learning is human bias on steroids
  - We seek out evidence that supports our existing point of view while avoiding information that contradicts it

8.4 Checking for multiple comparisons

- Come to Osorio Meirelles's workshop!
- The goal of adjusting for multiple comparisons is to reduce the number of false positives.
- Upshot: Every single individual p-value you get, including all the pvalues for the betas in a single model, is a (potential) target to be adjusted for multiple comparisons.
  - Bonferroni: "Family-wise error rate"; too strict
  - Benjamini-Hochburg: "False discovery rate"; less strict

9 Difference between statistical inference and machine learning

- https://www.coursera.org/lecture/statistical-genomics/inference-vs-prediction-8-52-PkWHh
- Model interpretability versus predictive power

10 Data Mining Pipeline

- Pre-processing, a.k.a. "data wrangling"

10.1 SEMMA

10.1.1 Sample

- Import the data
- Check the data types
- Partition into training & validation sets for predictive modeling
  - What is the sampling unit? Read? Person?

10.1.2 Explore

- Look for outliers
- Univariate descriptive statistics
- Bivariate descriptive statistics - include target variable

- Cluster - a descriptive model
- Check for multicolinearity

10.1.3 Modify

- Transform
- Impute
- Replacement
- Drop

10.1.4 Model (with metrics)

- Linear regression (Adjusted-R^2, p-values)
- Logistic regression (accuracy, p-values)
- Random forests, gradient-boosted trees, neural networks (accuracy)

10.1.5 Assess

- Model comparison for predictive model
- Visualize
- Report