Model Diagnostics

MAST90083 Computational Statistics and Data Mining

Karim Seghouane

School of Mathematics & Statistics
The University of Melbourne

Outline

General purpose of model diagnostics

- Supervised learning models are used to investigate/discover the relationship between a response/outcome/dependent variable y and a set of predictor/explanatory/independent/covariate variables $\mathbf{x}$, based on observations $\mathbf{D} = \{(y_i, \mathbf{x}_i), i = 1, \cdots, N\}$.

- Given a training data set there is generally more than one possible learning method or model

General purpose of model diagnostics

- There is therefore a need for a measure of the quality of these learning methods or models
- The "generalization performance" of a leaning method relates to its prediction capability on independent test data
- It gives a measure of the quality of the selected model
- It helps assess how well the model fits and if necessary, modify the model to improve the fit
- It guides the choice of a learning method or model
- Assessment of this performance is important and used in practice

General purpose of model diagnostics

- Assume a quantitative response y and a vector of predictors $\mathbf{x}$
- Given a training sample $\mathbf{D} = \{(y_i, \mathbf{x}_i), i = 1, \cdots, N\}$ we can estimate a prediction model $\hat{f}(\mathbf{x})$

- The cost for measuring the error or deviation between y and $\hat{f}(\mathbf{x})$ is

$$L\left(y, \hat{f}(\mathbf{x})\right) = \begin{cases} \left[y - \hat{f}(\mathbf{x})\right]^2 & \text{squared error} \\ \left|y - \hat{f}(\mathbf{x})\right| & \text{absolute error} \end{cases}$$

Training error vs. Generalization error

- The training error is the empirical loss or the average loss over the training sample

$$\overline{err} = \frac{1}{N}\sum_{i=1}^{N} L\left(y_i, \hat{f}\left(\mathbf{x}_i\right)\right)$$

- The generalization error is the expected prediction error over an independent test sample (test error)

$$Err = E\left[L\left(y, \hat{f}\left(\mathbf{x}\right)\right)\right]$$

- Interest: Test error of our estimated model $\hat{f}$

Training error vs. Generalization error

- The training error is not a good estimate of the test error
- More complex model $\rightarrow$ adapt to more complex structures
- Training error consistently decreases with the model complexity $\rightarrow$ dropping to zero for high enough complex model
- However, a model with zero training error is overfit to the training data $\rightarrow$ generalize poorly

Training error vs. Generalization error

- For qualitative or categorical response $G \in \{1, \ldots, K\}$ we model

$$p_k(\mathbf{x}) = Pr\left(G = k/\mathbf{x}\right) \text{ and } \hat{G}(\mathbf{x}) = \arg\max_k \hat{p}_k\left(\mathbf{x}\right)$$

- The loss functions are

$$L\left(G, \hat{G}(\mathbf{x})\right) = I\left(G \neq \hat{G}(\mathbf{x})\right), \quad 0-1 \text{ loss}$$

$$L\left(G, \hat{p}(\mathbf{x})\right) = -2\sum_{k=1}^{K} I\left(G = k\right) \log \hat{p}_k\left(\mathbf{x}\right) = -2 \log \hat{p}_G\left(\mathbf{x}\right)$$

Training error vs. Generalization error

- We are interested in the expected missclassification rate

$$Err = E\left[L\left(G, \hat{G}(\mathbf{x})\right)\right] \quad \text{or} \quad Err = E\left[L\left(G, \hat{p}(\mathbf{x})\right)\right]$$

- But in practice we have access to

$$\overline{Err} = \frac{-2}{N}\sum_{i=1}^{N} \log \hat{p}_{g_i}(x_i)$$

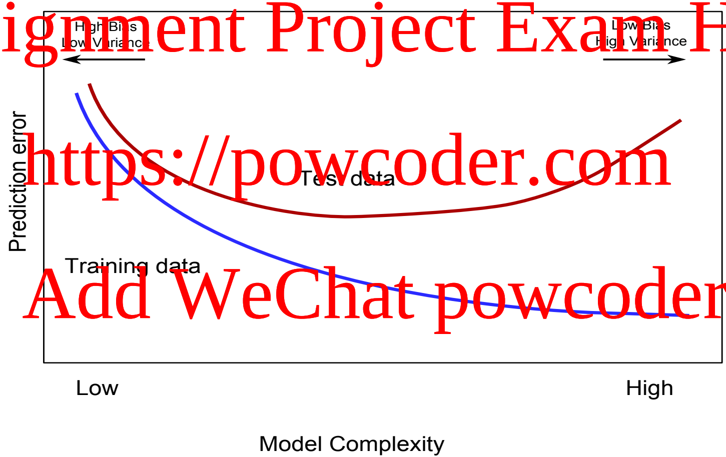
- We are interested in estimating the test error
- and find the model with the appropriate complexity

Training error vs. Generalization error

- The problem of estimating the test error in categorical response settings is similar to the quantitative case response setting on which we will focus.

- If there was a parameter α which controlled the complexity of the model, the aim is to find α that produces the minimum test error.

Training error vs. Generalization error



- The test error varies with the model complexity

Model Diagnostic with Data

There are two separate objects in supervised learning

- **Model Selection** estimating the performance of different models in order to choose the approximate best one; and
- **Model Assessment** having chosen or selected a model, estimating its prediction error (generalization error or performance) on new data.

Before we look at these however, we shall define some terms and discuss the bias-variance tradeoff and its relation to model complexity.

Model Diagnostic with Data

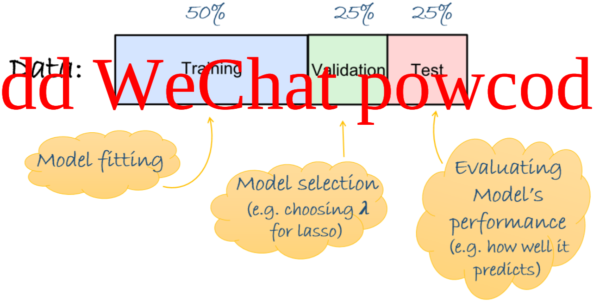
In a data-rich situation, the best approach for both problems is to randomly divide the dataset into three sets

- A **training** dataset is the set of data used to fit a model.
- A **validation** dataset is the one on which we check the performance of the model fitted from the training dataset. We use this to guide our model selection.
- A **test** dataset is the one on which we assess the prediction accuracy of the model found from a model selection procedure.

If the test set is also used to choose the model $\rightarrow$ the final model will underestimate the true test error

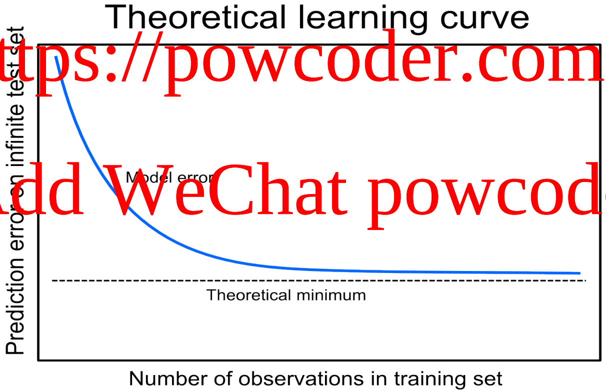
Model Diagnostic with Data

- There is no general rule on how to choose the number of samples of each data set. This could for example depends on signal-to-noise ratio in the data or the model complexity
- A typical split might be 50% for training and 25% each for validation and testing.



Accuracy versus sample size

By splitting up the data this way we are left with a significantly smaller number of observations with which to fit our model. This is sometimes a problem, although not always.



Accuracy versus sample size

In a number of cases, there is insufficient data to split it into three parts. The methods here

- Approximate the validation step analytically using model selection criteria or by
- Efficient sample re-use
- Provide an estimate of the test error of the final chosen model

Accuracy versus sample size

- The methods discussed next are designed for situations where there is insufficient data

- These methods approximate or include the validation step
  - analytically: C_p, AIC, BIC
  - by efficient sample re-use: cross-validation and bootstrap

Bias-variance decomposition

Assume $y = f(\mathbf{x}) + \varepsilon$ with $E(\varepsilon) = 0$, $\mathbf{Var}(\varepsilon) = \sigma^2$ and $\hat{f}(\mathbf{x})$ is a regression fit

- The expected prediction error of a fit $\hat{f}(\mathbf{x})$ at $\mathbf{x} = \mathbf{x}_0$

$$\mathrm{Err}(\mathbf{x}_0) = E\left\{\left(y - \hat{f}(\mathbf{x}_0)\right)^2\right\}$$

$$= \sigma^2 + \left[E\hat{f}(\mathbf{x}_0) - f(\mathbf{x}_0)\right]^2 + E\left[\hat{f}(\mathbf{x}_0) - E\hat{f}(\mathbf{x}_0)\right]^2$$

$$= \sigma^2 + \mathbf{Bias}^2\left(\hat{f}(\mathbf{x}_0)\right) + \mathbf{Var}\left(\hat{f}(\mathbf{x}_0)\right)$$

Note 1

Bias-variance decomposition

- The first term is the variance around $f(\mathbf{x}_0)$ and can not be reduced

- The second term is the squared bias, the average $\hat{f}(\mathbf{x}_0)$ differs from $f(\mathbf{x}_0)$
- The last term is the variance, the average deviation of $\hat{f}(\mathbf{x}_0)$ from its mean

Bias-variance decomposition

- The more complex we make $\hat{f}$, the lower the (squared) bias but the higher the variance
- The optimal model is the one that gives the best compromise between the second and third term

Bias-variance decomposition

- Suppose we have the choice of a range of models of differing complexity.
  - For example, the number of polynomial terms in a one dimensional linear regression $(x, x^2, x^3, \ldots)$.

- Suppose further that we have both a **training** dataset to fit our model and a **test** dataset on which to assess prediction accuracy.

- The more complex models will always fit the training data better, but will not necessarily improve test performance!
- The extra complexity allows the bias of the estimate to be reduced, but at a cost of extra variance associated with estimating parameters.

An Example of Bias-variance decomposition

For a linear model $\hat{f}(\mathbf{x}) = \hat{\beta}^\top \mathbf{x}$, $\beta \in \mathbb{R}^p$ estimated by least square

$$\mathrm{Err}(\mathbf{x}_i) = E\left\{\left(y - \hat{f}(\mathbf{x}_i)\right)^2\right\}$$

$$= \sigma^2 + \left[E\hat{f}(\mathbf{x}_i) - f(\mathbf{x}_i)\right]^2 + \|h(\mathbf{x}_i)\|^2\sigma^2$$

- While the variance changes with $\mathbf{i}$ its average (over the sample values $\mathbf{x}_i$) doesn't

Note 2

An Example of Bias-variance decomposition

$$\frac{1}{N}\sum_{i=1}^{N}\mathrm{Err}(\mathbf{x}_i) = \sigma^2 + \frac{1}{N}\sum_{i=1}^{N}\left[E\hat{f}(\mathbf{x}_i) - f(\mathbf{x}_i)\right]^2 + \frac{p}{N}\sigma^2$$

- For linear models fit by LS, the bias is zero
- For regularized fit, the bias is positive with aim to reduce the variance

Note 2

An illustration of bias-variance tradeoff

Suppose we have $n = 50$ observations with $x_i \stackrel{d}{=} \text{Uniform}(0, 2)$, and

$$y_i = \cos(2x_i) + \varepsilon_i \, ,$$

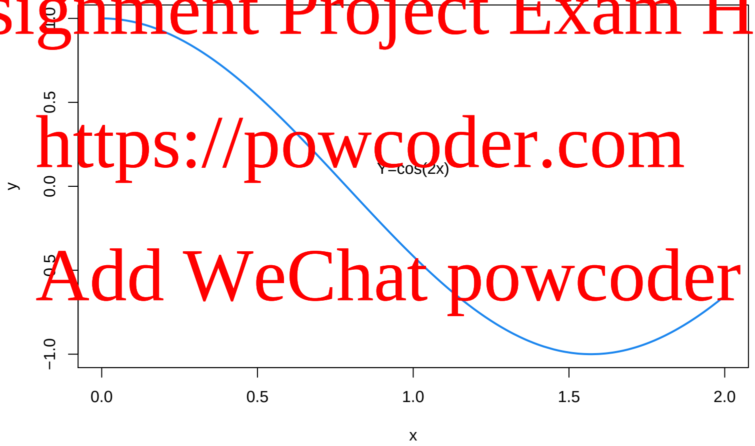
with $\varepsilon_i \stackrel{d}{=} N(0, 0.3^2)$.

We want to use a polynomial function of x_i to fit y_i, but we are not sure how many polynomial terms to use.

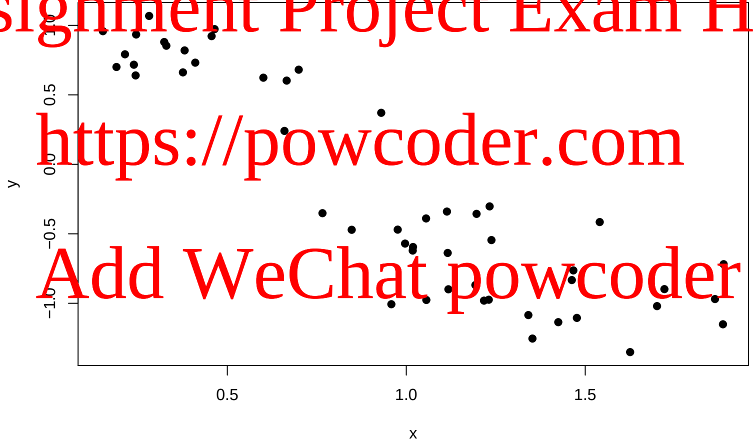
To test this, we fit on the data for linear, quadratic, cubic, quartic and quintic fits.

We then check how each fitted model performs on a separate **test** dataset of $n = 1000$. We repeat 100 times and average the results.

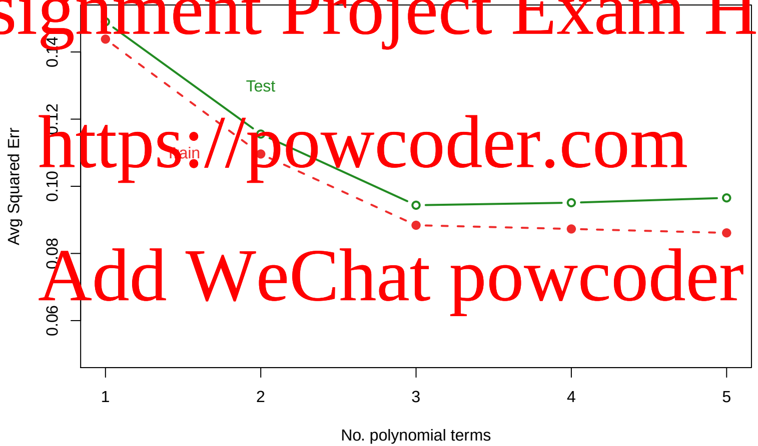
Target function



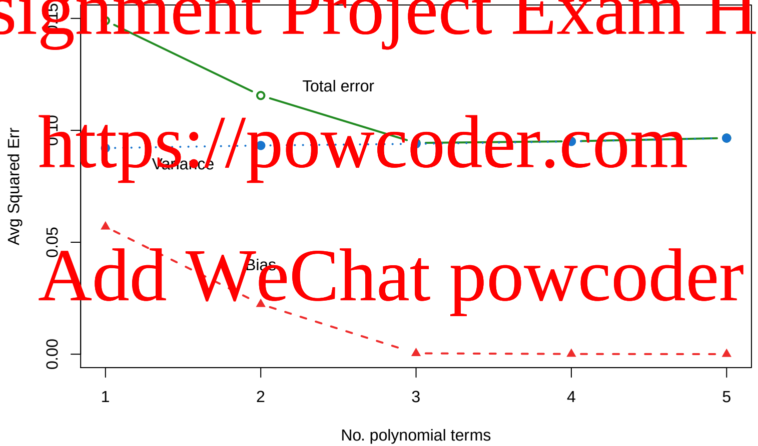
Example data (one of the 100 repeations)



Training and test error for increasingly complex models



Bias-variance decomposition for test data



Resume: model fitting, selection and assessment

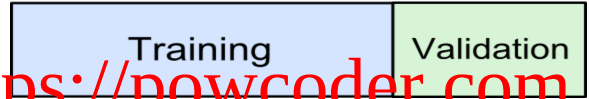
- A morale coming out of this example is that **a model that fits well on the training data is not indicative of true model performance.**
- Thus the average loss

$$\bar{err} = \frac{1}{N}\sum_{i=1}^{N} L\{y_i, \hat{f}(\mathbf{x}_i)\}$$

computed from the training data can be misleading.
- Rather, the average loss of the model should be computed on a separate "test" dataset.
- In reality, this means we should partition the data, fitting on one portion and testing performance on the other.

Resume: model fitting, selection and assessment

- On the other hand, suppose the data is split into a training and a validation set.



- In model selection we use the **training data** to fit each candidate model and choose as the selected model the one having the best performance on the validation set.

- The average loss on the validation set will be smallest for the selected model.
- But to assess the prediction accuracy of the selected model, we still need to calculate its average loss based on a separate "test" dataset.

Optimism

- The training error rate

$$\overline{err} = \frac{1}{N}\sum_{i=1}^{N} L\left(y_i, \hat{f}(x_i)\right)$$

- will not correctly reflect the true error

$$Err = E\left[L\left(y, \hat{f}(\mathbf{x})\right)\right]$$

- because the same data is used to fit the model and assess its error

Optimism

- The model obtained from $\mathbf{D} = \{(y_i, \mathbf{x}_i), i = 1, \cdots, N\}$ adapts to the training data

- The training error $\bar{err}$ is an optimistic estimate of the generalization error Err
- Err is a form of extra-sample error, since the test feature vectors don't need to coincide with the training vectors

Optimism

- The nature of the optimism can be seen when we consider

$$Err = \frac{1}{N}\sum_{i=1}^{N} E_{New}\left[L\left(y_i, \hat{f}\left(\mathbf{x}_i\right)\right)\right]$$

- where E_{New} indicates that we observe multiple new responses at each of the training points $\mathbf{x}_i$, $i = 1, \ldots, N$

- It better reflects the true error and therefore it is a better performance measure of a model

Optimism

- The optimism is defined as

$$op = E[Err - \bar{err}]$$

- and is positive since $\bar{err}$ is usually biased downward as an estimate of Err

- An obvious way to estimate the prediction error is to estimate the optimism and add it to the training error $\bar{err}$

Optimism

- A corrected estimate of Err is

$$\hat{Err} = \bar{err} + \hat{op}$$

- where $\hat{op}$ is an estimate of the optimism
- This corrected estimate provides a method to assess and select a model

Optimism

- For squared loss

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

- The amount by which $\bar{err}$ underestimates the true error depends on how strongly y_i affects its own prediction

- The harder we fit the data, the greater $\operatorname{cov}(\hat{y}_i, y_i)$ will be, thereby increasing the optimism

Note 3

Optimism

- The expected criterion is

$$Err = E_y\left(\overline{err}\right) + \frac{2}{N}\sum_{i=1}^{N} \operatorname{cov}\left(\hat{y}_i, y_i\right)$$

- which gives in the case of linear model fit with d input

$$Err = E_y\left(\overline{err}\right) + \frac{2d}{N}\sigma^2$$

Note 4

Relation with existing criteria

C_p statistics

$$C_p = \overline{err} + 2\frac{d}{N}\hat{\sigma}^2$$

- $\hat{\sigma}^2$ is the noise variance obtained from the mean squared error of a low bias model
- This criterion adjust the training error by a factor proportional to the number of parameter

Relation with existing criteria

- The Akaike information criterion

$$AIC = -\frac{2}{N}\,\mathrm{loglik} + \frac{2d}{N}$$

- and is equivalent to C_p for Gaussian models

$$AIC(\alpha) = \bar{err}(\alpha) + 2\frac{d(\alpha)}{N}\hat{\sigma}^2$$

- since $-2\mathrm{loglik}$ equals $\sum_i \left(y_i - f(x_i)\right)^2 / \sigma^2$ which is $N.\bar{err}/\sigma^2$

Relation with existing criteria

- The Bayesian information criterion

$$BIC = -\frac{2}{N}.\mathrm{loglik} + d\log(N)$$

- and for Gaussian models

$$BIC = \frac{N}{2}\left[\bar{err} + \frac{d}{N}.\log(N).\sigma^2\right]$$

- Therefore BIC is proportional to AIC and C_p with the factor 2 replaced by $\log(N)$
- BIC tends to penalize complex models more heavily, giving preference to simpler models in selection

The effective number of parameters

- The concept of "number of parameters" can be generalized where regularization is used

$$\hat{y} = Sy$$

- where for least squares

$$S = X\left(X^\top X\right)^{-1} X^\top$$

- and for ridge regression

$$S = X\left(X^\top X + \lambda I\right)^{-1} X^\top$$

The effective number of parameters

- The effective number of parameters is

$$d(S) = \mathrm{trace}(S)$$

- If S is an orthogonal projection matrix

$$\mathrm{trace}(S) = d$$

- the number of parameters

Motivations for Selection Criteria

Let M_0 be the model with density $f_{\boldsymbol{\beta}_0}$

- In the linear case for example

$$\mathbf{y} = X\boldsymbol{\beta}_0 + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}\left(0, \sigma_0^2 I\right)$$

- We have $f_{\boldsymbol{\beta}_0} = f(\mathbf{y}/\boldsymbol{\beta}_0) = f(\boldsymbol{\epsilon})$

Motivations for Selection Criteria

Given a class of candidate models $M = \{M_1, ..., M_K\}$, selection criteria aims to select a candidate model M_k as an approximation for M_0

- In the linear this becomes

$$\hat{\mathbf{y}} = X\hat{\beta}_k + \epsilon_k, \quad \epsilon_k \sim \mathcal{N}\left(0, \sigma_k^2 I\right)$$

- and we have $f_{\boldsymbol{\beta}_k} = f\left(\mathbf{y}/\hat{\beta}_k\right) = f\left(\epsilon_k\right)$

Motivations for Selection Criteria

C_p is a criterion derived using the L_2 norm as a basis for measuring the discrepancy

- The derivation of selection criteria requires methods or measures to quantify the separation between M_0 and M_k
- C_p is a criterion derived using the L_2 norm as a basis for measuring the discrepancy between M_0 and M_k

$$\Delta\left(M_0, M_k\right) = \left\|\mu_{M_0} - \mu_{M_k}\right\|^2 = L_2\left(M_k\right)$$

where μ_{M_0} and μ_{M_k} are the true and candidate model means

Motivations for Selection Criteria

Advantages

- L_2 depends only on the means of the models and not on the actual two densities
- This means that L_2 can be applied when errors are not normally distributed

Disadvantage

- L_2 is a matrix in certain multivariate models

Motivations for Selection Criteria

C_p provides an estimation of $E[J_k]$ where

$$J_k = \frac{1}{\sigma_0^2}\left(\hat{\beta}_k - \beta_0\right)^\top X^\top X\left(\hat{\beta}_k - \beta_0\right)$$

$$E\left[RSS_k/\sigma_0^2\right] = n - k + \frac{B_k}{\sigma_0^2}$$

$$E\left[\frac{RSS_k}{\sigma_0^2} - n + 2k\right] = k + \frac{B_k}{\sigma_0^2}$$

Motivations for Selection Criteria

Hence

$$C_p = \frac{RSS_k}{\sigma_0^2} - n + 2k$$

is unbiased for $E\left[J_k\right]$

In C_p, σ_0^2 is replaced by σ_K^2 obtained from the largest candidate model

Motivations for Selection Criteria

AIC is based on using the Kullback-Leibler divergence between the true and approximating probability density models as measure of discrepancy

$$
\begin{aligned}
E_0\left[\log \frac{f(\mathbf{y}/\beta_0)}{f(\mathbf{y}/\beta_k)}\right] &= \int f(\mathbf{y}/\beta_0) \log f(\mathbf{y}/\beta_0)\, d\mathbf{y} \\
&\quad - \int f(\mathbf{y}/\beta_0) \log f(\mathbf{y}/\beta_k)\, d\mathbf{y} \\
&= d(\beta_0, \beta_0) - d(\beta_k, \beta_0)
\end{aligned}
$$

Motivations for Selection Criteria

The discrepancy can then by measured using

$$d_n(\beta_k, \beta_0) = E_0\{-2\log f(\mathbf{y}_n/\beta_k)\}$$

Using the maximum likelihood $\hat{\beta}_k$

$$d_n\left(\hat{\beta}_k, \beta_0\right) = E_0\{-2\log f(\mathbf{y}_n/\beta_k)\}\,|_{\beta_k=\hat{\beta}_k}$$

Motivations for Selection Criteria

Akaike noted that $-2\log f\left(\mathbf{y}_n/\hat{\beta}_k\right)$ is baised and that the bias

$$E_0\left\{E_0\left\{-2\log f\left(\mathbf{y}_n/\beta_k\right)\right\}|_{\beta_k=\hat{\beta}_k}\right\}-E_0\left\{-2\log f\left(\mathbf{y}_n/\hat{\beta}_k\right)\right\}$$

can often be asymptotically estimated by twice the dimension of $\hat{\beta}_k$

Motivations for Selection Criteria

Therefore for

$$AIC = -2\log f\left(\mathbf{y}_n / \hat{\beta}_k\right) + 2k$$

we have

$$E_0\{AIC\} \simeq E_0\left\{d_n\left(\hat{\beta}_k, \beta_0\right)\right\}$$

Motivations for Selection Criteria

- The derivation of BIC is motivated using Bayesian arguments
- Let $f(M_k)$, $k \in \{1, ..., K\}$ denotes the discrete prior over the models $M_1, ..., M_K$
- Let $f(\beta_k | M_k)$ denotes a prior on β_k given the model M_k

Motivations for Selection Criteria

Applying Bayes rule gives

$$f(\mathbf{y}, \beta_k, M_k) = f(\mathbf{y}/\beta_k, M_k)\, f(\beta_k/M_k)\, f(M_k)$$

$$= f(\beta_k, M_k/\mathbf{y})\, f(\mathbf{y})$$

BIC aims to choose the model which is a posteriori most probable

Motivations for Selection Criteria

The posterior probability density for M_k

$$f\left(M_k/\mathbf{y}\right) = \frac{1}{f\left(\mathbf{y}\right)} f\left(M_k\right) \int f\left(\mathbf{y}/\boldsymbol{\beta}_k, M_k\right) f\left(\boldsymbol{\beta}_k/M_k\right) d\boldsymbol{\beta}_k$$

Considering minimizing

$$-2\log f\left(M_k/\mathbf{y}\right) = 2\log\left\{f\left(\mathbf{y}\right)\right\} - 2\log\left\{f\left(M_k\right)\right\}$$

$$-2\log\left\{\int f\left(\mathbf{y}/\boldsymbol{\beta}_k, M_k\right) f\left(\boldsymbol{\beta}_k/M_k\right) d\boldsymbol{\beta}_k\right\}$$

Motivations for Selection Criteria

- The first term is constant with respect to k and
- assuming uniform prior for $f(M_k)$ and $f(\beta_k/M_k)$

the BIC is obtained using a Taylor serie expension and a Laplace approximation of the resulting integral

$$-2\log f(M_k/\mathbf{y}) \approx -2\log f\left(\mathbf{y}/\hat{\beta}_k\right) + k\log n$$

Motivations for Selection Criteria

- BIC is an asymptotic approximation of $-2 \log f(M_k/\mathbf{y})$
- The model with minimum BIC is the model with the largest approximate posterior probability

We have discussed three type of criteria C_p, AIC and BIC, what is the difference ?

Motivations for Selection Criteria

AIC and C_p are asymptotically efficient

$$\lim_{n\to\infty} \frac{E_0\left[L_2\left(M_c\right)\right]}{E_0\left[L_2\left(M_k\right)\right]} = 1$$

M_c is the model that is the closest to the true model

BIC is consistent (asymptotically select with probability one, the model having the correct structure)

Motivations for Selection Criteria

In Bayesian applications, comparison between models are based on Bayes factors

Considering two models M_{k_1} and M_{k_2} the bayes factor B_{12} is the ratio of the posterior odds

$$\frac{f(M_{k_1}/\mathbf{y})}{f(M_{k_2}/\mathbf{y})}$$

If $B_{12} > 1$, M_{k_1} is favored by the data and if $B_{12} < 1$, then M_{k_2} is favored by the data

Model Evaluation

- A problem closely related to model selection is one of model evaluation

- Here, an investigator is less interested in the selection of a single model and more interested in assessing preference from the data toward each of the models in the candidate collection

Model Evaluation

As BIC approximates a transformation of a model's posterior probability, one can perform model evaluation by transforming BIC back to a posterior probability

$$f(M_k/\mathbf{y}) \approx \frac{\exp\left(-\frac{1}{2}BIC_k\right)}{\sum_{i=1}^{K}\exp\left(-\frac{1}{2}BIC_i\right)}$$

The set of posterior probabilities can be used as a model evaluation tool and assess the relative merits of the considered models

Model Averaging

This cam also be used in model averaging

- Consider inference on a parameter δ that is defined within each model in the collection of candidates models
- δ can be a prediction $f(\mathbf{x})$ at some fixed value $\mathbf{x}_0$
- Rather than taking a selected model as correct with probability one, model averaging allows a quantification of the uncertainty inherent to model selection

Model Averaging

- The posterior distribution on δ is found as a weighted average of the posterior distributions conditional on each model

$$f(\delta/\mathbf{y}) = \sum_{k=1}^{K} f(\delta/M_k, \mathbf{y}) f(M_k/\mathbf{y})$$

- with posterior mean

$$E(\delta/\mathbf{y}) = \sum_{k=1}^{K} E(\delta/M_k, \mathbf{y}) f(M_k/\mathbf{y})$$

Model Averaging

- This Bayesian prediction is a weighted average of the individual predictions with weights proportional to the posterior probability of each model
- The process of model averaging is seen to improve estimation and prediction which tend to be over-confident if one proceeds as if a selected model is correct with certainty

Cross-Validation

- The simplest and most widely used method for estimating the prediction error is cross-validation

- It estimates the generalization error

$$Err = E\left[L\left(\mathbf{y}, \hat{f}\left(\mathbf{x}\right)\right)\right]$$

- when $\hat{f}(\mathbf{x})$ is applied to an independent test sample from the joint distribution

Cross-Validation

- Suppose for now that we do not need the final model assessment on a test dataset, so only need to fit and validate a model

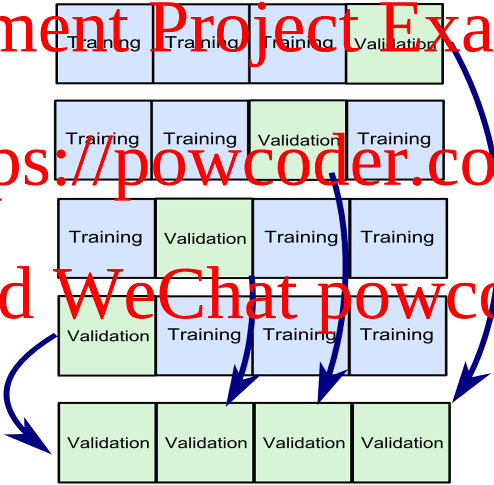
- K-fold cross validation uses part of the available data to fit the model and a different part to test it

Cross-Validation

- $K-1$ parts of the data are used to fit or learn the model and the k^{th} part is used to calculate the prediction error of the fitted model when predicting the k^{th} part of the data

- This is repeted for $k = 1, ..., K$ and the K estimates of the prediction error are combined (averaged)
- $\hat{f}^{-k}$ is widely used to denote the fitted model obtained with the k^{th} part of the data removed

Illustration of Cross-Validation



Cross-Validation

The case $K = N$ is known as leave-one-out cross-validation

$$CV = \frac{1}{N}\sum_{i=1}^{N} L\left(y_i, \hat{f}^{-k(i)}(\mathbf{x}_i)\right)$$

In this case $k(i) = i$, the fit is computed using all the data except the i^{th} pair $(y_i, \mathbf{x}_i)$.

In this case CV is approximately unbiased for the true prediction error with low bias

Cross-Validation

Given a set of models indexed by a tuning parameter α

$$CV(\alpha) = \frac{1}{N}\sum_{i=1}^{N} L\left(y_i, \hat{f}^{-k(i)}(\mathbf{x}_i, \alpha)\right)$$

- The curve $CV(\alpha)$ is used for tuning the parameter α
- Select $\hat{\alpha}$ that minimizes $CV(\alpha)$
- Use the model $f(\mathbf{x}, \hat{\alpha})$ is the final chosen model

Cross-validation

Leave-one-out vs. k-fold CV

- In practice $K = 5$ or $K = 10$ is usually sufficient.
- In situations where the sample size is not large, **leave-one-out CV** may be employed.
- k-fold preferable on leave-one-out CV
  - Save computational time.
  - Improves the accuracy due to bias-variance trade-off.
  - As $K \uparrow$ more obs. used to fit the model $\Rightarrow$ bias $\downarrow$.
  - But, the num. of obs. in the validation set $\downarrow$ $\Rightarrow \uparrow$ variance (less typical obs. / outliers have more influence).

Generalized Cross-validation

- GCV provides a convenient approximation to leave-one-out cross-validation for linear fitting under squared loss
- In linear fitting $\hat{\mathbf{y}} = S\mathbf{y}$ with least square $H = X\left(X^\top X\right)^{-1} X^\top$

- In the case of linear model

$$\frac{1}{N}\sum_{i=1}^{N}\left[y_i - \hat{f}^{-k(i)}(\mathbf{x}_i)\right]^2 = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{y_i - \hat{f}(\mathbf{x}_i)}{1 - S_{ii}}\right]^2$$

where

$$S_{ii} = \mathbf{x}_i^\top \left(X^\top X\right)^{-1} \mathbf{x}_i$$

Note 5

Generalized Cross-validation

The GCV approximation is

$$GCV(\hat{f}) = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{y_i - \hat{f}(\mathbf{x}_i)}{1 - \mathrm{trace}(S)/N}\right]^2$$

and takes the form

$$GCV(\hat{f}) = \bar{err} + \frac{2p}{N}\hat{\sigma}^2$$

where

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{f}(\mathbf{x}_i)\right)^2$$

Note 6

For more readings

- Summaries on LMS
- Chapters 7 & 8 from 'The elements of statistical learning' book.
- Chapters 6 from 'An introduction to statistical learning' book.