

Hastie, Tibshirani & Friedman: Elements of Statistical Learning Chapter 7.1-7.9
Model Assessment and Selection

CN700/March 4, 2008

Satyavarta

sat@cns.bu.edu

Auditory Neuroscience Laboratory,
Department of Cognitive and Neural Systems,
Boston University, Boston MA 02215



Outline: Model Assessment and Selection

- ⇒ ● Choosing Model Complexity
 - Model Assessment
 - Other Loss Functions
 - Arriving at a Model
 - Comparing Model Classes
 - Choosing Model Complexity
 - Bias–Variance Decomposition
 - Choosing Model Complexity
 - Bias–Variance Decomposition: Special Cases
 - Bias and Variance
 - Bias-Variance Tradeoff in Model Complexity
 - Bias-Variance with 0-1 Loss

- Bias-Variance with 0-1 Loss: kNN
- Bias-Variance Tradeoff with 0-1 Loss: Regression
- Optimism
- AIC in model selection
- Estimates of Model Complexity: # of Parameters
- Estimates of Model Complexity: Vapnik Chernovenkis Dimension
- Shatter
- Example: Error of models picked by criteria relative to best model
- References

Choosing Model Complexity

Model Assessment

☞ Given X , estimate Y as \hat{f}

Model Assessment

④ Given X , estimate Y as \hat{f}

④ Loss function squared error: $L(Y, \hat{f}(X)) = (Y - \hat{f}(X))^2$

Model Assessment

④ Given X , estimate Y as \hat{f}

④ Loss function squared error: $L(Y, \hat{f}(X)) = (Y - \hat{f}(X))^2$

④ Loss function absolute error: $L(Y, \hat{f}(X)) = |Y - \hat{f}(X)|$

Model Assessment

- Given X , estimate Y as \hat{f}
- Loss function squared error: $L(Y, \hat{f}(X)) = (Y - \hat{f}(X))^2$
- Loss function absolute error: $L(Y, \hat{f}(X)) = |Y - \hat{f}(X)|$
- Test error $\text{Err} = E[L(Y, \hat{f}(X))]$

Model Assessment

④ Given X , estimate Y as \hat{f}

④ Loss function squared error: $L(Y, \hat{f}(X)) = (Y - \hat{f}(X))^2$

④ Loss function absolute error: $L(Y, \hat{f}(X)) = |Y - \hat{f}(X)|$

④ Test error $\text{Err} = E[L(Y, \hat{f}(X))]$ ④

Model Assessment

- Given X , estimate Y as \hat{f}
- Loss function squared error: $L(Y, \hat{f}(X)) = (Y - \hat{f}(X))^2$
- Loss function absolute error: $L(Y, \hat{f}(X)) = |Y - \hat{f}(X)|$
- Test error $\text{Err} = E[L(Y, \hat{f}(X))]$
- Training error $e\bar{r}r = \frac{1}{N} \sum_{i=1}^N L(y_i, \hat{f}(x_i))$

Model Assessment

- Given X , estimate Y as \hat{f}
- Loss function squared error: $L(Y, \hat{f}(X)) = (Y - \hat{f}(X))^2$
- Loss function absolute error: $L(Y, \hat{f}(X)) = |Y - \hat{f}(X)|$
- Test error $\text{Err} = E[L(Y, \hat{f}(X))]$
- Training error $e\bar{r}r = \frac{1}{N} \sum_{i=1}^N L(y_i, \hat{f}(x_i))$

Other Loss Functions

☛ 0–1 loss: $L(G, \hat{G}(X)) = I(G \neq \hat{G})(X)$

☛ Log likelihood loss: $L(G, \hat{p}(X)) = -2 \log \hat{p}_G(X)$

Arriving at a Model

- ② Model training
- ② Model Selection
- ② Model assessment

Data



Arriving at a Model

- ② Model training Training set
- ② Model Selection
- ② Model assessment



Arriving at a Model

- Model training Training set
- Model Selection Validation set
- Model assessment



Arriving at a Model

- ② Model training Training set
- ② Model Selection Validation set
- ② Model assessment Test set



Comparing Model Classes

☺ Data rich Validation



☺ Data poor Approximate validation

- ★ Analytically: AIC, BIC, MDL, SRM
- ★ Efficient Sample re-use: cross-validation, bootstrapping

Choosing Model Complexity

Bias–Variance Decomposition

Assumptions

★ $Y = f(X) + \epsilon, \epsilon \sim N(0, \sigma_\epsilon^2)$

★ Squared Loss error

Error

$$\begin{aligned} Err(x_0) &= E[(Y - \hat{f}(x_0))^2 | X = x_0] \\ &= \sigma_\epsilon^2 + [E\hat{f}(x_0)^2 - f(x_0)]^2 + E[\hat{f}(x_0) - E\hat{f}(x_0)]^2 \\ &= \sigma_\epsilon^2 + Bias^2(\hat{f}(x_0)) + Var(\hat{f}(x_0)) \end{aligned}$$

Choosing Model Complexity

Bias–Variance Decomposition: Special Cases

☛ k-nearest neighbor fit

$$Err(x_0) = \sigma_\epsilon^2 + \left[f(x_0) - \frac{1}{k} \sum_{l=1}^k f(x_l) \right]^2 + \sigma_\epsilon^2/k$$

☛ Linear model fit $\hat{f}(x) = \hat{\beta}^T x$

$$Err(x_0) = \sigma_\epsilon^2 + \left[f(x_0) - E\hat{f}_p(x_0) \right]^2 + \|h(x_0)\|^2 \sigma_\epsilon^2$$

☛ Linear Model family: further decomposition of bias

$$\begin{aligned} \left[f(x_0) - E\hat{f}_p(x_0) \right]^2 \\ \left[f(x_0) - E\hat{f}_\alpha(x_0) \right]^2 &= [f(x_0) - \beta_*^T x_0]^2 + [\beta_*^T x_0 - E\hat{\beta}_\alpha^T x_0]^2 \\ &= [\text{Model Bias}]^2 + [\text{Estimation Bias}]^2 \end{aligned}$$

Bias and Variance

Bias-Variance Tradeoff in Model Complexity

- Model Sample variance
- Best model (*) Total error, model bias, model variance
- Restricted model (restricted) estimation bias
- Choose restricted model if $B_{est} + B_* + \text{Var}_{restricted} < B_* + \text{Var}_*$

Figure: Hastie et al. 7.2

Bias-Variance with 0-1 Loss

Assumptions

★ $Y = f(X) + \epsilon, \epsilon \sim N(0, \sigma_\epsilon^2)$

★ **Squared Loss error**

Error

$$\begin{aligned} Err(x_0) &= E[(Y - \hat{f}(x_0))^2 | X = x_0] \\ &= \sigma_\epsilon^2 + [E\hat{f}(x_0)^2 - f(x_0)]^2 + E[\hat{f}(x_0) - E\hat{f}(x_0)]^2 \\ &= \sigma_\epsilon^2 + Bias^2(\hat{f}(x_0)) + Var(\hat{f}(x_0)) \end{aligned}$$

Bias-Variance with 0-1 Loss: kNN

Figures: Hastie et al. 7.3 Prediction error (red), bias² (green) and variance (blue)

Bias-Variance Tradeoff with 0-1 Loss: Regression _____

Figures: Hastie et al. 7.3 Prediction error (red), bias² (green) and variance (blue)

Optimism

☛ Training error $e\bar{r}r = \frac{1}{N} \sum_{i=1}^N L(y_i, \hat{f}(x_i))$

☛ True error $Err = E[L(Y, \hat{f}(X))]$

☛ Model adapts to data $\Rightarrow e\bar{r}r \leq Err$

☛ In-sample error $Err_{in} = \frac{1}{N} \sum_{i=1}^N E_y E_Y^{new} L(Y_i^{new}, \hat{f}(x_i))$

Y_i^{new} new responses observed at each training point $x_i, i = 1, 2, \dots, N$

☛ Optimism $op \equiv Err_{in} - E_y(e\bar{r}r)$

Estimating In-sample error using Optimism

② $Err_{in} = E_y(e\bar{r}r) + op$

② For squared error, 0–1 loss, and other loss functions

$$op = \frac{2}{N} \sum_{i=1}^N Cov(\hat{y}_i, y_i)$$

Tighter the data is fit, higher the optimism

② $Err_{in} = E_y(e\bar{r}r) + \frac{2}{N} \sum_{i=1}^N Cov(\hat{y}_i, y_i)$

② For linear fit with d inputs: $\sum Cov(\hat{y}_i, y_i) = d\sigma_\epsilon^2$

② $Err_{in} = E_y(e\bar{r}r) + \frac{2}{N} d\sigma_\epsilon^2$

Estimates of in-sample prediction error

☛ C_p statistic

$$C_p = E_y(e\hat{r}r) + \frac{2}{N}d\hat{\sigma}_\epsilon^2$$

☛ Estimate $\hat{\sigma}_\epsilon$ from a low-bias model

☛ For logistic regression, using binomial likelihood

$$e\hat{r}r = -\frac{2}{N}\loglik$$

$$Err_{in} = -\frac{2}{N}\loglik + 2\frac{d}{N}$$

$$\triangleq AIC$$

Estimates of in-sample prediction error (contd) _____

☛ Akaike Information Criterion (AIC): maximize likelihood \Rightarrow minimize $-\text{likelihood}$

☛ For Gaussian model,

$$AIC = C_p$$

☛ In general, for a family of models with tuning parameter α ,

$$AIC(\alpha) = e\hat{r}r + 2\frac{d(\alpha)}{N}\sigma_\epsilon^2$$

☛ $d(\alpha)$ is the effective number of parameters, e.g. $d(S) = \text{trace}(S)$

AIC in model selection

Figures: Hastie et al. 7.4

Pick model with smallest AIC

More Estimates of in-sample prediction error _____

☛ Bayes Information Criterion (BIC)

$$BIC = -2\loglik + \log(N)d$$

★ Penalizes complexity more heavily than $AIC = -\frac{2}{N}\loglik + 2\frac{d}{N}$

★ Asymptotically optimal: picks correct model (if it lies in the family) as $N \leftarrow \infty$

☛ Minimum Description Length: Formally the same as BIC, motivated by Information theory

$$Descriptionlen = \arg \min \text{len}(\text{encoded message}) + \text{len}(\text{encoding parameters})$$

Estimates of Model Complexity: # of Parameters _____

Estimates of Model Complexity: # of Parameters _____

☞ $Y = \beta_0 + \beta_1 X$

Estimates of Model Complexity: # of Parameters _____

④ $Y = \beta_0 + \beta_1 X$

④ $Y = I(\sin(\alpha_1 x + \alpha_0)),$

Estimates of Model Complexity: Vapnik Chernovenkis Dimension —

The VC dimension of the class $\{f(x, \alpha)\}$ is defined to be the largest number of points (in some configuration) that can be shattered by members of $\{f(x, \alpha)\}$.

Shatter

A set of points is said to be shattered by a class of functions if, for any binary labeling, a member of the class can perfectly separate them

Shatter

A set of points is said to be shattered by a class of functions if, for any binary labeling, a member of the class can perfectly separate them



Shatter

A set of points is said to be shattered by a class of functions if, for any binary labeling, a member of the class can perfectly separate them



Max points shattered: 3

Shatter

A set of points is said to be shattered by a class of functions if, for any binary labeling, a member of the class can perfectly separate them



Max points shattered: 3



Shatter

A set of points is said to be shattered by a class of functions if, for any binary labeling, a member of the class can perfectly separate them



Max points shattered: 3



Max points shattered: ∞

Example: Error of models picked by criteria relative to best model __

Figures: Hastie et al. 7.7

References

- ④ Hastie, Tibshirani and Friedman. The Elements of Statistical Learning. Springer-Verlag, 2001, pp. 193–213.<http://www-stat.stanford.edu/~tibs/ElemStatLearn/>