Advanced Regression: Some Notes on Regularization

Máster Data Science

Advanced Regression

How to improve simple linear models when the dimension is high?

- Try to improve the interpretability while attaining good predictive performance
- Need to replace least squares with some alternative fitting tools
- Need to balance prediction accuracy versus model interpretability (feature selection)

Collinearity

- Phenomenon due to redundant information about the response because some predictors in the model are correlated
- Usually, more information (more variables) is not necessary better, because adding more variables will result in inability to visualize that information
- That implies a close-to-singularity matrix $X'X$
- That implies overfitting
- That implies beta's are estimated with high noise
- That implies confusing and misleading results about the effects

Collinearity

- In high dimension (p/n is large), collinearity appears almost sure
- Estimation of beta's not reliable
- Predictions for response not reliable if predictors are outside the range of historial data. But reliable within the range
- Hence, how can we estimate with some accuracy β?
- And, are all the p variables really needed to predict y?

To explain or to predict?

In regression/classification, there are three sources of uncertainty:

- The error in the coefficients when the linear approximation is true (estimation error)
- The error in the linear approximation when the true model is non-linear, or contains other variables (model bias)
- The noise in the DGP: Data $=$ Model $+$ Noise (irreducible error)

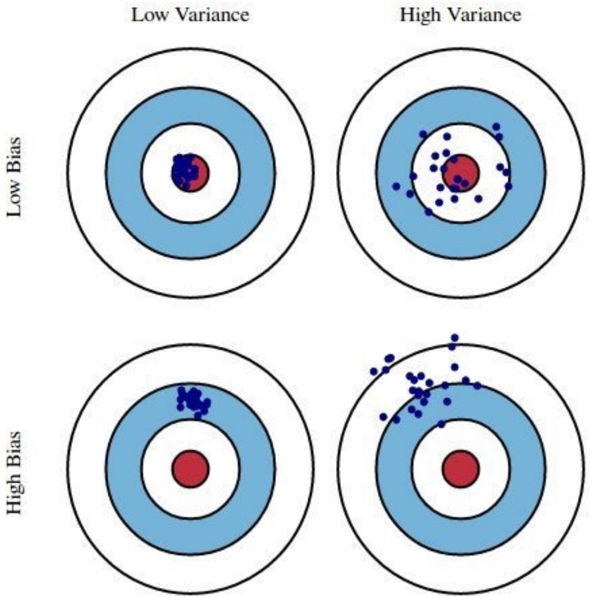
$$\boxed{(\text{Prediction Error})^2 = \sigma^2 + \text{Bias}^2 + \text{Var}}$$

To explain or to predict?

$$\boxed{(\text{Prediction Error})^2 = \sigma^2 + \text{Bias}^2 + \text{Var}}$$

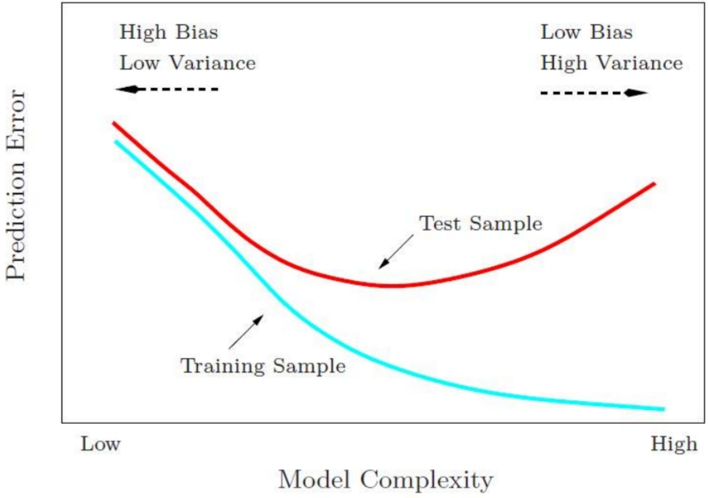
- Statistics:
  - Focus on minimizing Bias (by assuming knowledge about population, DGP)
  - Hence, able to obtain formulas for Var that provides explanation (inference, effects of predictors on response)
  - The Var can be large in practice
- Machine Learning:
  - Focus on minimizing $\text{Bias}^2 + \text{Var}$
  - No assumptions needed (discover knowledge), hence no formulas and no explanation
  - But good prediction performance

Predictions: Bias vs Variance



Overfitting

When a model fits or predict very well the training data bat bad the testing data (p is large compared with n)



Overfitting and underfitting in practice

- Consider the following true data generating process (DGP):

$$y = x_1\beta_1 + \cdots + x_p\beta_p + \epsilon = x'_{\text{true}}\beta + \epsilon$$

where $E(\epsilon) = 0$ and $\text{Var}(\epsilon) = \sigma^2$. The corresponding OLS estimator is denoted by $\hat{\beta}_{\text{true}}$

- **Overfitting**: the model is estimated with more variables than needed ($q > p$ variables):

$$y = x_1\beta_1 + \cdots + x_q\beta_q + \epsilon = x'_{\text{over}}\beta_{\text{over}} + \epsilon$$

If $\hat{\beta}_{\text{over}}$ denotes the OLS estimator, then the prediction is unbiased but with larger variance:

$$E(x'_{\text{over}}\hat{\beta}_{\text{over}}) = x'_{\text{true}}\beta, \quad \text{Var}(x'_{\text{over}}\hat{\beta}_{\text{over}}) \geq \text{Var}(x'_{\text{true}}\hat{\beta}_{\text{true}})$$

- **Underfitting**: the model is estimated with less variables than needed ($q < p$ variables). Then, the prediction is biased but with smaller variance, i.e.

$$E(x'_{\text{under}}\hat{\beta}_{\text{under}}) \neq x'_{\text{true}}\beta, \quad \text{Var}(x'_{\text{under}}\hat{\beta}_{\text{under}}) \leq \text{Var}(x'_{\text{true}}\hat{\beta}_{\text{true}})$$

Overfitting and Collinearity

Collinearity increases the overfitting effect
Detection:

- Some variables are significant in simple regressions but non-significant in multiple regressions
- The p-value for the F-test is significant but many p-values for t-tests are insignificant
- I.e. we can trust the global F-test but not the individual t-tests
- Large condition number of $X'X$ (i.e. greater than 30)

Regression Tools in High Dimension

- Variable selection
- Regularization (shrinkage estimation)
- Dimension Reduction

- **Regularization Methods**

Regularization Methods

- How can we estimate with some accuracy β?
- Main idea: penalize coefficient estimates, i.e. shrink them to 0
- With this framework, no need to select variables previously
- Main tools: Ridge, Lasso, Elastic Net, etc.

Ridge Regression

- Ridge regression: used in high dimension to mitigate overfitting (even if $p > n$)
- Also known as Tikhonov regularization: ill-conditioned problems

$$\text{minimize} \quad ||y - X\beta||_2^2 + \rho||\beta||_2^2$$

  where ρ is a tuning parameter, to be calibrated separately
- Explicit solution: $\hat{\beta} = (X^T X + \rho I)^{-1} X^T y$
- It adds some bias to the estimation to reduce a lot the variance: better MSE than OLS
- It is better the data matrix X is centered previously (no estimation of β_0, we do not want to shrink it). Then, $\hat{\beta}_0 = \bar{y}$
- It is also better to standardize the data, in order to make the estimation scale-invariant
- Low computational cost, good prediction accuracy, but dense solution (no variable selection)

The Lasso

- Lasso regression: used in high dimension to mitigate overfitting (even if $p > n$)
- L_1 regularization: sparse solutions

$$\text{minimize}_\beta \quad \frac{1}{2}||y - X\beta||_2^2 + \rho||\beta||_1$$

- No explicit solution: non-differentiable problem
- It adds some bias to the estimation to reduce a lot the variance: better MSE than OLS
- Attains sparsity (model selection at the same time)
- Again, it is convenient the data matrix X is centered previously (no estimation of β_0, we do not want to shrink it). Then, $\hat{\beta}_0 = \bar{y}$
- State-of-the-art tool in Big Data Analytics

The Lasso

- Efficient equivalent (differentiable) formulation:

$$\begin{aligned} \text{minimize}_{t,\beta} \quad & ||y - X\beta||_2^2 + \rho t^T e \\ \text{subject to} \quad & -t \leq \beta \leq t \\ & t \geq 0 \end{aligned}$$

- Many ways to estimate the lasso regression: solving previous quadratic optimization problem, solving the original (but non-differentiable) problem, using the LARS, using coordinate descent, . . .

- With non-prior information, ridge regression attains less variance than lasso (with similar bias)

- If real model is sparse, then lasso performs better

Cardinality constraints or L_0 regression

- Based on 0-norm penalty:

$$\text{minimize}_\beta \quad \frac{1}{2}||y - X\beta||_2^2 + \rho||\beta||_0$$

  where $||\beta||_0$ is the number of non-zero elements in β

- $||\cdot||_0$ is not really a norm...

- This formulation is equivalent to best subset selection

- Can improve computational efficiency by using good MIP techniques (optimization)

- Not quite stable in practice

Regularization: Elastic Net

- Based on 1 and 2-norm penalties:

$$\text{minimize}_\beta \quad \frac{1}{2}||y - X\beta||_2^2 + \rho_1||\beta||_1 + \rho_2||\beta||_2^2$$

- The 1-norm controls sparsity
- The 2-norm stabilizes the regularization path
- But we need to calibrate two parameters...
- Recommended packages for regularization tools: glmnet (in R), scikit.learn (in Python)

Referencias:

- Korosteleva O. (2004). *Advanced Regression Models with SAS and R* Chapman and Hall/CRC. ISBN: 978-1-138-04901-7
- Zou, H., Hastie, T. (2005). *Regularization and variable selection via the elastic net*. J. R. Statistic Soc. B, 67(2), 301.
- Hastie, T., Tibshirani, R., Wainwright, M. (2015) *Statistical Learning with Sparsity: the Lasso and Generalizations*, CRC Press. ISBN: 978-1-498-71216-3 pdf here
- Hastie, T., Tibshirani, R., Friedman, J. (2009). *The elements of Statistical Learning. Data Mining, Inference and Prediction.* Springer. ISBN: 978-0-387-84858-7.