Machine Learning 101

A Practitioner's Guide

Onyi Lam

UCSD

What is Machine Learning?

Types of Machine Learning Tasks

- Detection (patterns, event)
  - Detect plagiarism
- Prediction (predict the future)
  - Targeted cash transfer to most in need
  - Predict student at risk of not graduating on time

Types of Learning

- Supervised Learning
- Unsupervised Learning

A Different Approach

- We (mostly) don't care about the structure of the model
- We don't (necessarily) want the model that best fits the data we've already seen, but rather the model that will perform the best on new data.
- We can put whatever variables we want, and as many as we like, into a model.

Typical Work Flow

1. Data Preprocessing
2. Feature Engineering
3. Designing Cross-Validation Schemes
4. Looping Through Models
5. Model Evaluation
6. Selecting the Best Model
7. Generate Prediction and Understand Important Features

Data Preprocessing

Missing Values

- Remove
- Replace with some value (mean, median..etc?)

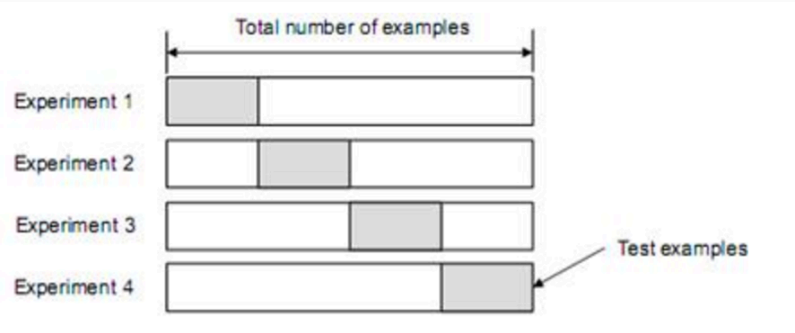
Non-Numeric Values

- Dummify
- Convert to numbers
- Combine categories

Splitting Data into training set and test set

Cross-Validation

- Repeated random sub-sampling validation
- k-fold



- Random Assignment
- Split by cohort, year ... etc

Testing Different Models

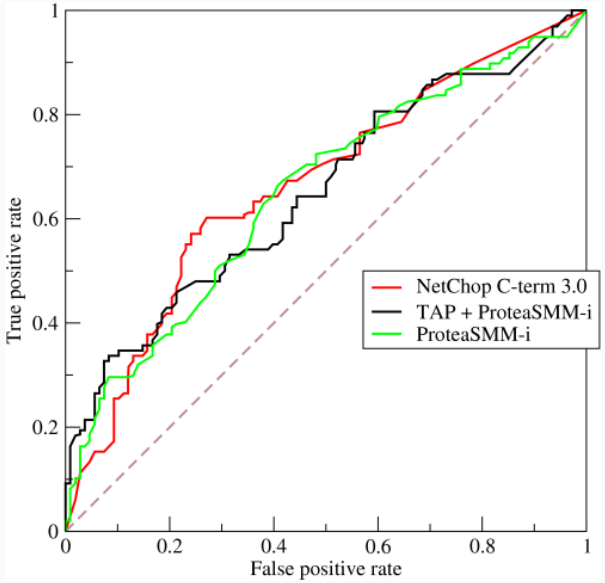
Common Models

- Logistic Regression
- Decision Tree
- Random Forest
- KNN

Model Evaluation

Common Metrics

- AUC



- Precision

$$PRE = \frac{TP}{TP + FP} \tag{1}$$

- Recall

$$REC = \frac{TP}{P} = \frac{TP}{FN + TP} \tag{2}$$

Generate Prediction

- Use best performing best to generate prediction
- Understand the predictive features