Lecture #5: Regularization

CS 109A, STAT 121A, AC 209A: Data Science

Pavlos Protopapas   Kevin Rader

Lecture Outline

Review

Behind Ordinary Lease Squares, AIC, BIC

Regularization: LASSO and Ridge

Bias vs Variance

Regularization Methods: A Comparison

Review

Model Selection

Model selection is the application of a principled method to determine the complexity of the model, e.g. choosing a subset of predictors, choosing the degree of the polynomial model etc.

A strong motivation for performing model selection is to avoid overfitting, which we saw can happen when

- there are too many predictors:
  - the feature space has high dimensionality
  - the polynomial degree is too high
  - too many cross terms are considered
- the coefficients values are too extreme

Stepwise Variable Selection and Cross Validation

Last time, we addressed the issue of selecting optimal subsets of predictors (including choosing the degree of polynomial models) through:

- stepwise variable selection - iteratively building an optimal subset of predictors by optimizing a fixed model evaluation metric each time,
- cross validation - selecting an optimal model by evaluating each model on multiple validation sets.

Today, we will address the issue of discouraging extreme values in model parameters.

Behind Ordinary Lease Squares, AIC, BIC

Likelihood Functions

We've been using AIC/BIC to evaluate the explanatory powers of models, and we've been using the following formulae to calculate these criteria

$$\text{AIC} \approx n \cdot \ln(\text{RSS}/n) + 2J$$
$$\text{BIC} \approx n \cdot \ln(\text{RSS}/n) + J \cdot \ln(n)$$

where J is the number of predictors in model.

Intuitively, AIC/BIC is a loss function that depends both on the predictive error, RSS, and the complexity of the model. We see that we prefer a model with few parameters and low RSS.

But why do the formulae look this way - what is the justification?

Likelihood Functions

Recall that our statistical model for linear regression in vector notation is

$$y = \beta_0 + \sum_{j=1}^{J} \beta_i x_i + \epsilon = \boldsymbol{\beta}^\top \boldsymbol{x} + \epsilon.$$

It is standard to suppose that $\epsilon \sim \mathcal{N}(0, \sigma^2)$. In fact, in many analyses we have been making this assumption. Then,

$$y | \boldsymbol{\beta}, \boldsymbol{x}, \epsilon \sim \mathcal{N}(\boldsymbol{\beta}^\top \boldsymbol{x}, \sigma^2).$$

Can you see why?

Note that $\mathcal{N}(y; \boldsymbol{\beta}^\top \boldsymbol{x}, \sigma^2)$ is naturally a function of the model parameters $\boldsymbol{\beta}$, since the data is fixed. We call

$$\mathcal{L}(\boldsymbol{\beta}) = \mathcal{N}(y; \boldsymbol{\beta}^\top \boldsymbol{x}, \sigma^2)$$

the ***likelihood function***, as it gives the likelihood of the observed data for a chosen model $\boldsymbol{\beta}$.

Maximum Likelihood Estimators

Once we have a likelihood function, $\mathcal{L}(\beta)$, we have strong incentive to seek values of β to maximize $\mathcal{L}$.

Can you see why?

The model parameters that maximizes $\mathcal{L}$ are called ***maximum likelihood estimators (MLE)*** and are denoted:

$$\beta_{MLE} = \underset{\boldsymbol{\beta}}{\operatorname{argmax}}\, \mathcal{L}(\beta)$$

The model constructed with MLE parameters assigns the highest likelihood to the observed data.

Maximum Likelihood Estimators

But how does one maximize a likelihood function?

Fix a set of n observations of J predictors, **X**, and a set of corresponding response values, **Y**; consider a linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \epsilon$.

If we assume that $\epsilon \sim \mathcal{N}(0, \sigma^2)$, then the likelihood for each observation is

$$\mathcal{L}_i(\boldsymbol{\beta}) = \mathcal{N}(y_i; \boldsymbol{\beta}^\top \boldsymbol{x}_i, \sigma^2)$$

and the likelihood for the entire set of data is

$$\mathcal{L}(\boldsymbol{\beta}) = \prod_{i=1}^{n} \mathcal{N}(y_i; \boldsymbol{\beta}^\top \boldsymbol{x}_i, \sigma^2)$$

Through some algebra, we can show that maximizing $\mathcal{L}(\boldsymbol{\beta})$ is equivalent to minimizing MSE:

$$\boldsymbol{\beta}_{MLE} = \underset{\boldsymbol{\beta}}{\text{argmax}}\ \mathcal{L}(\boldsymbol{\beta}) = \underset{\boldsymbol{\beta}}{\text{argmin}} \frac{1}{n} \sum_{i=1}^{n} |y_i - \boldsymbol{\beta}^\top \boldsymbol{x}_i|^2 = \underset{\boldsymbol{\beta}}{\text{argmin}}\ RSS$$

Minimizing MSE or RSS is called ***ordinary least squares***.

Information Criteria Revisited

Using the likelihood function, we can reformulate the information criteria metrics for model fitness in very intuitive terms.

For both AIC and BIC, we consider the likelihood of the data under the MLE model against the number of explanatory variables used in the model

$$g(J) - \mathcal{L}(\boldsymbol{\beta}_{MLE})$$

where g is a function of the number of predictors J. Individually,

$$AIC = J - \ln(\mathcal{L}(\boldsymbol{\beta}_{MLE}))$$

$$BIC = \frac{1}{2}J \ln(n) - \ln(\mathcal{L}(\boldsymbol{\beta}_{MLE}))$$

In the formulae we'd been using for AIC/BIC, we approximate $\mathcal{L}(\boldsymbol{\beta}_{MLE})$ using the RSS.

Regularization: LASSO and Ridge

Regularization: An Overview

The idea of regularization revolves around modifying the loss function L; in particular, we add a ***regularization term*** that penalizes some specify properties of the model parameters

$$L_{reg}(\beta) = L(\beta) + \lambda R(\beta),$$

where λ is a scalar that gives the weight (or importance) of the regularization term.

Fitting the model using the modified loss function L_{reg} would result in model parameters with desirable properties (specified by R).

LASSO Regression

Since we wish to discourage extreme values in model parameter, we need to choose a regularization term that penalizes parameter magnitudes. For our loss function, we will again use MSE.

Together our regularized loss function is

$$L_{LASSO}(\beta) = \frac{1}{n}\sum_{i=1}^{n} |y_i - \boldsymbol{\beta}^\top \boldsymbol{x}_i|^2 + \lambda \sum_{j=1}^{J} |\beta_j|.$$

Note that $\sum_{j=1}^{J} |\beta_j|$ is the ℓ_1 norm of the vector $\boldsymbol{\beta}$

$$\sum_{j=1}^{J} |\beta_j| = \|\boldsymbol{\beta}\|_1$$

Hence, we often say that L_{LASSO} is the loss function for ***ℓ_1 regularization***.

Finding model parameters $\boldsymbol{\beta}_{LASSO}$ that minimize the ℓ_1 regularized loss function is called ***LASSO regression***.

Ridge Regression

Alternatively, we can choose a regularization term that penalizes the squares of the parameter magnitudes.

Then, our regularized loss function is

$$L_{Ridge}(\beta) = \frac{1}{n}\sum_{i=1}^{n} |y_i - \boldsymbol{\beta}^\top \boldsymbol{x}_i|^2 + \lambda \sum_{j=1}^{J} \beta_j^2.$$

Note that $\sum_{j=1}^{J} \beta_j^2$ is related to the ℓ_2 norm of $\boldsymbol{\beta}$

$$\sum_{j=1}^{J} \beta_j^2 = \|\boldsymbol{\beta}\|_2^2$$

Hence, we often say that L_{Ridge} is the loss function for ***ℓ_2 regularization***.

Finding model parameters $\boldsymbol{\beta}_{Ridge}$ that minimize the ℓ_2 regularized loss function is called ***ridge regression***.

Choosing λ

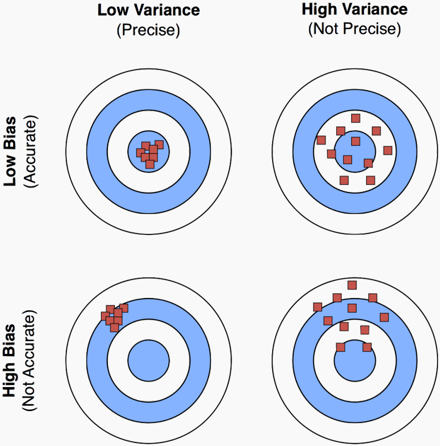
In both ridge and LASSO regression, we see that the larger our choice of the ***regularization parameter*** λ, the more heavily we penalize large values in β,

1. If λ is close to zero, we recover the MSE, i.e. ridge and LASSO regression is just ordinary regression.
2. If λ is sufficiently large, the MSE term in the regularized loss function will be insignificant and the regularization term will force β_{Ridge} and β_{LASSO} to be close to zero.

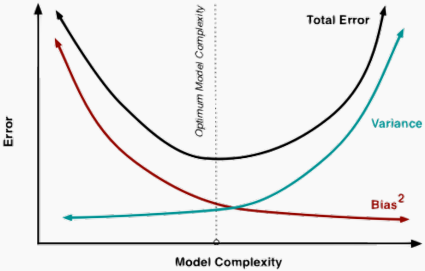
To avoid ad-hoc choices, we should select λ using cross-validation.

Bias vs Variance

Bias vs Variance

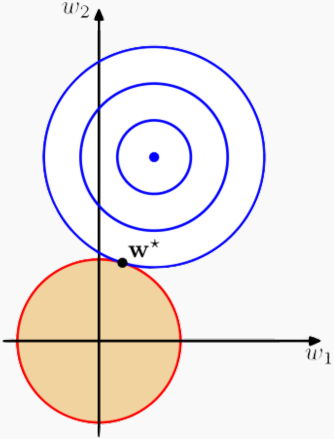


The Bias/Variance Trade-off



Regularization Methods: A Comparison

The Geometry of Regularization



Variable Selection as Regularization

Since LASSO regression tend to produce zero estimates for a number of model parameters - we say that LASSO solutions are ***sparse*** - we consider LASSO to be a method for variable selection.

Many prefer using LASSO for variable selection (as well as for suppressing extreme parameter values) rather than stepwise selection, as LASSO avoids the statistic problems that arises in stepwise selection.

An Comparative Example

Bibliography

1. Bolelli, L., Ertekin, S., and Giles, C. L. ***Topic and trend detection in text collections using latent dirichlet allocation***. In European Conference on Information Retrieval (2009), Springer, pp. 776-780.
2. Chen, W., Wang, Y., and Yang, S. ***Efficient influence maximization in social networks. In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (2009)***, ACM, pp. 199-208.
3. Chong, W., Blei, D., and Li, F.-F. ***Simultaneous image classification and annotation. In Computer Vision and Pattern Recognition***, 2009. CVPR 2009. IEEE Conference on (2009), IEEE, pp. 1903-1910.
4. Du, L., Ren, L., Carin, L., and Dunson, D. B. ***A bayesian model for simultaneous image clustering, annotation and object segmentation***. In Advances in neural information processing systems (2009), pp. 486-494.
5. Elango, P. K., and Jayaraman, K. ***Clustering images using the latent dirichlet allocation model.***
6. Feng, Y., and Lapata, M. ***Topic models for image annotation and text illustration***. In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (2010), Association for Computational Linguistics, pp. 831-839.
7. Hannah, L. A., and Wallach, H. M. ***Summarizing topics: From word lists to phrases.***
8. Lu, R., and Yang, Q. ***Trend analysis of news topics on twitter***. International Journal of Machine Learning and Computing 2, 3 (2012), 327.