Lecture Outline

Review

Applications of Model Selection

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

Cross Validation

Predictor Selection: Cross Validation

Rather than choosing a subset of significant predictors using stepwise selection, we can use K-fold cross validation:

- create a collection of different subsets of the predictors
- for each subset of predictors, compute the cross validation score for the model created using only that subset
- select the subset (and the corresponding model) with the best cross validation score
- evaluate the model one last time on the test set

Degree Selection: Stepwise

We can frame the problem of degree selection for polynomial models as a predictor selection problem: which of the predictors $\{x, x^2, \ldots, x^M\}$ should we select for modeling?

We can apply stepwise selection to determine the optimal subset of predictors.

Degree Selection: Cross Validation

We can also select the degree of a polynomial model using K-fold cross validation.

- consider a number of different degrees
- for each degree, compute the cross validation score for a polynomial model of that degree
- select the degree, and the corresponding model, with the best cross validation score
- evaluate the model one last time on the test set

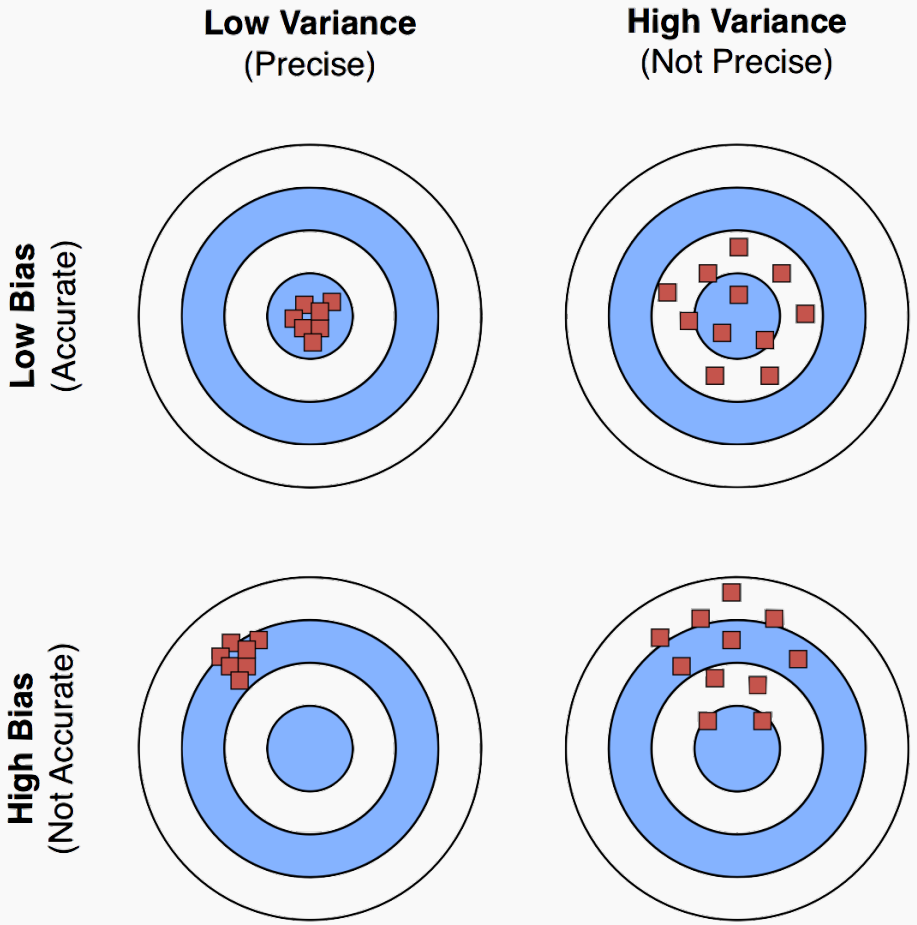
kNN Revisited

Recall our first simple, intuitive, non-parametric model for regression - the kNN model. We saw that it is vitally important to select an appropriate k for the data.

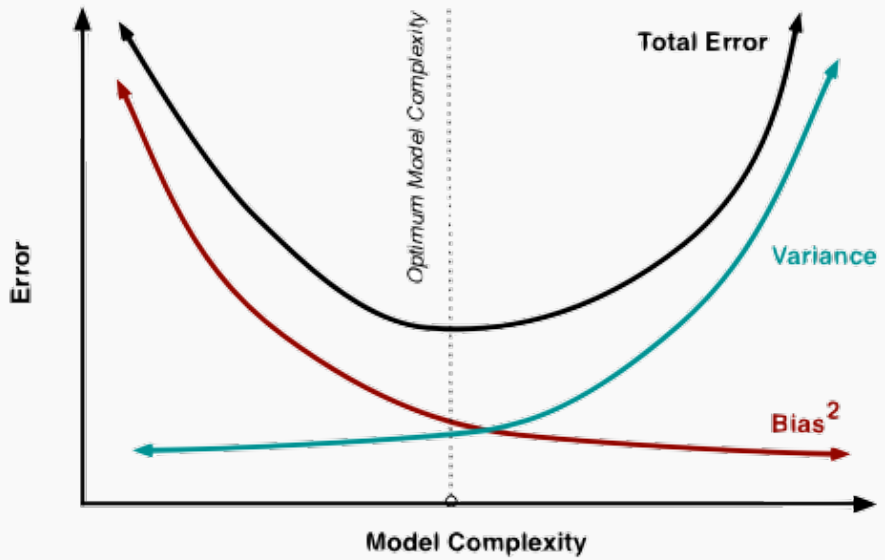
If the k is too small then the model is very sensitive to noise (since a new prediction is based on very few observed neighbors), and if the k is too large, the model tends towards making constant predictions.

A principled way to choose k is through K-fold cross validation.

Bias vs Variance



The Bias/Variance Trade-off



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 specified 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 $\boldsymbol{\ell_1}$ ***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 $\boldsymbol{\ell_2}$ ***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.

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.