MATH 547/ BIOINF 547: Mathematics of Data

Date: March 18, 2019

Due date: March 26, 2019 (11:59pm)

Data problem from Defense Advanced Research Projects Agency (DARPA)

From Dr. Sri Kumar, Brian Sandberg, Information Innovation Office (I2O)

Dataset name: 38_sick dataset

Description: Thyroid disease records supplied by the Garavan Institute and J. Ross Quinlan, New South Wales Institute, Syndney, Australia. 1987. Source: Dua, D. and Graff, C. (2019). UCI Machine Learning Repository http://archive.ics.uci.edu/ml. Irvine, CA: University of California, School of Information and Computer Science.

Task type: Binary classification (tabular)

- Given patient information, can you predict if patient will have thyroid disease ('sick') or not ('negative')
- Prepare data and construct best performing ML model to extract information from data and make predictions

Metric (measuring goodness of fit): F1 Macro

Input:

- Training dataset: 38_sick_train.csv
- Test dataset: 38_sick_test.csv (blind dataset; test your model and deliver predictions, see Output)
- Ground truth: 38_sick_ground_truth.csv (not provided)

Dataset has 29 attributes. Column 'd3mIndex' can be ignored/removed

Output:

Create single output file with the name 'id_predictions.csv'

- id = unique identifier per student (1_, 2_, etc)
- single .csv file with a single column of predictions
- each row is a prediction with the value of either 'negative' or 'sick'
- file should contain 754 rows (predictions)

Data scientist or automated machine learning (AutoML) system will need to address various data challenges:

- Strategy to impute missing values
- Transform categorical / text data into numerical data for estimator
  - dummy explanatory variable / indicator variable / one hot encoding
  - useful for time series analysis, seasonal analysis, economic forecasting, bio-medical studies, credit scoring, fraud detection

  - check consistency of encoded vectors between train and test data

- Project/decompose multivariate data to a lower dimensional space (set of successive orthogonal components that explain a maximum amount of the variance)
- Deal with label imbalance
  - success based only on underlying class distribution (accuracy paradox)
  - accuracy is not a good metric for predictive models (use precision and recall, f1 score, confusion matrix)
  - can't ignore the lesser class
  - ratio of 'negative' to 'sick' labels in training set is more than 15:1
- Avoid overfitting
  - capture noise in training data, which doesn't represent the true properties of the data
  - bias-variance tradeoff to control errors (simple v. complex models; high bias is off center; high variance is scattered)
  - cross-validation (generalize to unseen data)
- Select and tune best performing estimator
  - given the task and characteristics of the data, which machine learning algorithm should be searched first
  - search optimization strategies
  - Automl goes through 100Ks of iterations of training and validation to pick the best model
  - how does data transforms (properties of data) influence estimator selection
- Understand/explain model (model interpretation)
  - model complexity
  - explain important features and success/failure of mdoel to build user trust and address problems like bias and leakage
  - how well can human (1) consistently estimate what a model will predict, (2) understand and follow the model's prediction, and (3) detect when a model has made a mistake
  - what are the top most important features (e.g. using relative feature importance measures)?