Step-wise and Embedded Methods

Lesson 6 – Section 4

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Overview

- Step-wise model selection
- Embedded method

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Step-wise Model

Step-wise Model (Feature) Selection (1)

- Forward:
  - –Start with a model with only inception
  - –Add one feature in the model at each step
  - –At each step, the variable that can maximally reduce the residual sum of squares (RSS) is chosen as the feature to add in the model.

Step-wise Model (Feature) Selection (2)

- Backward:
  - –Start with a model with all features
  - –Remove one feature from the model at each step
  - –At each step, the variable that can minimally increase the residual sum of squares (RSS) is chosen as the feature to remove from the model.

Step-wise Model (Feature) Selection (3)

- Both:
  - –At each step, will check whether add a feature, or remove a feature

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

How to Select the Best Model (Feature Set)?

- Akaike information criterion (AIC)
  - –k: number of coefficients to estimate in the model
  - –L: likelihood of the training data based on the model

$$\mathrm{AIC} = 2k - 2\ln(\hat{L})$$

- Bayesian information criterion $\mathrm{BIC} = \ln(n)k - 2\ln(\hat{L}).$

- Choose the model that has the minimal AIC or BIC
- AIC tends to choose a larger model than BIC
  - –AIC has less penalty on the complexity of model (k) than BIC

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Embedded Method

Embedded Method

- **Lasso** (least absolute shrinkage and selection operator)

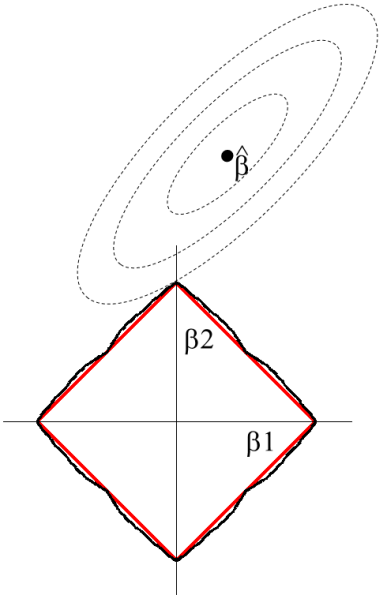
$$\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{N} \|y - X\beta\|_2^2 \right\} \text{ subject to } \|\beta\|_1 \leq t.$$

$$\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{N} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\}$$

- Based on the second equation, we are penalizing on the complexity of the model (The sum of absolute values of the coefficients)

Why LASSO Can Select Features?

- Assuming only 2 X variables
- $\hat{\boldsymbol{\beta}}$ is the coefficient vector where there is no penalty
- Ellipsoid is the contour of MSE when coefficients change
- Very likely, some contour will meet with $|\beta_1| + |\beta_2| \leq t$ at the corner
- At the corner, the coefficients of some variables are set to 0
- These variables are de-selected



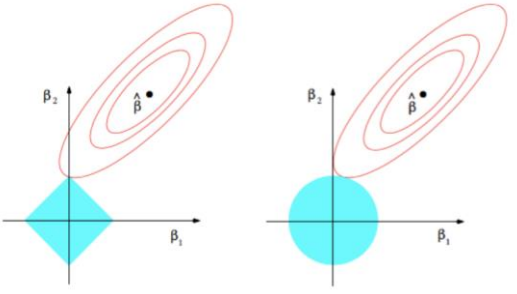
LASSO and Ridge Regression

- Ridge Regression

$$\text{minimize} \sum_{i=1}^{n}(y_i - \beta^\top \mathbf{z}_i)^2 \text{ s.t. } \sum_{j=1}^{p}\beta_j^2 \leq t$$

- Ridge Regression can be helpful when Z is highly correlated
  - $(Z^TZ)^{-1}$ does not exist, or is very sensitive to noise
  - $(Z^TZ + \lambda I_p)$ is always inversible.
- But Ridge Regression just shrinks variables, it does not select variables

$$\hat{\beta}_\lambda^{\text{ridge}} = (\mathbf{Z}^\top\mathbf{Z} + \lambda\mathbf{I}_p)^{-1}\mathbf{Z}^\top\mathbf{y}$$



Feature Selection and Engineering
Optimality?

In theory the goal is to find an optimal set of features, one that maximizes the scoring function…

In real world applications this is usually not possible

- For most problems it is computationally intractable to search the whole space of possible feature subsets
- One usually has to settle for approximations of the optimal subset
- Most of the research in this area is devoted to finding efficient search-heuristics

Summary

- Feature engineering
  - Categorical variables:
    - One-hot encoding
    - Risk values of categorical variables
  - Recency, Frequency, and Monetary (RFM) framework
- Feature selection
  - Filter based: Mutual Information
  - Step-wise: Forward, Backward, Both
  - Embedded: LASSO (L1 Regularization)

Feature Engineering, and Feature Selection

Lesson 6

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON