Regression analysis and resampling methods

Nicholas Karlsen and Thore Espedal Moe
University of Oslo
(Dated: September 24, 2020)

An abstract abstract.

I. INTRODUCTION

In essence, Linear Regression is the process of taking points from a function, or a set of measurements and mapping them to coordinates in a choosen basis in order to create an approximation, or model of your original dataset.

II. THEORY

See Hastie *et al.* [1]

A. Linear Regression

- describe the general problem
- brief description of design matrix
- brief introduction to the cost function
- discuss different choices of bases. explain why $\mathbb{P}_n$ is often a good choice. Perhaps also touch on overfitting.

1. Ordinary Least Squares

In ordinary least squares, we aim to find an optimal set of parameters $\hat{\boldsymbol{\beta}} = [\hat{\beta}_0, \dots, \hat{\beta}_n]^T$ such that the L^2 norm $\|\mathbf{y} - \mathrm{X}\boldsymbol{\beta}\|_2$ is minimal, where the L^2 norm is induced by the inner product

$$\|\mathbf{u}\|_2^2 = \sum_i u_i^2 = \mathbf{u}^T\mathbf{u} \tag{1}$$

This defines the cost function for OLS, which may be written as

$$C_{OLS}(\boldsymbol{\beta}) = (\mathbf{y} - \mathrm{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathrm{X}\boldsymbol{\beta}) \tag{2}$$

In order to find its minima, we differentiate wrt to $\boldsymbol{\beta}$ and assert that $\partial_{\boldsymbol{\beta}} C_{OLS} = 0$ for the optimal predictor. Taking the partial derivative yields

$$\frac{\partial}{\partial \boldsymbol{\beta}} C_{OLS}(\boldsymbol{\beta}) = -2\mathrm{X}^T(\mathbf{y} - \mathrm{X}\boldsymbol{\beta}) \tag{3}$$

Which we then set to 0

$$\mathrm{X}^T\mathbf{y} = \mathrm{X}^T\mathrm{X}\boldsymbol{\beta} \tag{4}$$

taking the inverse of $\mathrm{X}^T\mathrm{X}$ on both sides then yields the optimal $\boldsymbol{\beta}$ as

$$\hat{\boldsymbol{\beta}} = (\mathrm{X}^T\mathrm{X})^{-1}\mathrm{X}^T\mathbf{y} \tag{5}$$

2. Ridge Regression

$$C_R(\boldsymbol{\beta}) = \|\mathbf{y} - \mathrm{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2 \tag{6}$$

3. Lasso Regression

$$C_L(\boldsymbol{\beta}) = \|\mathbf{y} - \mathrm{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 \tag{7}$$

B. Singular Value Decomposition

- Discuss problems of X^TX becoming singular in OLS, and how we use SVD to work around it.

C. Resampling

Resampling methods are ways in which we can generate new statistics from our existing data, which as the name suggests implies sampling new data sets from our already existing data. By doing so, we may gain new insigts about our data which may not be available through regular analysis, particularly in situations where we are limited by the number of data points.

Here, we will focus on two of many such techniques.

1. Cross Validation

In the cross-validation resampling method, we split our data set S into k equally sized subsets $s_1, \dots, s_k$. We then for each $i = 1, \dots, k$ assign the i-th subset as the test set and the remaining $k-1$ subsets as the training set and compute the statistics in the usual way. Then at the end, we compute the mean value of the k sets of statistics. A visual representation of this process can be seen in Fig. 1. When doing cross-validation, typical choices of k are 5 and 10 [1]. Which one is better will depend on how the error scales with the size of the training set, as

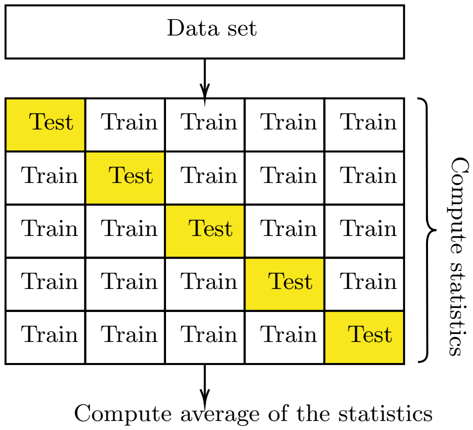

FIG. 1: Visual representation of k-fold cross-sampling for $k = 5$

such, choosing a suitable k requires some analysis. The cross-validation resampling provides a good estimate for the mean error of our estimates.

- Discuss this in more detail $\rightarrow \Delta Err$ wrt to number of data points

2. *Bootstrap*

In the bootstrap resampling method, we sample our data set $S = \{s_1, \dots s_N\}$ N-number of times, in particular, we allow sampling the same s_i multiple times. In this way, we generate new datasets in which some points are underweighted and others overweighted with respect to the original dataset S. So we effectively are generating small pertubations of the original dataset, which we may then use to compute new statistics. In particular, this technique enables us to investivate the variance of our model wrt small pertubations of the predictors. **This may be worded more elegantly**

- touch on rates of convergence

D. The Bias-Variance Tradeoff

III. RESULTS

IV. DISCUSSION

V. CONCLUSION

[1] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed. (2009).