

## Lecture 8: Variable selection and model verification

---

CSI 5v93: Introduction to machine learning

Baylor University  
Computer Science Department

Dr. Greg Hamerly  
<http://cs.baylor.edu/~hamerly/>

CSI 5v93: Introduction to machine learning, Lecture 8 – p. 1/18

## Announcements

---

- Homework 2 due February 8th

CSI 5v93: Introduction to machine learning, Lecture 8 – p. 2/18

## Questions?

---

Questions from last time:

- For stepwise selection, why not use the  $t$  statistic instead of the  $F$  statistic? – see homework, exercise 3.1
- More examples – coming as we move along
- How do we choose the best model using ridge regression?
- How do we use ridge regression?
- What is cross-validation? How does it work?
- How can correlated variables negate one another?

## Chapter 3: Linear methods for regression

---

- 3.1 – Introduction
- 3.2 – Linear regression models and least squares
- 3.3 – Multiple regression from simple univariate regression
- 3.4 – Subset selection and coefficient shrinkage
- 3.5 – Computational considerations

## Coefficient shrinkage

---

Subset selection of variables:

- eliminates variables completely by testing if their coefficients are sufficiently close to zero
- trying all subsets is a difficult, combinatorial problem

An alternative is to “shrink” the coefficient values so that variables become less important.

This can turn the combinatorial, discrete problem of subset selection into a smoother problem which can be optimized easily.

This also solves the problem of correlated variables whose coefficients negate one another and have high variance.

## Ridge regression

---

Ridge regression adds a squared penalty to each coefficient:

$$\begin{aligned}RSS(\lambda, \beta) &= \sum_{i=1}^n \left( y_i - \beta_0 - \sum_{j=1}^d x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^d \beta_j^2 \\ &= (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta) + \lambda \beta^T \beta\end{aligned}$$

So we want to choose the  $\beta$  that minimizes the above:

$$\begin{aligned}\hat{\beta} &= \arg \min_{\beta} \left\{ \sum_{i=1}^n \left( y_i - \beta_0 - \sum_{j=1}^d x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^d \beta_j^2 \right\} \\ &= (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{y}\end{aligned}$$

Where  $\mathbf{I}$  is the  $d \times d$  identity matrix.

## The problem of correlated variables

---

Let  $y = \beta_1 x_1 + \beta_2 x_2$  be our model

If  $x_1$  and  $x_2$  are independent, then both  $\beta_1$  and  $\beta_2$  are necessary to represent  $y$ .

If  $x_1 = x_2$ , or  $x_1$  is some multiple of  $x_2$ , then they are *correlated*.

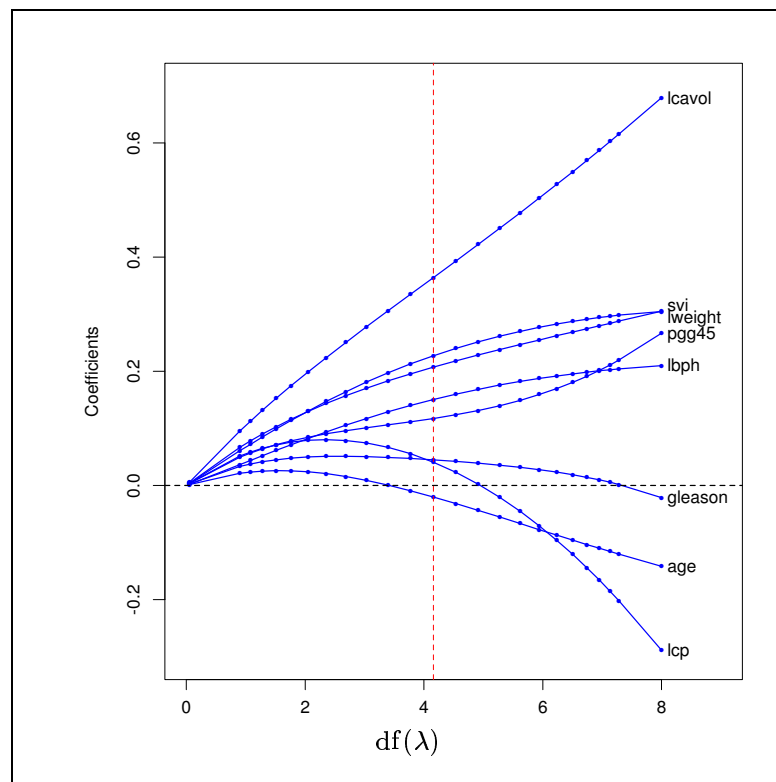
If they're correlated, then EITHER  $\beta_1$  or  $\beta_2$  could be used to accurately represent  $y$ . The  $\beta_1$  and  $\beta_2$  are not unique. This causes instability.

Ridge regression solves this by adding  $\lambda$  to the diagonal of  $\mathbf{X}^T \mathbf{X}$ .

CSI 5v93: Introduction to machine learning, Lecture 8 – p. 7/18

## Example of ridge regression

---



CSI 5v93: Introduction to machine learning, Lecture 8 – p. 8/18

## Model selection

How do we choose the best model for a machine learning problem?

Choosing a model means choosing the model **type** and the **parameters** for that model.

There are different model types:

- linear regression
- k-nearest neighbor
- many others we will look at

There are different parameters that we can select:

- use all parameters
- use a subset (subset selection)
- shrink the coefficients
- create new parameters as combinations of the old ones
- other ways to select parameters...

## Determining the best model

How do we determine which of two models is better?

- often we use some measure of the model's **accuracy**

There are lots of possibilities:

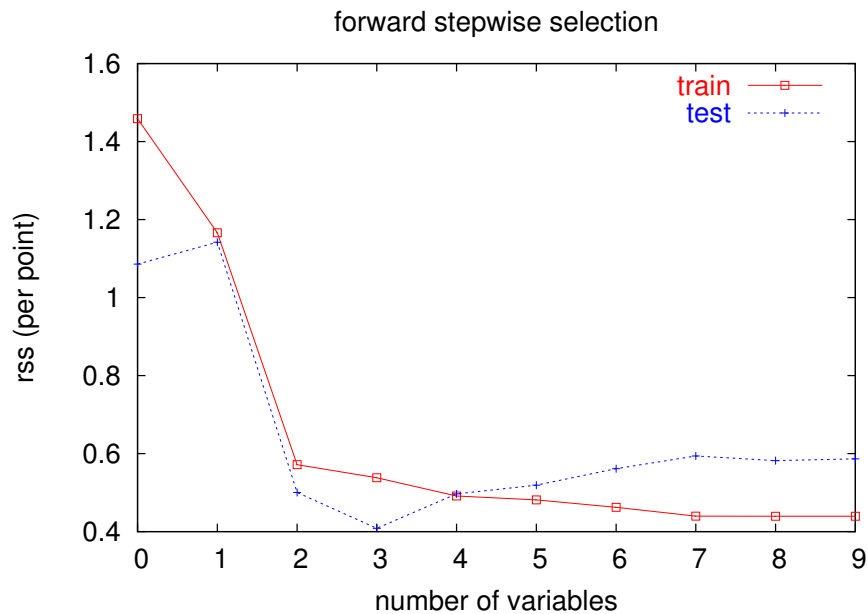
- sum-of-squared error (aka RSS, for regression problems)
- sum of absolute error (for regression problems)
- prediction accuracy (for classification problems)
- likelihood
- other?

All of these metrics could be measured on:

- the training set
- the test set
- a holdout validation set
- other?

## Example: RSS for model selection (cancer dataset)

---



Simple linear regression on the training data with different numbers of parameters. Forward stepwise selection uses RSS on training data to select the next parameter.

CSI 5v93: Introduction to machine learning, Lecture 8 – p. 11/18

## Verifying a model with held-out data

---

A simple and reasonable way to choose a model is based on “hold-out” data or a “validation set”.

Model selection with a hold-out set:

- Given training data  $X$
- Divide the training data into  $X_T$  and  $X_V$  (the training and validation sets)
- Learn different models on  $X_T$
- Test each model on  $X_V$
- Choose the model with the lowest error/highest accuracy on  $X_V$

What are the advantages and disadvantages to this approach?

CSI 5v93: Introduction to machine learning, Lecture 8 – p. 12/18

## Cross-validation

---

Cross-validation takes the hold-out set one step further.

Divide your training data  $X$  into  $k$  disjoint parts (aka folds):

$$X = X_1 \cup X_2 \cup \dots \cup X_k$$

Cross validation procedure:

- loop over  $i = 1 \dots k$ :
  - train on  $\{X - X_i\}$
  - validate on  $X_i$  (obtain an accuracy or error on  $X_i$ )
- report the  $k$  errors

It's common to use  $k = n$  (called leave-one-out cross validation), or  $k = 10$  (10-fold cross validation)

## Cross validation and choosing the model

---

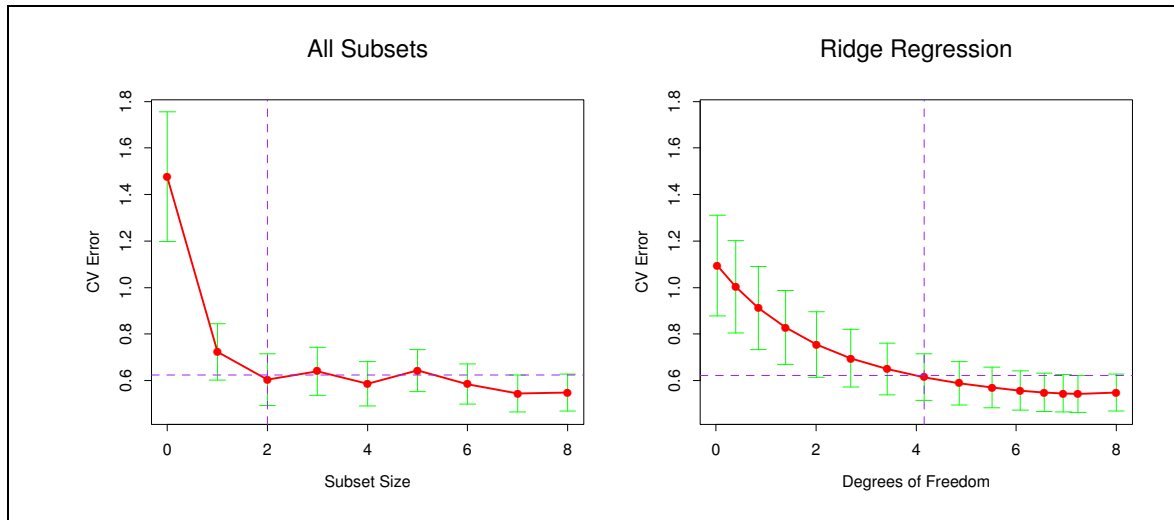
Cross validation produces  $k$  different errors on the  $k$  folds *for each model*.

Taking the average of the errors will give an idea of the expected error for other similar test sets.

Finding the standard deviation of the errors gives an idea of how much the model varies in its error.

## Examples of variable selection

---



Error bars are obtained by cross-validation.

The least complex model within 1 standard error of the best average error is chosen.

CSI 5v93: Introduction to machine learning, Lecture 8 – p. 15/18

## Cross validation, nearest neighbors, and repeats

---

Think about the 1-nearest neighbor classifier.

Imagine that your training data has 2 copies of every input example.

What if you use leave-one-out cross validation with this model and data?

CSI 5v93: Introduction to machine learning, Lecture 8 – p. 16/18

## Practical issues for homework 2

---

- center your data (input and output) before using ridge regression (see the book, p. 60)
- The first column of the data is the true label, so you don't want to use that value as an input, because it makes the problem trivial!
- Make sure you understand the ZIP data before you work with it.
- Confusion matrix

## 2-minute journal

---

Please write a response to the following on a piece of paper and hand it in immediately. Please make it anonymous (no names). Write about:

- major points you learned today
- areas not understood or requiring clarification