Feature selection for Predictive Modeling

Karthik Prakash

kprakas2@stanford.edu

Motivation

- Predictive models are trained on a scheduled basis on the recent dataset.
- WHY WOULD SOMEONE TOUCH CODE THAT IS WORKING IN PRODUCTION!!!
- Lots of things change in few months in Industry. The features you thought was the most important might be obsolete because the data changed.
- You are BURNING money hosting these obsolete features. For real-time model predictions you need the feature values in super-fast database which is not CHEAP.
- Goal: Build a framework that gives meaningful insights of the features which will aid in maintaining feature data quality.

Data

CONSTRAINTS WHILE CHOOSING DATA
- A simple dataset with few features so that it is easy to do feature selection experiments.
- NO correlated features.
- Synthetically generate data to vary dataset size and features.

SYNTHETIC DATA GENERATION
- All features were univariate normal distributions separately.
- Sample from Normal Distribution for each feature to produce large datasets with same distribution.
- Verified Data distribution of synthetic and original data are identical.
- No change in the model accuracy for synthetic data compared to original data.

SYNTHETIC Features Added
- Random features.
- Redundant features.
- Low Variance features
- Zero Variance features

Technique

Recursive Feature Elimination

$X = \text{training data}$

for $n$ in range(number_of_features):

$\theta = \text{train_logistic_regression}(X)$

least_important_feature = min value in $\theta$

$X$ = drop(least important feature)

return feature ranks

Variance Threshold Elimination

- All the features below certain variance threshold are eliminated.

Chi Square Statistics

- chi-squared stats between each non-negative feature and class.

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Tree Based Feature Selection

- ExtraTreeClassifier trained on the data.
- Feature importance scores available from the trained tree classifier.

Analysis

- Most effective way of selecting feature if cross validated with original model
- Effective for redundant features, low variance features, random features.
- EXPENSIVE: N! number of training jobs.

- Easiest way to identify “aging” features losing importance over a period of time.
- Not effective on any other type of irrelevant features.

- Good to find low variance features and random features
- Not effective against redundant features
- Not effective when feature values have negative numbers

- Regression Models not supported
- Not effective against redundant feature
- Effective against Low variance and ransom features.

Future

- I would make this an open source framework with self-serve usage documentation.
- Automatic Model training on different feature selection datasets and comparison in model accuracy, prediction time and training time if a given feature set is selected or not.

Feature Selection Report

Results

- KNN Classifier Models trained on different datasets sizes and different feature counts
- Dataset Size has high impact on training time
- Irrelevant feature count has impact on training time
- Irrelevant feature count has impact on prediction time

- Decision Tree classifier trained on different dataset sizes and different feature counts
- Irrelevant feature count has impact on training time
- Irrelevant feature count has impact on prediction time