Week04 - Summary

Introduction to Feature and Model Selection

Goals of Feature Selection

- **Prediction accuracy:** especially when $p > n$
  - Where p is the number of features and n denotes number of observations
- **Model interpretability**
  - Removing irrelevant or poor features (that is, by setting the corresponding coefficient estimates to zero) → we can obtain a model that is more easily interpreted

Approaches for Feature Selection

1. **Subset selection**
   - Identify a subset of the p predictors that we believe to be related to the response or class (y)
   - FIt a classification or regression model on the reduced set of variables

2. **Shrinkage**
   - It is primarily used for regression models
   - Fit a model involving all p predictors
   - Some coefficients are shrunk towards zero
   - This shrinkage (also known as regularisation) has the effect of reducing variance and can also be used for feature selection

3. **Dimension reduction**

- We project the p predictors into M-dimensional subspace, $M < p$

Linear Model (Feature) Selection

- The model containing all of the predictors will always have the smallest RSS, since these quantities are related to the training error
- We wish to choose a model with low test error, not a model with low training error

Indirect: Mallows C_p and BIC

Mallow's $C_p = \frac{1}{n}(RSS + 2d\hat{\sigma}^2)$

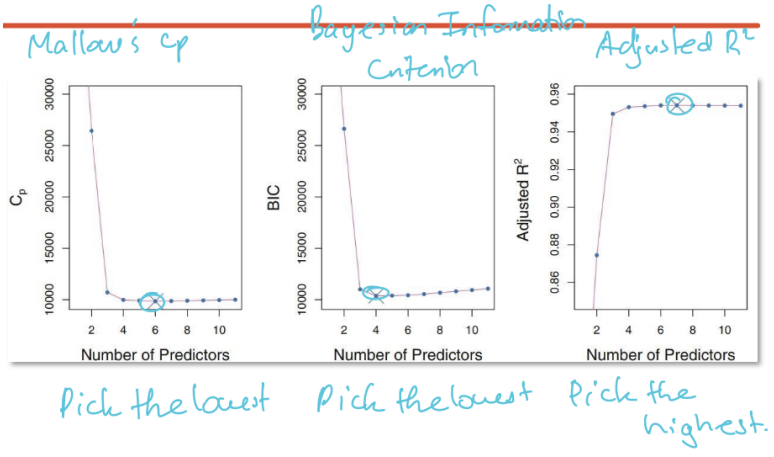
- d is the total number of predictors
- $\hat{\sigma}^2$ is an estimate of the variance of ε

Bayesian information criterion: BIC $= \frac{1}{n}(RSS + \log(n)d\hat{\sigma}^2)$

- Like C_p, the BIC will tend to take on small value for model with a low test error, and so generally we select model that has the lowest BIC value

Notice that BIC replaces the $2d\hat{\sigma}^2$ used by C_p with a $\log(n)d\hat{\sigma}^2$ term, where n is the number of samples

Since $\log n > 2$ when $n > 7$, the BIC statistic generally places a heavier penalty on models with many variables, and hence results in the selection of smaller models than C_p



Direct: Test Set and Cross-Validation

- Each of the procedures returns a sequence of models M_k indexed by model size $k = 0, 1, 2, \ldots$ Our job here is to select k. Once selected, we will return model M_k
- We compute the validation set error or the cross-validation error for each model M_k .
    - Select the k for which the resulting estimated test error is smallest
- This procedure has an advantage relative to C_p and BIC, in that it provides direct estimate of the test error and doesn't require an estimate of the error variance σ^2

Stepwise Methods

Forward Stepwise Selection

- Forward stepwise selection begins with a model containing no predictors, and then adds predictors to the model, one-at-a-time, until all of the predictors are in the model
- In particular, at each step the variable that gives the greatest additional improvement to the fit is added to the model

Backward Stepwise Selection

- Like forward stepwise selection, backward stepwise selection provides an efficient alternative to best subset selection
- However, unlike forward stepwise selection;
    - Begins with the full model containing all p predictors
    - Iteratively removes the least useful predictor, one-at-a-time

Exhaustive Searches

- Exhaustive search is the only technique guaranteed to find the predictor variable subset with the best evaluation criterion

- Since we look over the whole model space, we can identify the best model(s) at each model size
- Sometimes known as the best subsets model selection
- Loss component is (typically) the residual sum of squares
- Main drawback: exhaustive searching is computationally intensive

Best Subset Selection

- Consider as an example linear regression
  - $\mathcal{M}_0: Y = \beta_0 + \varepsilon$ null model
  - $\mathcal{M}_p: Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$ full model

1. Denote $\mathcal{M}_0$ to be the null model
   - Contains no predictors
2. For $k = 1,2, \ldots, p$
   - Fit all $\binom{p}{k}$ models that contain exactly k predictors
   - Denote $\mathcal{M}_k$ the best among the $\binom{p}{k}$ models
     - Measured as best against some metric (smallest residual sum of squares or highest accuracy etc.)
3. Select the single best model among the $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_p$
   - Using cross-validated prediction error or residual sum of squares etc.

Feature and Model Selection

Shrinkage or Regularisation Methods

- Regularisation methods shrink estimated regression coefficients by imposing a penalty on their sizes
- There are different choices for the penalty
- The penalty choice drives the properties of the method
- Helpful to find solutions for ill-posed problems or to prevent overfitting

Penalty

- Recall that the least squares fitting procedure estimates $\beta_0, \beta_1, \ldots, \beta_p$ using the values that minimise:

$$RSS = \sum_{i=1}^{n}\left(Y_i - \beta_0 - \sum_{j=1}^{p}\beta_j X_{ij}\right)^2 = ||Y - \mathbf{X\beta}||_2^2$$

- Shrinkage methods seek to minimise

  $RSS + \lambda R(\beta)$

  for some tuning parameter λ and penalty function R

Ridge Regression

- The ridge regression coefficient estimates $\hat{\beta}_R$ are the values that minimise

$$RSS + \lambda \sum_{j=1}^{p} \beta_j^2$$

where $\lambda \geq 0$ is a **tuning** parameter, to be determined separately

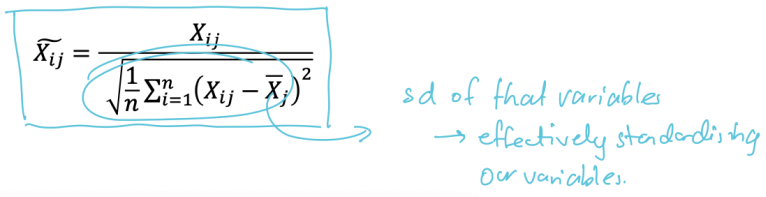
- As with least squares, ridge regression seeks coefficient estimates that fit the data well, by making the RSS small
  - However, the second term, $\lambda \sum_{j=1}^{p} \beta_j^2$, is called a shrinkage penalty.
    - Is small when $\beta_1, \ldots, \beta_p$ are close to zero, and so it has the effect of shrinking the estimates of β_j towards zero
- The tuning parameter λ serves to control the relative impact of these two terms on the regression coefficient estimates
- Selecting a good value for λ is critical; cross-validation can be used for this

Ridge Regression: Scaling of Predictors

- The standard least squares coefficient estimates are scale invariant

- Multiplying X_j by a constant c simply leads to a scaling of the least squares coefficient estimates by a factor of $1/c$
    - In other words, regardless of how the j^{th} predictor is scaled, X_j, $\hat{\beta}_j$ will remain the same

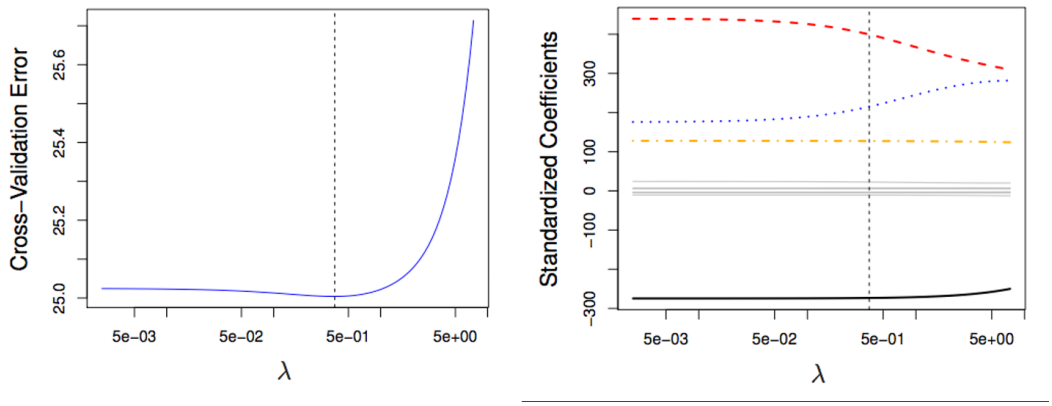
- In contrast, the ridge regression coefficients estimates can change substantially when multiplying a given predictor by a constant
    - Due to the sum of squared coefficients term in the penalty part of the ridge regression objective function
- Therefore, it is best to apply ridge regression after standardising the predictors, using a formula such as below:



Selecting Tuning Parameters

- As for subset selection we require a method to determine which of the models under consideration is the best
- That is, we require a method selecting a value for the tuning parameter λ or equivalently
- Cross-validation provides a simple way to tackle this problem. We choose a grid of λ values, and compute the cross-validation error rate for each value of λ
- We then select the tuning parameter value for which the cross-validation error is smallest
- Finally, the model is re-fit using of the available observations and the selected value of the tuning parameter

Credit Data Example



First plot:

- Cross-validation errors that result from applying ridge regression to the credit data set with a range of λ values

Second plot:

- Coefficient estimates as a function of λ. The vertical dashed lines indicate the best value λ selected by cross-validation

Ridge Penalty

- Trades some bias in the parameter estimates for reduced variance
- Particularly useful with highly correlated predictors
- Does not perform feature selection: all variables are still included in the model

Lasso

- The Lasso is a relatively recent alternative to ridge regression that can be used for feature selection
- The lasso coefficient minimise the quantity:

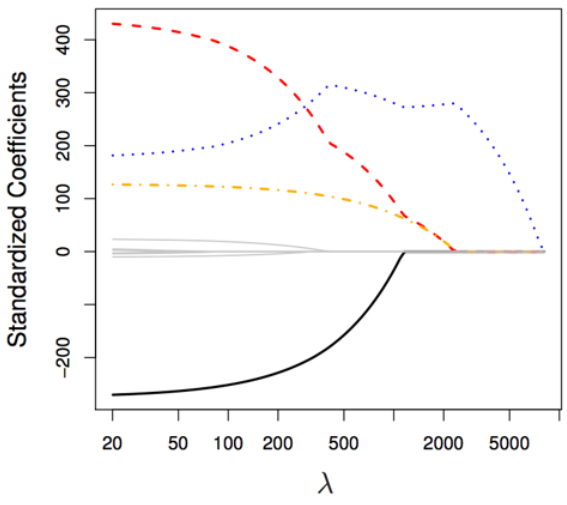
$$RSS + \lambda \sum_{j=1}^{p} |\beta_j|$$

- The lasso uses an ℓ_1 penalty instead of the ℓ_2 penalty used for ridge regression
  - The ℓ_1 norm of a coefficient vector:

$$\boldsymbol{\beta} \text{ is } ||\boldsymbol{\beta}||_1 = \sum_{j=1}^{p} |\beta_j|.$$

- As with ridge regression, the lasso shrinks the coefficient estimates towards zero
- However, in the case of the lasso, ℓ_1 penalty has the effect of forcing some of the coefficient estimates to be exactly equal to zero when the tuning parameter λ is sufficiently large
- Hence, much like best subset selection, the lasso performs **feature selection**
- We say that the lasso yields **sparse** models —that is, models that involve only a subset of variables
- As in ridge regression, selecting a good value of λ for the lasso is critical; cross-validation is again the method of choice

Example: Credit Data Set



Extension: Elastic Net

- A more general model, the Elastic net, combines the ridge and the lasso penalties
- It solves the following penalised minimisation problem:

$$\text{argmin}_{\boldsymbol{\beta}} ||Y - \mathbf{X}\boldsymbol{\beta}||_2^2 + \lambda\left((1-\alpha)/2||\boldsymbol{\beta}||_2^2 + \alpha||\boldsymbol{\beta}||_1\right)$$

- Can consider it a weighted combination (mixture) of ℓ_1 and ℓ_2 penalties