

# Lecture 11: Model and Variable Selection

Nancy R. Zhang

Statistics 203, Stanford University

February 16, 2010

## Training versus Test Performance

Given  $n \times p$  matrix  $X$  and response  $Y$ , we obtain a fitted model:

$$\hat{Y} = \hat{f}_{X,Y}(X).$$

For example, in the linear regression models we have been studying,

$$\hat{f}_{X,Y}(X) = X\hat{\beta}_{X,Y} = X(X'X)^{-1}X'Y.$$

The predicted values for  $Y$  are based on regression parameters that were fit using  $X, Y$ .

How would the model perform on unseen data?

$$EPE \equiv E[Y_{\text{new}} - \hat{Y}_{\text{new}}]^2 \quad ?$$

The expectation is taken over everything that is random:

$$X, Y, X_{\text{new}}, Y_{\text{new}}.$$

Good training performance does not imply good test performance.

## Bias-Variance Trade-off

$X$ ,  $Y$  used to fit the model are called “training data”, and “unseen” data used to estimate prediction error are called “test data”.

$$\text{Truth: } y = f(x) + \epsilon, \quad \text{Var}(\epsilon) = \sigma^2.$$

Estimate  $f(\cdot)$  using  $\hat{f}_{X,Y}$ .

$$\begin{aligned} \text{EPE} &\equiv E[Y_{\text{new}} - \hat{Y}_{\text{new}}]^2 \\ &= E[Y_{\text{new}} - \hat{f}_{X,Y}(X_{\text{new}})]^2 \\ &= E[(Y_{\text{new}} - f(X_{\text{new}})) + (f(X_{\text{new}}) - E\hat{f}_{X,Y}(X_{\text{new}})) + (E\hat{f}_{X,Y}(X_{\text{new}}) - \hat{f}_{X,Y}(X_{\text{new}}))]^2 \\ &= E[Y_{\text{new}} - f(X_{\text{new}})]^2 + E[f(X_{\text{new}}) - E\hat{f}_{X,Y}(X_{\text{new}})]^2 \\ &\quad + E[E\hat{f}_{X,Y}(X_{\text{new}}) - \hat{f}_{X,Y}(X_{\text{new}})]^2 \\ &= \sigma^2 + (\text{Model bias})^2 + \text{Model variance} \end{aligned}$$

As model complexity increases, bias *decreases* while variance *increases*. How to achieve a balance?

## Bias-Variance Trade-off

If you care more about  $\hat{\beta}$ :

$$\begin{aligned}MSE &\equiv E[\beta - \hat{\beta}]^2 = [E(\hat{\beta}) - \beta]^2 + E[\hat{\beta} - E(\hat{\beta})]^2 \\ &= \text{Bias}(\hat{\beta})^2 + \text{Var}(\hat{\beta}).\end{aligned}$$

For linear regression, **assuming that you've got the correct model**, the least squares estimates had 0 bias:

$$E[\hat{\beta}] = \beta,$$

so

$$MSE = \text{Var}(\hat{\beta}).$$

In a multiple regression, for any  $\hat{\beta}_i$ :

$$MSE(\hat{\beta}_i) = \text{Var}(\hat{\beta}_i) = \frac{\sigma^2}{\|r_{i, i}\|^2},$$

where  $r_{i, j}$  are the residuals of regression  $X_i$  on all other predictors.

# Estimating the prediction error

## 1 Asymptotic approximations

- ▶ Mallows  $C_p$ :

$$C_p = SSE + 2p\hat{\sigma}^2,$$

where  $p$  is the number of predictors. This is an unbiased estimate for  $EPE$ .

- ▶ Akaike's Information Criterion:

$$AIC = -2\loglik + 2p.$$

Reduces to  $C_p$  for linear models. Will be useful later for non-linear models.

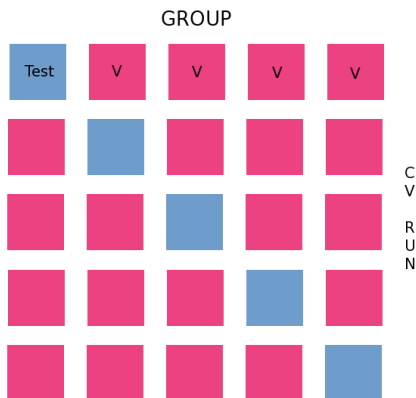
## 2 Cross-validation (next slide)

Select the model that *minimizes* the  $C_p$  or AIC.

The BIC is another useful model selection criterion, but is *not* based on prediction error (more later).

# Cross-Validation

Idea: Use a part of the data for training, the other part for testing.



## More on the $C_p$

$$C_p = SSE + 2p\hat{\sigma}^2.$$

Sometimes also written as:

$$C_p = SSE + 2p\hat{\sigma}^2 - n\hat{\sigma}^2,$$

but the last term doesn't change across models.

- 1  $\hat{\sigma}^2$  is usually estimated from the largest model.
- 2 Usual practice: plot  $C_p$  versus  $p$ , choose model with minimum.

# Bayes Information Criterion

Asymptotic approximation for

$$-2 \log P(\text{Model}|\text{Data}) \propto -2 \log \left\{ P(\text{Model}) \int_{\theta \in \Theta} P(\text{Data}|\theta) f(\theta) d\theta \right\}$$

$$\text{BIC} = -2l(\hat{\theta}) + p \log n,$$

where

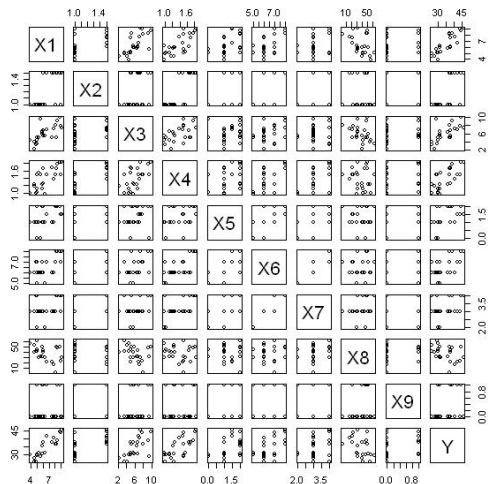
- $l(\cdot)$  is the log likelihood function. For linear regressions assuming Gaussian errors,

$$\text{BIC} \approx \text{SSE}/\hat{\sigma}^2 + p \log n.$$

- $\hat{\theta}$  is the likelihood maximizing value, for linear regressions this is the least squares estimate.
- $p$  is the number of parameters in the model.

Pick the model that minimizes the BIC.

# Exploring the model space



Property values data:

Y: Sales price  
X<sub>1</sub>: Local taxes  
X<sub>2</sub>: # bathrooms  
X<sub>3</sub>: Lot size  
X<sub>4</sub>: Living space  
X<sub>5</sub>: Garage  
X<sub>6</sub>: # rooms  
X<sub>7</sub>: # bedrooms  
X<sub>8</sub>: Property age  
X<sub>9</sub>: # fireplaces

$2^9 = 512$  possible models!

# Exploring the model space

- 1 Forward selection:
  - 1 Start with null model.
  - 2 Repeat: add variable with the most significant F-test.
  - 3 End when no variable has F-test p-value  $< \alpha$ .
- 2 Backward elimination:
  - 1 Start with full model.
  - 2 Repeat: delete variable with the least significant F-test.
  - 3 End when all variables have F-test p-value  $< \alpha$ .
- 3 Forward + Backward: Same as forward procedure, with option of deleting a variable at each step.
- 4 All subsets: possible when number of possible predictors is small ( $< 20$ ).

# Model Shrinkage Methods

- 1 Bias variance trade off:

$$EPE = \sigma^2 + (\text{Model bias})^2 + \text{Model variance}$$

$$MSE \equiv E[\beta - \hat{\beta}]^2 = \text{Bias}(\hat{\beta}) + \text{Var}(\hat{\beta}).$$

- 2 Mallows  $C_p$  statistic:

$$C_p = SSE + 2p\hat{\sigma}^2.$$

The second term is a “penalty” for model size.

- 3 Today: penalties based on  $\hat{\beta}$ .
  - 1 Ridge regression
  - 2 LASSO

# Ridge Regression

Ridge was developed first. It is based on the idea of constrained minimization:

$$\text{Minimize: } \sum_{i=1}^n \left( Y_i - \beta_0 - \sum_{j=1}^p X_{ij}\beta_j \right)^2$$

$$\text{Subject to: } \sum_{j=1}^p \beta_j^2 < C.$$

By the Lagrange multiplier method, this is equivalent to:

$$\text{Minimize: } \sum_{i=1}^n \left( Y_i - \beta_0 - \sum_{j=1}^p X_{ij}\beta_j \right)^2 + \lambda_C \sum_{j=1}^p \beta_j^2.$$

The second term is a penalty that depends on  $\|\beta\|^2$ .

# Ridge Regression

Ridge regression:

$$\text{Minimize: } \sum_{i=1}^n \left( Y_i - \beta_0 - \sum_{j=1}^p X_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^p \beta_j^2.$$

- 1 In statistics this is also called “shrinkage”: you are shrinking  $\|\beta\|^2$  towards 0.
- 2  $\lambda$  is a shrinkage parameter that you have to choose.
- 3 The Ridge solution  $\hat{\beta}_{\text{ridge}}$  is easy to solve, because the above is still a quadratic function in  $\beta$ .

# Ridge Solutions

Ridge loss function:

$$f(\beta) = \sum_{i=1}^n \left( Y_i - \beta_0 - \sum_{j=1}^p X_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^p \beta_j^2.$$

In matrix notation:

$$\begin{aligned} f(\beta) &= (Y - X\beta)'(Y - X\beta) + \lambda\beta'\beta \\ &= \beta'[X'X + \lambda I]\beta - \beta'X'Y - Y'X\beta + Y'Y \end{aligned}$$

Solving  $f'(\beta) = 0$  gives you:

$$\hat{\beta}_{\text{ridge}} = (X'X + \lambda I)^{-1}X'Y.$$

# Ridge Solutions

- 1 Whereas the least squares solutions  $\hat{\beta} = (X'X)^{-1}X'Y$  are unbiased if model is correctly specified, ridge solutions are *biased*

$$E[\hat{\beta}_{\text{ridge}}] \neq \beta.$$

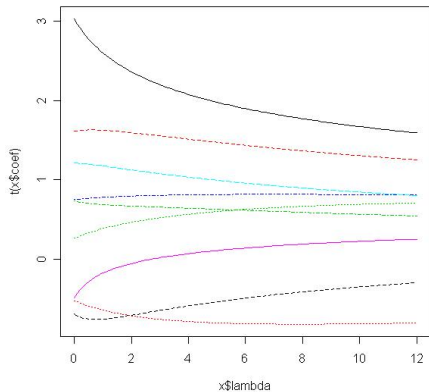
- 2 However, at the cost of bias, Ridge reduces the variance, and thus might reduce MSE.

$$MSE = Bias^2 + Variance$$

- 3 Ridge solutions are hard to interpret, because it is not sparse.

Sparse: some  $\beta_j$ 's are set exactly to 0.

# Ridge solutions versus lambda



## $L_1$ penalties

What if we constrain the  $L_1$  norm instead of the Euclidean norm?

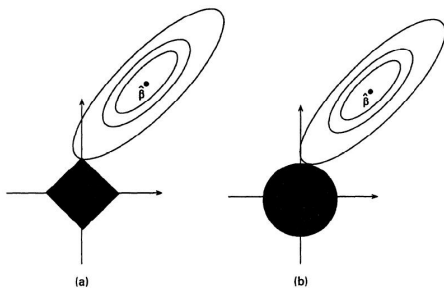
$$\text{Minimize: } \sum_{i=1}^n \left( Y_i - \beta_0 - \sum_{j=1}^p X_{ij} \beta_j \right)^2$$

$$\text{Subject to: } \sum_{j=1}^p |\beta_j| < C.$$

This is a subtle, but important change.

$$\text{Minimize: } \sum_{i=1}^n \left( Y_i - \beta_0 - \sum_{j=1}^p X_{ij} \beta_j \right)^2 + \lambda_C \sum_{j=1}^p |\beta_j|.$$

The above is termed Lasso regression (Tibshirani, 1996).



Comparing Lasso and Ridge, from Tibshirani (1996).

# Lasso

Lasso loss function is no longer quadratic, but is still convex:

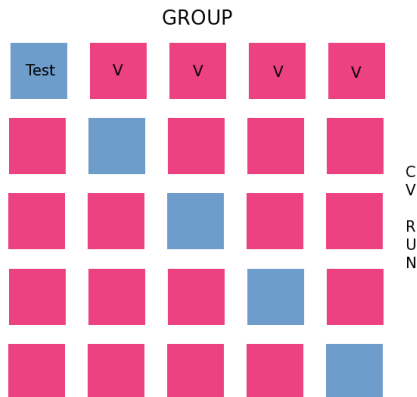
$$\text{Minimize: } \sum_{i=1}^n \left( Y_i - \beta_0 - \sum_{j=1}^p X_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^p |\beta_j|.$$

- 1 Unlike Ridge, there is no analytic solution for the LASSO.
- 2 Efron et al. (2002) gave an efficient algorithm `lars` to solve the Lasso.
- 3 Lasso solutions are sparse.

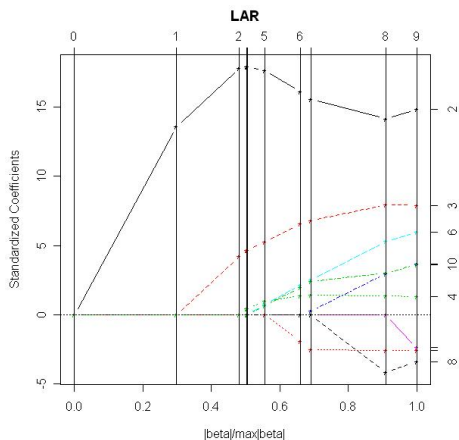
# Selecting the shrinkage parameter $\lambda$

$\lambda$  can be selected based on any of the model selection criteria we have discussed.

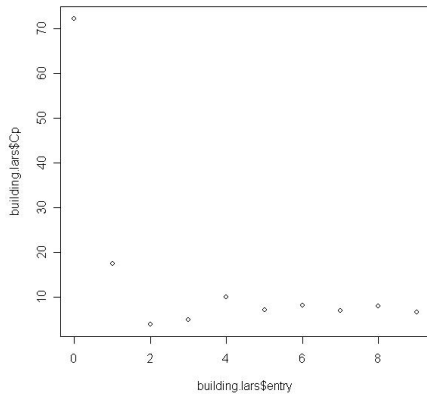
- 1  $C_p$  included in the output of `lars`.
- 2 Cross-validation (`cv.lars`).



# Lars output



# Lars $C_p$



# Lars cross validation prediction error

