One Drug Sensitivity Dataset Analysis

Yue Jiang
Zhenzhen Li
Lizhang Miao

Hong Kong University of Science and Technology

April 10, 2017

Overview

Outline

Introduction

- One-drug dataset.
- 60 covariates vs logarithm of IC50s
- 542 observations $\rightarrow$ training set (500) + validation set (42)

Outline

Naive Mean Method

- inspired by "binary" features $\rightarrow$ clustering?
- distance: inner product of features vector
- use neighbors to make prediction: local sample mean

Elastic Net

- 60 covariates $\rightarrow$ variable selection?
- combine L_1 and L_2 penalty $\rightarrow$ elastic net
-

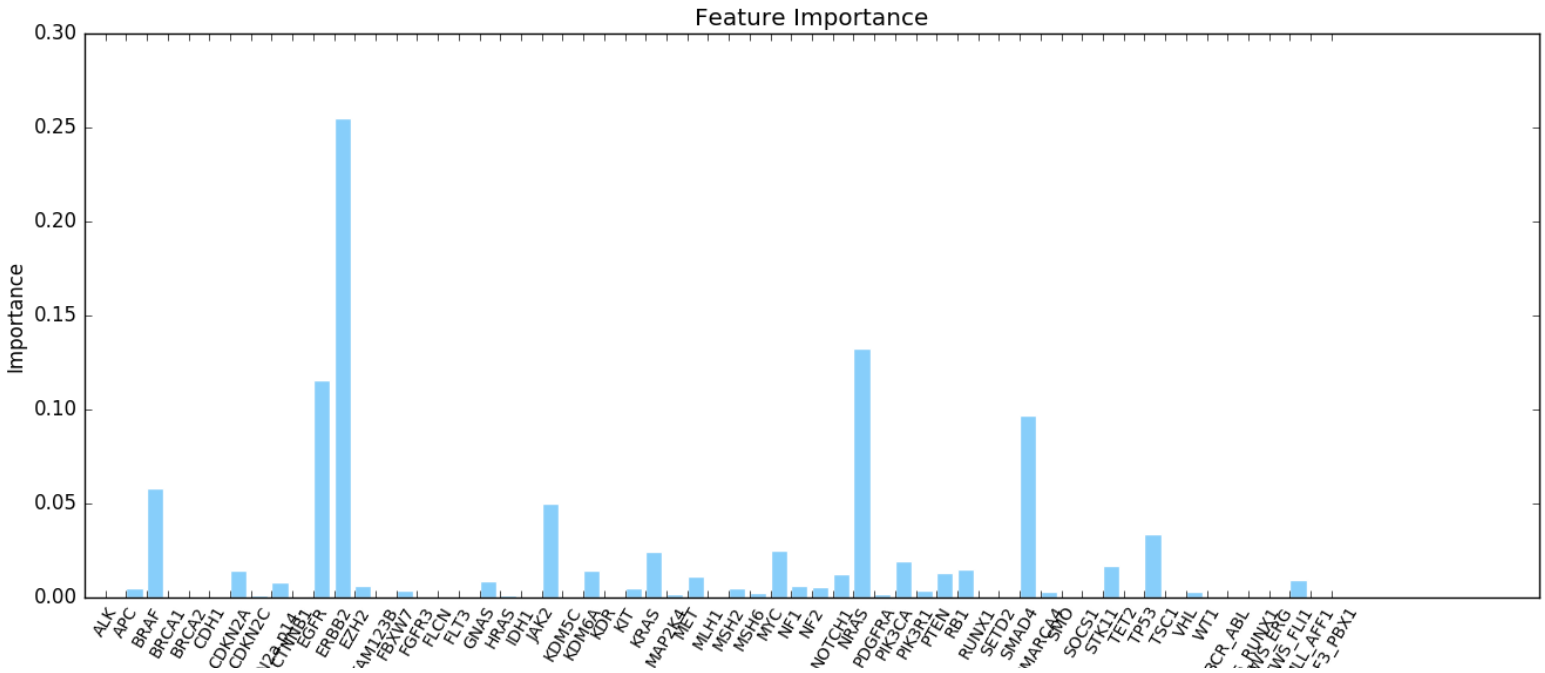
$$\hat{\beta} = argmin_{\beta} \left(\frac{1}{2N} \sum_{i=1}^{N} \left(y_i - x_i^T \beta \right)^2 + \lambda \left((1-\alpha) \frac{1}{2} ||\beta||_2^2 + \alpha ||\beta||_1 \right) \right),$$

- hyperparameters: α, λ
- perform 10-fold cross validation to find the optimal λ for each α.
- Optimal α: 0.997 $\rightarrow$ close to LASSO

Bagging and Random Forest

- A single decision trees might tend to overfit. To improve stability, bootstrap-aggregated decision trees combine the results of many decision trees, which reduces variance and improves generalization.
- Ensembling trees using a bootstrap samples of the data and selecting a random subset of predictors to use at each decision split.
- Performance is not good when data is too sparse

Bagging and Random Forest



Gradient Boosted Regression Tree

- boosting: ensemble a group of weak estimators
- starting from a weak estimator, fit a regression tree to pseudo-residuals and add to previous model
- hyperparameters:
  - number of trees (number of iterations)
  - minimum number of observations in the leaf node
  - maximum tree depth
- choose the set of hyper-parameters that can minimize the validation error

Local Regression Tree

- Direct regression method such as Lasso regression, Elastic regression, SVR don't achieve satisfying result - no method achieve MSE less than 3.0, large noise? Too much hidden information as inputs are all 0-1? (some samples even share the same feature but different IC50). Local regression might help.
- Feature: As some genes are correlated, raise 60 features to 3660 features(added with quadratic feature) might help.
- distance: difference of features taking L2 norm. (Too spare, the range of inner product is small)
- use neighbors to make prediction: local regression tree

Outline

Summary

- Are the cell line names useful?
- What's the appropriate definition of "distance"?

Thank you!