Model selection and regularization

- Linear regression model: $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$.
- When $n \gg p$, provided that there is a linear relationship between response and predictors, least square estimators have low bias and low variance.
- When $p \geq n$, more variability in the least squares fit, variance can be infinite. Also computational issues.
- Overfitting: Choose a model with low bias and high variance. The model will fit the training data very well, however, due to high variance it does poor on unseen observations.
- Model interpretability: Many variables used in MLR model are infact not associated with the response. Removing these varaibles can help interpreting the model.

Subset Selection

- Identify a subset of the p predictors that could be related to the response. Then a regression model can be fit using this subset of predictors.
- Fit a separate regression model using each possible combination of the p predictors. There are 2^p of them.
- Algorithm: Best subset selection
  1. Let M_0 be the null model. The model simply uses the sample mean to predict for each observation.
  2. For $k = 1, 2, \cdots, p$: Fit all $\binom{p}{k}$ models that contains exactly k predictors. Pick the best one (with smallest RSS or R^2), call it M_k.
  3. Select the best model from $M_0, M_1, \cdots, M_p$ using a validation set.

Forward stepwise selection

- Computationally efficient alternative to best subset selection by considering a much smaller set of models.
- Algorithm:
  1. Let M_0 be the null model, with no predictors.
  2. For $k = 0, \cdots, p-1$: Consider $p-k$ models that augment the predictors in M_k with one additional predictor. Choose the best one (with smallest RSS or highest R^2), call it M_{k+1}.
  3. Select the best model among $M_0, \cdots, M_p$ using cross validated prediction error or adjusted R^2.

Backward stepwise selection

- Start with a full model, then iteratively removes the least useful predictor one at a time.
- Algorithm:
    1. Let M_p be the full model, with all P predictors.
    2. Let $k = p, p-1, \cdots, 1$: Consider all k models that contain all but one of the predictors in M_k (k-1 of them), choose the best among these k models , call it M_{k-1} (the one with smallest RSS or higest R^2).
    3. Select the best model among $M_0, \cdots, M_p$ using cross validated prediction error or adjusted R^2.

Regularizatin methods

- Best subset selection, forward stepwise, and backward stepwise can still be very slow.
- Alternative: Fit a linear regression model with certain constraints or regularization on the coefficient estimates.
- When $p \gg n$, the normal equation $X^T X\beta = X^T Y$ might be hard to solve computationally. Regularization techniques could be a remedy for this problem.

Ridge regression

- Vector norm: $\|x\|_p = (\sum_{i=1}^{n} |x_i|^p)^{1/p}$
- Euclidean norm: $\|x\|_2 = \sqrt{\sum_{i=1}^{n} |x_i|^2}$

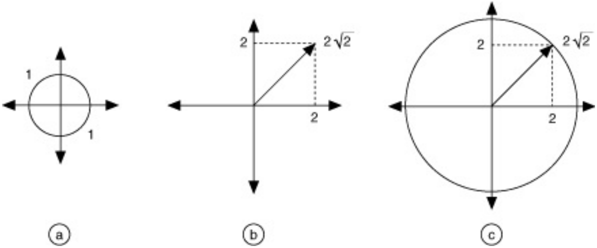


Figure: http://zone.ni.com/

- min $\frac{1}{2}\|y - X\beta\|_2^2$ s.t: $\|\beta\|_2^2 \leq s$. for some $s > 0$.
- Basicly guessing where the solution could be.

- The problem above is equivalent to:

$$\min \frac{1}{2}\|y - X\beta\|_2^2 + \frac{1}{2}\lambda\|\beta\|_2^2, \lambda \geq 0.$$

- $\lambda \geq 0$ is a tuning parameter.
- The first part of objective functioc: $\frac{1}{2}\|y - X\beta\|_2^2$: look for parameters β that fit the data well.
- The second part: $\frac{1}{2}\lambda\|\beta\|_2^2$ control the impact of the penalty term. When $\lambda = 0$, its MLR. When λ is large, it tends to shrink coefficients to zero.

Ridge regression

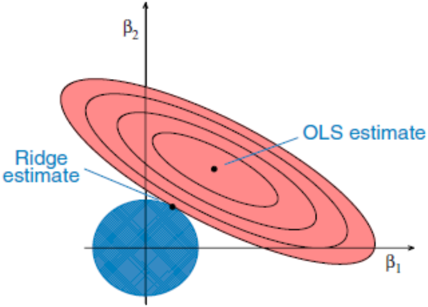
- 

Figure: https://onlinecourses.science.psu.edu

- Notice that the shrinkage penalty is applied to $\beta_1, \beta_2, \cdots, \beta_p$ only, but not the intercept β_0

Ridge regression solution

- For now, lets ignore the intercept term:

$$\min f(\beta) = \frac{1}{2}\|Y - X\beta\|_2^2 + \frac{\lambda}{2}\langle \beta, \beta \rangle$$

- Take derivative:

$$\begin{aligned}\frac{\partial f}{\partial \beta} &= X^T X\beta - X^T Y + \lambda\beta \\ &= (X^T X + \lambda\mathbb{I})\beta - X^T Y.\end{aligned}$$

$\mathbb{I}$ is the identity matrix.

- Set this derivative to 0: $(X^T X + \lambda\mathbb{I})\beta = X^T Y \rightarrow \beta = (X^T X + \lambda\mathbb{I})^{-1} X^T Y.$

Ridge regression solution

- When we consider the intercept β_0, the problem will be:

$$\min \frac{1}{2}\|y - X\beta - \beta_0\|_2^2 + \frac{\lambda}{2}\|\beta\|_2^2$$

Ridge regression and gradient descent method

- Gradient descent method for ridge regression.

Why ridge regression improve over least squares

- As parameter λ increasees, the flexibility of ridge regression model decreases, leading to decreased variance but increased bias.
- Remember bias-variance trade off?
- At one extreme λ very small, close to 0, "my guess is the solution is in this very big sphere". At the other extreme, for a large value of λ, the solution is restricted to a very small sphere, thus flexibility decreases.
- Shrinkage of ridge solution leads to substantial reduction in the variance of predictions, at the expense of a slight increase in bias.

The lasso

- Ridge solution will include all p predictors in the final model.
- The final model will be hard to interpret since it is very likely that only a small number of predictors are important.
- The lasso (least absolute shrinkage and selection operator):

$$\min \frac{1}{2}\|y - X\beta\|_2^2 \text{ s.t: } \|\beta\|_1 \leq s.$$

$$\min \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1., \lambda \geq 0.$$

Lasso

- $\|\beta\|_1 = |\beta_1| + |\beta_2| + \cdots + |\beta_p|$
- The lasso shrinks the coefficient estimates towards 0. When λ is sufficiently large enough, some of the coefficients will be forced to be exactly equal to 0.
- The lasso can perform variable selection.
- The final model will be sparse - involve only a subset of the original set of predictors.

Lasso

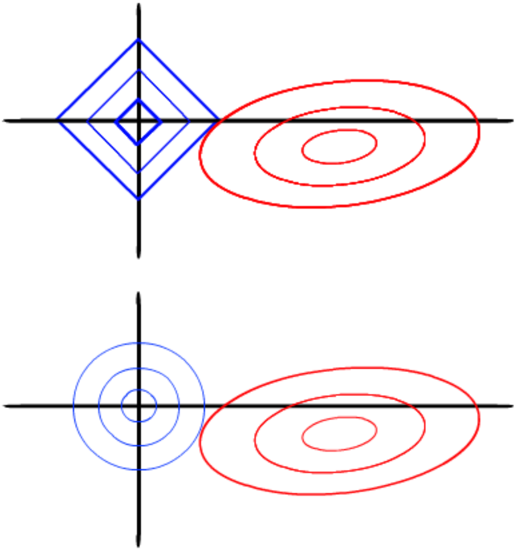


Figure: http://stats.stackexchange.com/

Ridge vs. Lasso

- If you have evidence that all the predictors you have in hand are important, ridge solution is better.
- If only a small subset of the predictors are important, lasso will be better.
- Elastic net regularization:

$$\min \frac{1}{2}\|y - X\beta\|_2^2 + \frac{\lambda_2}{2}\|\beta\|_2^2 + \lambda_1\|\beta\|_1, \lambda_1 \geq 0, \lambda_2 \geq 0.$$

- Elastic net combines the advantages of lasso and ridge regression.

Solution to lasso

- Unlike ridge regression, lasso does not have an explicit solution.
- Lasso solution can be found using iterative methods.
- Gradient descent method revisited.