Variable Selection and Regularization

Christoph Kern

Mannheim Machine Learning Modules

c.kern@uni-mannheim.de

March 21 and 22, 2018

UNIVERSITY OF MANNHEIM

Outline

Introduction

Selecting Features for Prediction

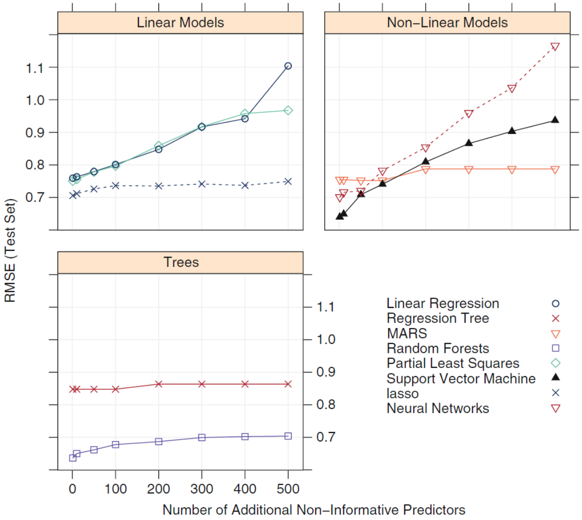
- Prediction problems might involve data with many potential features (and less domain knowledge)
  - Contrasts theory-based hypothesis testing perspective
  - Relates to building parsimonious models (e.g. with respect to AIC, BIC)
- Interest in developing sparse models to improve:
  - Model interpretability
  - Prediction accuracy
    - Performance in a new **test set** given estimation in **training data**

$\rightarrow$ Remove non-informative variables to improve effectiveness

Is feature selection necessary?

- Regression models
  - Run into problems in high dimensions (large p, small n)
    - $p \approx n$: Overfitting and poor prediction performance
    - $p > n$: OLS can not be estimated
- Tree-based models
  - Involve build-in feature selection
  - Less affected by irrelevant features
- Support vector machines, neural networks
  - Negatively affected by irrelevant features

Figure: Consequences of non-informative predictors



Kuhn & Johnson (2013)

Feature selection methods

- Wrapper
  - Search algorithms that add and/or remove predictors to optimize performance
  - e.g. forward, backward, and best subset selection
- Filter
  - Test individual predictors outside of the predictive model
  - e.g. t-tests, r, χ^2
- ℓ_1 **regularization**

Stepwise Variable Selection

Algorithm 1: Classical forward selection

Set p-value threshold τ;
Initialize empty model;
repeat
    for *each predictor not in the model* **do**
        Add the predictor to the current model;
        Estimate the statistical significance of the new term;
    end
    if *the smallest p is less than τ* **then**
        Include the corresponding predictor in the model;
    end
until *no significant predictor remains outside the model*;

There are a number of **serious** problems here!

1. Multiple testing issue
2. Objective function does not focus on prediction accuracy
3. Prone to performance evaluation bias

Adjusting stepwise selection approaches

1. Usage of performance measures instead of *p*-values
2. Implement feature selection in a (proper) **resampling setting**
3. Interweave feature selection in the model-building process

Algorithm 2: Forward selection with resampling

Set the number of resampling iterations;
Set the number of features p;
Initialize empty model;
for *each resampling iteration* **do**
    Partition data into training and hold-out set;
    for $k = 0, \ldots, p - 1$ **do**
        Consider all $p - k$ models that add an additional predictor to the current model;
        Choose the best among these models in terms of loss in the training data;
    end
end
Determine the best model over all hold-out sets;

Regularization

Penalized regression models

- (Even) regression models can be over-parameterized (large p, small n)
- Shrinkage / Regularization methods
  - Consider model complexity in the estimation process by...
  - ...shrinking regression coefficients towards zero

$\rightarrow$ Ridge regression & Lasso

OLS regression

$$\hat{\beta}_{OLS} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 \right\}$$

→ Minimizes ("only") RSS

Ridge regression

$$\hat{\beta}_{ridge} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$
$$= RSS + \lambda \sum_{j=1}^{p} \beta_j^2$$

→ Introduces a **shrinkage penalty**: Fit - complexity trade-off

OLS regression

$$\hat{\beta}_{OLS} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 \right\}$$

$\rightarrow$ Minimizes ("only") RSS

Ridge regression

$$\begin{aligned} \hat{\beta}_{ridge} &= \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\} \\ &= RSS + \lambda \sum_{j=1}^{p} \beta_j^2 \end{aligned}$$

$\rightarrow$ Introduces a **shrinkage penalty**: Fit - complexity trade-off

Comparing OLS and Ridge regression

$RSS = (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta)$

$RSS(\lambda) = (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) + \lambda\beta'\beta$

$$\hat{\beta}_{OLS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

$$\hat{\beta}_{ridge} = (\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}'\mathbf{y}$$

As a result...

- OLS regression needs $\mathbf{X}$ to be of full column rank
- Ridge regression (still) allows matrix inversion due to $\lambda\mathbf{I}$

Other penalties are possible

Ridge regression

- Penalty on ℓ_2 norm of $\boldsymbol{\beta}$
- $\|\boldsymbol{\beta}\|_2 = \sqrt{\sum_{j=1}^{p} \beta_j^2}$

Lasso (Least Absolute Shrinkage and Selection Operator)

$$\hat{\beta}_{lasso} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

- Penalty on ℓ_1 norm of β
- $\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$

Other penalties are possible

Ridge regression

- Penalty on ℓ_2 norm of $\boldsymbol{\beta}$
- $\|\boldsymbol{\beta}\|_2 = \sqrt{\sum_{j=1}^{p} \beta_j^2}$

Lasso (Least Absolute Shrinkage and Selection Operator)

$$\hat{\boldsymbol{\beta}}_{lasso} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

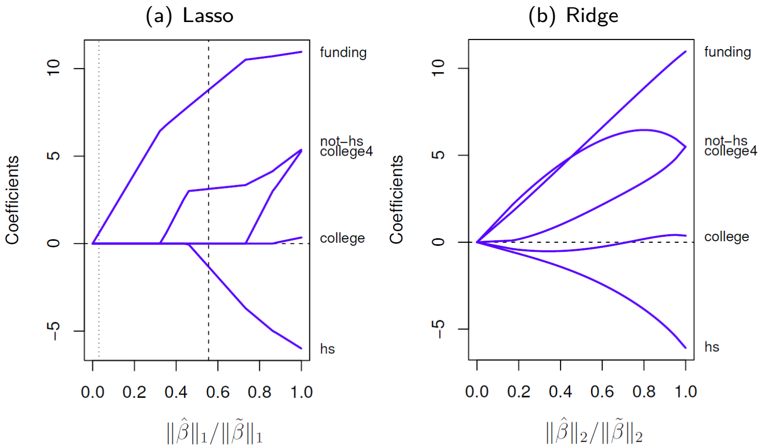
- Penalty on ℓ_1 norm of $\boldsymbol{\beta}$
- $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p} |\beta_j|$

Increasing the penalty on model complexity

- $\lambda = 0$
  - Models are equivalent to OLS
- $\lambda \to \infty$
  - Ridge regression ($RSS + \lambda\|\beta\|_2^2$)
    - Coefficients are shrunken towards zero
    - Shrinks coefficients of correlated predictors towards each other
  - Lasso ($RSS + \lambda\|\beta\|_1$)
    - Coefficients are eventually shrunken exactly to zero (i.e. performs **variable selection**)
    - Erratic paths for correlated predictors

$\rightarrow$ The penalty λ is a tuning parameter

Figure: Coefficient paths



Efron & Hastie (2016)

A compromise between ridge and lasso

$$\underset{\beta_0,\beta}{\text{minimize}}\left\{\frac{1}{2}\sum_{i=1}^{n}(y_i-\beta_0-x_i'\beta)^2+\lambda\left[\ \tfrac{1}{2}(1-\alpha)\|\beta\|_2^2+\alpha\|\beta\|_1\right]\right\}$$

Elastic net

- Introduces a mixing parameter $\alpha \in [0,1]$
  - $\alpha = 0$: Ridge regression
  - $\alpha = 1$: Lasso
- α is an additional tuning parameter

Standardization

- Contribution to penalty term dependent on scale
- Ridge and Lasso typically applied with standardized features

Group lasso

- Standard lasso considers predictors independently
- Group lasso in- or excludes groups of variables together

Categorical outcomes

$$\underset{\beta_0,\beta}{\text{minimize}} \left\{ -\frac{1}{n}\mathcal{L}(\beta_0, \beta; \mathbf{y}, \mathbf{X}) + \lambda \|\beta\|_1 \right\}$$

- Shrinkage can be applied with binary and multinomial outcomes...
- ...by introducing a penalty in the likelihood function

Tuning and Cross-Validation

Lasso regression modeling process

1. Choose a series of λ values
2. Estimate a sequence of penalized regression models
   - Since we are interested in the best prediction model for new data...
   - ...this sequence is estimated in a Cross-Validation loop
3. Choose the best λ based on step 2
4. Re-fit model with chosen λ on full training data

$\rightarrow$ Data is split into training and validation set(s) for **model tuning**

Cross-Validation with the Lasso

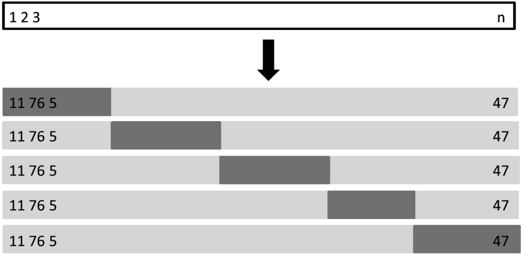
1. Split the data into k sets at random
2. Fit a sequence of regularized models using $k-1$ parts of the data
3. Estimate model performances on the holdout set
4. Repeat step 2 & 3 k times

Cross-validated errors (κ indicates data partitions)

$$CV(\hat{f}_\lambda) = \frac{1}{n}\sum_{i=1}^{n} L(y_i, \hat{f}_\lambda^{-\kappa(i)}(x_i))$$

with $L(y_i, \hat{y}_i) = (y_i - \hat{y}_i)^2$ for regression problems.

Figure: 5-Fold Cross-Validation with training set and validation set (example)

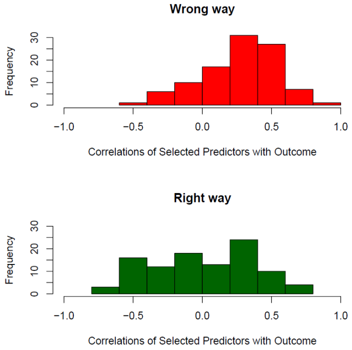


James et al. (2013)

Cross-Validation done wrong

- **Never** separate feature selection and CV
  - CV *after* selection on full data *biases* performance measures
  - Hold-out samples are no longer independent test sets
- Include feature selection within the CV loop
- Unsupervised screening on full data is valid

Figure: Correlations of y with unrelated x's with incorrect and correct CV



Hastie et al. (2009)

Summary

- Ridge, lasso and elastic net penalize model complexity
  - Can be used to fit sparse and stable models
  - Typically applied in large p, small n situations
  - Utilize Cross-Validation for parameter tuning
- Statistical inference after feature selection?
  - Selection needs to be taken into account (Taylor & Tibshirani 2015)!
- Beware of pitfalls when applying (stepwise) variable selection

Software Resources

Resources for R

- Stepwise selection e.g. available in `leaps` and `klaR` package
- Standard package for ridge regression, lasso and elastic net: `glmnet`
- Group lasso penalization implemented in `grpreg` and `gglasso`
- Tools for post-selection inference: `selectiveInference`

References

Efron, B. and Hastie, T. (2016). Sparse Modeling and the Lasso. In Efron, B. and Hastie, T. (Eds.), *Computer Age Statistical Inference: Algorithms, Evidence and Data Science* (pp. 298–324). New York, NY: Cambridge University Press

Hastie, T., Tibshirani, R., Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, NY: Springer.

Hastie, T., Tibshirani, R., Wainwright, M. (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. Boca Raton, FL: Chapman & Hall/CRC.

James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). *An Introduction to Statistical Learning*. New York, NY: Springer.

Kuhn, M., Johnson, K. (2013). *Applied Predictive Modeling*. New York, NY: Springer.

Taylor, J. and Tibshirani, R. (2015). Statistical Learning and Selective Inference. *Proceedings of the National Academy of Sciences, 112*(25), 7629–7634.