Model Selection

Topics in Data Mining
Fall 2015

Bruno Ribeiro
Goal

- Model Selection
- Model Assessment
In Training Phase There is Often a Better Model

Train Validation Test

Train

Test

Real-world Error

Model assessment

© 2015 Bruno Ribeiro
Measuring Error

- Input: $X$
- Output: $Y$
- Estimator: $f(\cdot; \hat{\theta})$
  - E.g.: $f(\cdot; \hat{\theta}) = \hat{\theta}_0 + \hat{\theta}_1 x + \hat{\theta}_2 x^2 + \ldots$

- Examples of loss function:
  $$L(Y, f(X; \hat{\theta})) = \begin{cases} (Y - f(X; \hat{\theta}))^2 & \text{squared error} \\ |Y - f(X; \hat{\theta})| & \text{absolute error} \end{cases}$$
Measuring Errors: Loss Functions

- Typical classification loss functions
  - 0-1 Loss:
    \[ L(Y, f_c(X; \hat{\theta})) = 1\{Y \neq f_c(X; \hat{\theta})\} \]
  - Log-likelihood:
    \[ L(Y, f_c(X; \hat{\theta})) = -2 \log p_c(X; \hat{\theta}) \]
The Problem of Overfitting

FIGURE 7.1. Behavior of test sample and training sample error as the model complexity is varied. The light blue curves show the training error, while the light red curves show the conditional test error for 100 training sets of size 50 each, as the model complexity is increased. The solid curves show the expected test error and the expected training error.

Test error, also referred to as generalization error, is the prediction error over an independent test sample:

\[ \text{Err}_T = E \left[ L(Y, \hat{f}(X)) \right] \bigg| T \]  

(7.2)

where both \( X \) and \( Y \) are drawn randomly from their joint distribution (population). Here the training set \( T \) is fixed, and test error refers to the error for this specific training set. A related quantity is the expected prediction error (or expected test error):

\[ \text{Err} = E \left[ L(Y, \hat{f}(X)) \right] = E[\text{Err}_T] \]  

(7.3)

Note that this expectation averages over everything that is random, including the randomness in the training set that produced \( \hat{f} \).

Estimation of \( \text{Err}_T \) will be our goal, although we will see that \( \text{Err} \) is more amenable to statistical analysis, and most methods effectively estimate the expected error. It does not seem possible to estimate conditional training error.
Linear Regression

\[ y = ax + b \]

\( n \) observations
Cubic Regression

\[ y = ax^3 + bx^2 + cx + d \]

Better

n observations
Degree $n-1$ Polynomial

$$y = ax^{n-1} + bx^{n-2} + \ldots$$

Even better?

$n$ observations
Model Selection and Assessment

- **Model Selection**: Estimating performances of different models to select the best one

- **Model Assessment**: Having chosen a model, estimating the prediction error on new data
Model Selection and Assessment

- In data-rich scenarios split the data:

- In data-poor scenarios: approximate validation step
  - analytically
    - AIC, BIC, MDL
  - via sample re-use
    - cross-validation
    - Leave-one-out
    - K-fold
    - bootstrap
Model vs Hypotheses

- Simple hypothesis: single probability distributions (or functions)
  - Model with specific parameters
- Composite hypothesis: family of probability distributions (or functions)
  - A Model
Why Look for More Restricted Models?

Bias x Variance

![Diagram showing the behavior of bias and variance. The model space is the set of all possible predictions from the model, with the "closest fit" labeled with a black dot. The model bias from the truth is shown, along with the variance, indicated by the large yellow circle centered at the black dot labeled "closest fit in population." A shrunken or regularized fit is also shown, having additional estimation bias, but smaller prediction error due to its decreased variance.]

© 2015 Bruno Ribeiro

The Elements of Statistical Learning
Jerome H. Friedman, Robert Tibshirani, and Trevor Hastie
Decomposing Test Error

Model: \( Y = f(X; \theta) + \epsilon; \quad E[\epsilon] = 0 \quad \text{Var}(\epsilon) = \sigma^2_\epsilon \)

For squared error loss & additive noise:

\[
\text{Err}(x_0) = E[(Y - f(x_0; \theta))^2] = x_0
\]

\[
= \sigma^2_\epsilon + \left( Ef(x_0; \hat{\theta}) - f(x_0) \right)^2 + E \left[ f(x_0; \hat{\theta}) - Ef(x_0; \hat{\theta}) \right]^2
\]

\[
= \sigma^2_\epsilon + \text{Bias}^2(f(x_0; \hat{\theta})) + \text{Var}(f(x_0; \hat{\theta}))
\]

Irreducible error of target Y

Expected squared deviation of estimate around its own mean

© 2015 Bruno Ribeiro
Model Bias

For linear models (e.g., Ridge), bias can be further decomposed:

\[ \beta^* \] is the best fitting linear approximation.

For standard (unconstrained) linear regression, estimation bias = 0

\[
\text{Bias}^2(f(x_0; \hat{\theta})) = E_{x_0}[f(x_0; \theta) - E\hat{\beta}^T x_0]^2
\]

\[
= E_{x_0}[f(x_0; \theta) - \beta^*_T x_0]^2 + E_{x_0}[\beta^*_T x_0 - E\hat{\beta}^T x_0]^2
\]

\( \beta^* \) is the best linear approximation.

\( \beta^* = \arg \min_{\beta} \text{Bias}^2(f(x_0; \theta)) \)

Average Model Bias

Average Estimation Bias

© 2015 Bruno Ribeiro
Optimism of the Training Error Rate

- Typically: training error rate < true error
  (same data is being used to fit the method and assess its error)

\[
Err_{\text{train}} = \frac{1}{N} \sum_{i=1}^{N} L(y_0(i), f(x_0(i); \hat{\theta})) < Err(X) = E[L(Y, f(X; \hat{\theta}))]
\]

overly optimistic

© 2015 Bruno Ribeiro
Estimating Test Error

- Can we estimate the discrepancy between $\text{Err}(x_0)$ and $\text{Err}(X)$?

- Suppose we measured new values of $y$ at same input $x$

$\text{Err}_{\text{in}}$ --- In-sample error:

$$E_Y[\text{Err}_{\text{in}}] = \frac{1}{N} \sum_{i=1}^{N} E_Y E_{Y_{\text{new}}} L(Y_{\text{new}}(i), f(x_0(i); \hat{\theta}))$$

$$= E_Y[\text{Err}_{\text{train}}] + \frac{2}{N} \sum_{i=1}^{N} \text{Cov}(\hat{y}_i, y_i)$$

Expectation over $N$ new responses at each $x_i$

Adjustment for optimism of training error
Measuring Optimism

If model: $Y = f(X; \theta) + \epsilon; \quad E[\epsilon] = 0 \quad \text{Var}(\epsilon) = \sigma^2_\epsilon$
with $|\theta| = d$

Then,

$$E_Y[Err_{in}] = E_Y[Err_{train}] + \frac{2}{N} d \sigma^2_\epsilon$$

- Optimism grows linearly with model dimension
- Optimism decreases as training sample size increases
Ways to Estimate Prediction Error

- In-sample error estimates:
  - AIC
  - BIC
  - MDL

- Extra-sample error estimates:
  - Cross-Validation
    - Leave-one-out
    - K-fold
  - Bootstrap
Estimates of In-Sample Prediction Error

- General form of the in-sample estimate:
  \[ Err_{in} = Err_{train} + \text{optimism error} \]

- For linear fit:
  \[ C_p = Err_{train} + \frac{2d}{N} \hat{\sigma}^2 \]
  Known as the Marllow's \( C_p \) statistic
AIC & BIC

Akaike Information Criterion (AIC)

\[ AIC = -\frac{2}{N} \cdot \log \text{lik} + 2 \cdot \frac{d}{N} \]

Bayesian Information Criterion (BIC)

\[ BIC = -2 \log \text{lik} + (\log N)d \]
**AIC & BIC**

\[ AIC = \log \text{lik}(Data \mid MLE \text{ params}) - (\# \text{ of parameters}) \]

AIC does not assume model is correct  
Better for small \# samples  
Does not converge to correct model

\[ BIC = \log \text{lik}(Data \mid MLE \text{ params}) - \frac{\log N}{2} (\# \text{ of parameters}) \]

BIC assumes model is correct  
Converges to correct parametrization of model

© 2015 Bruno Ribeiro
MDL (Minimum Description Length)

- Find hypothesis (model) that minimizes
  - $H(M) + H(D|M)$
  - $H(M)$ – length in bits of description of model
  - $H(D|M)$ – length in bits of description of data encoded by model

- If model probabilistic:

$$ length = -\log \Pr(y \mid \theta, M, X) - \log \Pr(\theta \mid M) $$

- **MDL principle**: choose the model with the minimum description length

- **Equivalent to maximizing the posterior**: $\Pr(y \mid \theta, M, X) \cdot \Pr(\theta \mid M)$
Estimation of Extra-Sample Err

- Cross Validation
- Bootstrap
Cross-Validation

For the \( k \)th part (third above), we fit the model to the other \( K-1 \) parts of the data, and calculate the prediction error of the fitted model when predicting the \( k \)th part of the data. We do this for \( k=1,2,\ldots,K \) and combine the \( K \) estimates of prediction error.

\[
CV(\hat{f}) = \frac{1}{N} \sum_{i=1}^{N} L(y(i), f(x_0^{(-K)}(i); \hat{\theta}))
\]

Typical choices of \( K \) are 5 or 10 (see below). The case \( K=N \) is known as leave-one-out cross-validation. In this case \( \kappa(i) = i \), and for the \( i \)th observation the fit is computed using all the data except the \( i \)th.

Given a set of models \( f(x, \alpha) \) indexed by a tuning parameter \( \alpha \), denote by \( \hat{f}^{(-k)}(x, \alpha) \) the \( \alpha \)th model fit with the \( k \)th part of the data removed. Then for this set of models we define

\[
CV(\hat{f}, \alpha) = \frac{1}{N} \sum_{i=1}^{N} L(y(i), f(x_0^{(-K)}(i); \hat{\theta}))
\]

The function \( CV(\hat{f}, \alpha) \) provides an estimate of the test error curve, and we find the tuning parameter \( \hat{\alpha} \) that minimizes it. Our final chosen model is \( f(x, \hat{\alpha}) \), which we then fit to all the data.
How many folds?

- Computation increases
- Variance decreases
- Bias decreases
- Leave-one-out

- $k$ fold
- $k$ increases
Cross-Validation: Choosing K

Popular choices for K: 5, 10, N=leave one out

FIGURE 7.8. Hypothetical learning curve for a classifier on a given task: a plot of $1 - \text{Err}$ versus the size of the training set $N$. With a dataset of 200 observations, fivefold cross-validation would use training sets of size 160, which would behave much like the full set. However, with a dataset of 50 observations fivefold cross-validation would use training sets of size 40, and this would result in a considerable overestimate of prediction error. On the other hand, with $K = 5$ say, cross-validation has lower variance. But bias could be a problem, depending on how the performance of the learning method varies with the size of the training set. Figure 7.8 shows a hypothetical “learning curve” for a classifier on a given task, a plot of $1 - \text{Err}$ versus the size of the training set $N$. The performance of the classifier improves as the training set size increases to 100 observations; increasing the number further to 200 brings only a small benefit. If our training set had 200 observations, fivefold cross-validation would estimate the performance of our classifier over training sets of size 160, which from Figure 7.8 is virtually the same as the performance for training set size 200. Thus cross-validation would not suffer from much bias. However if the training set had 50 observations, fivefold cross-validation would estimate the performance of our classifier over training sets of size 40, and from the figure that would be an underestimate of $1 - \text{Err}$. Hence as an estimate of $\text{Err}$, cross-validation would be biased upward. To summarize, if the learning curve has a considerable slope at the given training set size, five- or tenfold cross-validation will overestimate the true prediction error. Whether this bias is a drawback in practice depends on the objective. On the other hand, leave-one-out cross-validation has low bias but can have high variance. Overall, five- or tenfold cross-validation are recommended as a good compromise: see Breiman and Spector (1992) and Kohavi (1995). Figure 7.9 shows the prediction error and tenfold cross-validation curve estimated from a single training set, from the scenario in the bottom right panel of Figure 7.3. This is a two-class classification problem, using a lin-
Wrong Way to Do Cross-validation

- Use cross-validation to estimate anything in the model or for feature selection
- Use cross-validation to hunt for new models

The human regression machine:
Bootstrap: Main Concept

**Step 1:** Draw samples with replacement

**Step 2:** Calculate the statistic

**FIGURE 7.12.** Schematic of the bootstrap process. We wish to assess the statistical accuracy of a quantity $S(Z)$ computed from our dataset.

$Z = (z_1, z_2, \ldots, z_N)$

$Z^*_1$, $Z^*_2$, $Z^*_B$

$S(Z^*_1)$, $S(Z^*_2)$, $S(Z^*_B)$

Bootstrap replications

Bootstrap samples

Training sample

The Elements of Statistical Learning
Jerome H. Friedman, Robert Tibshirani, and Trevor Hastie
References

- The Elements of Statistical Learning: Data Mining, Inference, and Prediction
  By Trevor Hastie, Robert Tibshirani, Jerome Friedman