Best Subset Selection

If $y_1, \ldots, y_N \in \mathbb{R}$ and $x_1, \ldots, x_N \in \mathbb{R}^p$ and

$$\mathrm{RSS}(q) = \min_{i_1,\ldots,i_q} \min_{\beta_1,\ldots,\beta_q} \sum_{j=1}^{q} (y_j - (\beta_1 x_{j,i_1} + \ldots + \beta_q x_{j,i_q}))^2$$

is the least residual sum of squares for using any q dimensional submodel, then what q should I choose?

RSS is a relevant measure for comparison within all q-dimensional models, but decreases monotely with q.

Model Selection

The most important question that we have not dealt with yet is:

How do we select an appropriate model among several models of different/incommensurable complexity?

This is model selection. What is this besides an estimation problem?

Model Selection

We can always put all considered models into a single parameter set Θ and consider predictors f_θ for $\theta \in \Theta$. With e.g.

$$\hat{\theta} = \operatorname*{argmin}_{\theta \in \Theta} \frac{1}{N} \sum_{i=1}^{N} L(y_i, f_\theta(x_i))$$

the empirical loss minimizer $f_{\hat{\theta}}$ could be a bad predictor. We divide the parameter space $\Theta = \Theta_0 \cup \Theta_1$ into disjoint sets.

- Asymmetric test theoretic approach: $\hat{\theta}_1 \in \Theta_1$ is preferred over $\hat{\theta}_0 \in \Theta_0$ only if it is improbable as measured by a p-value that $\hat{\theta}_0$ is as good as $\hat{\theta}_1$.
- Symmetric prediction based approach: Consider

  $$\mathrm{EPE}(f_{\hat{\theta}_i}) = E(L(Y, f_{\hat{\theta}_i}(X))|\mathbf{X}, \mathbf{Y})$$

  and choose the one with the least expected prediction error.

Optimization versus Simplicity

One can say that we have two opposing philosophical directions of how to draw inference from empirical data:

- Trust all aspects of the information in the data on the relation between y and x as implemented in parameter estimators based on pure optimization.
- Keep it simple. Don't choose a complicated model over a simpler model if the simpler model suffice (Occam's razor).

Model selection/test theory work by the second principle to compensate for the fact that the optimization procedure may overfit to the given dataset.

Model or Predictor Assessment

Another problem of some importance is

Having fitted a final predictor $\hat{f}$, how will it actually perform?

We have the training error

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

which generally underestimates $\text{EPE}(\hat{f})$. The generalization or test error is

$$\text{Err} = E(\text{EPE}(\hat{f}))$$

is the expected EPE.

$\text{EPE}(\hat{f})$ can only really be estimated if we have an independent test dataset.

Figure 7.1 – The Bias-Variance Tradeoff

Realizations of training error $\overline{\mathrm{err}}$ and expected prediction error $\mathrm{EPE}(\hat{f})$ estimated on an independent test dataset as functions of model complexity. Also the estimates of the expectation of $\overline{\mathrm{err}}$ and $\mathrm{EPE}(\hat{f})$ are shown.

The Train-Validate-Test Idea

In a data rich situation we split the data before doing anything else into three subsets.

- On the training data we estimate all parameters besides tuning parameters (model complexity parameters).
- On the validation data we estimate prediction error for the estimated predictors and optimize over tuning parameters and models.
- On the test data we estimate the expected prediction error for the chosen predictor – no model selection here, please.

Problem: We are almost never in a data rich situation.

Can we justify to throw away data that can be used for estimation and thus reduction of variance to estimate parameters of secondary importance?

Figure 7.2 – Space of Models

Figure 7.3 – Quadratic Loss vs. 0-1 Loss

Setup

In the following discussion $(X_1, Y_1), \ldots, (X_N, Y_N)$ denote N i.i.d. random variables, with X_i a p-dimensional vector.

A concrete realization is denoted $(x_1, y_1), \ldots, (x_N, y_N)$ and we use boldface, e.g. $\mathbf{Y} = (Y_1, \ldots, Y_N)^T$ and $\mathbf{y} = (y_1, \ldots, y_N)^T$ to denote vectors.

We can not distinguish in notation between $\mathbf{X}$ – the matrix of random variables $X_1, \ldots, X_N$ – and $\mathbf{X}$ – the matrix of a concrete realization $x_1, \ldots, x_N$.

Mallows' C_p

With $\hat{\mathbf{f}} = P\mathbf{y}$ where P is a projection onto a d-dimensional subspace define

$$\overline{\text{err}} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{\mathbf{f}}_i)^2 = \frac{1}{N}||\mathbf{y} - \hat{\mathbf{f}}||^2.$$

By a standard decomposition

$$\frac{1}{N}E(||\mathbf{Y}^{\text{new}} - \hat{\mathbf{f}}||^2|\mathbf{X}) = \frac{1}{N}E(||\mathbf{Y} - \hat{\mathbf{f}}||^2|\mathbf{X}) + \frac{2d}{N}\sigma^2$$

The in-sample error

$$\text{Err}_{\text{in}} = \frac{1}{N}E(||\mathbf{Y}^{\text{new}} - \hat{\mathbf{f}}||^2|\mathbf{X})$$

can thus be estimated by

$$C_P = \widehat{\text{Err}}_{\text{in}} = \overline{\text{err}} + \frac{2d}{N}\hat{\sigma}^2.$$

Mallows' C_p

$$C_P = \hat{\text{Err}}_{\text{in}} = \bar{\text{err}} + \frac{2d}{N}\hat{\sigma}^2.$$

is an equivalent of Mallows' C_p statistics – with $\hat{\sigma}^2$ estimated from a "low-bias" model with p degrees of freedom;

$$\hat{\sigma}^2 = \frac{1}{N-p}||\mathbf{y} - Q\mathbf{y}||^2$$

where Q is a projection on a p-dimensional space.

If $\mathbf{S}_\lambda$ is a smoother and $\hat{\mathbf{f}} = \mathbf{S}_\lambda \mathbf{y}$ one can generalize C_p as

$$\hat{\text{Err}}_{\text{in}} = \bar{\text{err}} + \frac{2\text{trace}(\mathbf{S}_\lambda)}{N}\hat{\sigma}^2$$

with $\hat{\sigma}^2$ estimated from a "low-bias", or small λ, model, e.g.

$$\hat{\sigma}^2 = \frac{1}{N - \text{trace}(2\mathbf{S}_\lambda - \mathbf{S}_\lambda^2)}||\mathbf{y} - \mathbf{S}_\lambda \mathbf{y}||^2.$$

Using C_p

The classical use of C_p is when $\mathbf{X}$ is $N \times p$ of rank p and $Q = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$.

For any choice of d columns we compute C_p and select the model with the smallest value of C_P.

This is equivalent to best subset selection for each d followed by choosing d that minimizes

$$NC_p = \mathrm{RSS}(d) + \frac{2d}{N-p}\mathrm{RSS}(p)$$

As a function of d the normal defintion of C_p,

$$\tilde{C}_p = \frac{NC_p(N-p)}{\mathrm{RSS}(p)} - N = \frac{(N-p)\mathrm{RSS}(d)}{\mathrm{RSS}(p)} + 2d - N,$$

is a monotonely increasing function of C_p.

Generalization Error

Instead of the in-sample error we can consider the generalization or test error

$$\text{Err} = E(L(Y, \hat{f}(X))) = E(E(L(Y, \hat{f}(X))|\mathbf{X}, \mathbf{Y})) = E(\text{EPE}(\hat{f}))$$

Here (X, Y) is independent of $(X_1, Y_1), \ldots, (X_N, Y_N)$ that enter through $\hat{f}$.

Err is the expectation over the dataset of the expected prediction error for the estimated predictor $\hat{f}$.

A small value of Err tells us that the estimation methodology is good and will on average result in estimators with a small EPE. It does not guarantee that a concrete realization $\hat{f}$ has a small EPE!

In-sample Error and Generalization Error

Note the decomposition

$$\begin{aligned}
\mathrm{Err} &= E(L(Y, \hat{f}(X))) \\
&= E(E(L(Y, \hat{f}(X))|\mathbf{X}, X)) \\
&\simeq E\left(\frac{1}{N}\sum_{i=1}^{N} E(L(Y, \hat{f}(x_i))|\mathbf{X}, X = x_i)\right) \\
&= E\left(\frac{1}{N}\sum_{i=1}^{N} E(L(Y_i^{\mathrm{new}}, \hat{f}(x_i))|\mathbf{X})\right) \\
&= E\left(\mathrm{Err}_{in}\right)
\end{aligned}$$

If the p-dimensional model is unbiased $E(\hat{\sigma}^2|\mathbf{X}) = \sigma^2$ and

$$E(C_p) = E(\hat{\mathrm{Err}}_{\mathrm{in}}) = E(\bar{\mathrm{err}}) + \frac{2d}{N}\sigma^2 = E(\mathrm{Err}_{\mathrm{in}}),$$

in which case C_p can also be seen as an estimator of Err.

Linear Smoothers

For $\hat{\mathbf{f}} = \mathbf{S}_\lambda \mathbf{Y}$ with $\mathbf{S}_\lambda = \mathbf{S}_\lambda(\mathbf{X})$ not depending upon $\mathbf{Y}$ we had

$$\hat{\mathrm{Err}}_{\mathrm{in}} = \bar{\mathrm{err}} + \frac{2\mathrm{trace}(S)}{N}\hat{\sigma}^2,$$

It justifies the definition of $\mathrm{trace}(\mathbf{S}_\lambda)$ as the effective degrees of freedom for model selection – but $\mathrm{trace}(\mathbf{S}_\lambda)$ is now $\mathbf{X}$-dependent.

Likelihood Loss

The generalized decision theoretic setup has sample spaces E and F, action space $\mathcal{A}$, decision rule $f : E \to \mathcal{A}$ and loss functions $L : F \times \mathcal{A}$. If h_a for $a \in \mathcal{A}$ denotes a collection of densities on F we define the minus-log-likelihood loss function as

$$L(y, a) = -\log h_a(y)$$

The empirical loss for $(x_1, y_1), \ldots, (x_N, y_N)$ when using decision rule f is

$$\frac{1}{N} \sum_{i=1}^{N} L(y_i, f(x_i)) = -\log \prod_{i=1}^{N} h_{f(x_i)}(y_i)$$

With $\mathcal{F}$ a class of decision rules empirical risk minimization over $\mathcal{F}$ coincides with conditional maximum likelihood estimation of $f \in \mathcal{F}$. Expected prediction error equals the expectation of (conditional) cross entropies.

$$\mathrm{EPE}(f) = \int \underbrace{\int -\log h_{f(x)}(y) g(y|x) \mathrm{d}y}_{\text{cross entropi}} g_1(x) \mathrm{d}x$$

Akaike's Information Criteria – AIC

We take $\mathcal{A} = \{f_\theta(x, \cdot)\}_{\theta \in \Theta, x \in E}$ with Θ being d-dimensional and $f_\theta : E \times F \to [0, \infty)$ such that $f_\theta(x, \cdot)$ is a probability density on F. Let $\hat{\theta}_N$ denote the MLE.

With likelihood loss we define the equivalent of the in-sample error

$$\mathrm{Err}_{\mathrm{loglik,in}} = -\frac{1}{N} \sum_{i=1}^{N} E(\log f_{\hat{\theta}_N}(x_i, Y_i^{\mathrm{new}}) | \mathbf{X})$$

Then one derives the approximation

$$\mathrm{Err}_{\mathrm{loglik,in}} \simeq \frac{1}{N} E(l_N(\hat{\theta}_N)) + \frac{d}{N}$$

where the minus-log-likelihood function in $\hat{\theta}_N$

$$l_N(\hat{\theta}_N) = -\frac{1}{N} \sum_{i=1}^{N} \log f_{\hat{\theta}_N}(x_i, y_i)$$

is the equivalent of $\overline{\mathrm{err}}$ when using likelihood loss.

AIC

$$\text{AIC} = \frac{2}{N} l_N(\hat{\theta}_N) + \frac{2d}{N}$$

We use AIC for model selection by choosing the model among several possible that minimizes AIC.

Assumptions and extensions:

- The models considered must be true. If they are not, d must in general be replaced by a more complicated quantity d^* leading to the model selection criteria

$$\text{NIC} = \frac{2}{N} l_N(\hat{\theta}_N) + \frac{2d^*}{N}.$$

- For linear regression with Gaussian errors and fixed variance $d^* = d$ even when the model is wrong, but this does not hold in general, e.g. logistic regression.
- The estimator $\hat{\theta}_N$ must be the MLE. Extensions to non-MLE and non-likelihood loss setups are possible with d replaced again by a more complicated d^*.

AIC

- For model comparison there is theoretical evidence that

$$\mathrm{AIC}_1 - \mathrm{AIC}_2 = \frac{2}{N}(l_N^1(\hat{\theta}_N^1) - l_N^2(\hat{\theta}_N^2)) + \frac{2(d_1 - d_2)}{N}$$

  can be a (much) better approximation of the difference in $\mathrm{Err}_{\mathrm{loglik,in}}$ when the models are nested than if the models are non-nested.
- The deviance equals $2l_N(\hat{\theta}_N)$ up to an additive constant – often the value of twice the minus-log-likelihood in the "saturated model". For model comparisons we can replace $2l_N(\hat{\theta}_N)$ by the deviance, but make sure that all models considered use the same reference model/additive constant in their definition of the deviance.

Figure 7.4 – AIC Used for Model Selection

This figure provide some empirical justification of using AIC in a context where there is no theoretical justification. The 0-1 loss is not a minus-log-likelihood.

Practical BIC

With the same framework as for AIC

$$\mathrm{BIC} = 2l_N(\hat{\theta}_N) + d\log(N)$$

We choose among several models the one with the smallest BIC.

Up to the scaling by $1/N$, BIC is from a practical point of view AIC with 2 replaced by $\log(N)$. The theoretical derivation is, however, completely different.

For $N > e^2 \simeq 7.4$, BIC penalizes complex models more than simple models compared to AIC.

AIC in Unconditional Models

With $(P_\theta)_{\theta\in\Theta}$ a parameterized family of probability measures on $E \times F$, h_θ the joint density of (X, Y),

$$l_N(\theta) = -\sum_{i=1}^{N} \log h_\theta(x_i, y_i)$$

the joint minus-log-likelihood function, and $\hat{\theta}_N$ the MLE then

$$\mathrm{AIC} = \frac{2}{N} l_N(\hat{\theta}_N) + \frac{2d^*}{N}.$$

Under suitable regularity conditions

$$E(\mathrm{AIC}) = -E(\log h_{\hat{\theta}_N}(X, Y)) = E(\underbrace{-E(\log h_{\hat{\theta}_N}(X, Y)|\mathbf{X}, \mathbf{Y})}_{\text{cross entropy}})$$

is the expected cross entropy of $P_{\hat{\theta}}$ from the true distribution of (X, Y) If the model is true then $d^* = d$ where d is the dimension of Θ.

Other Ideas and Methods

An alternative to the approximations of expectations we can consider upper bounds.

Let $\eta \in [0, 1]$ and $h(\mathcal{F})$ denote a number – a complexity measure – for the class $\mathcal{F}$ such that with probability at least $1 - \eta$

$$EPE(\hat{f}) \leq \bar{\text{err}} + g(h(\mathcal{F}), \bar{\text{err}})$$

The theory by Vapnik based on the Vapnik-Chervonenkis dimension (VC dimension) provides such upper bounds. General upper bounds are nice but almost always extremely pessimistic. The theoretical justification is extremely hard. But it does not rule out the practical use of the upper bounds for model selection ....