Model Selection – Optimization versus Simplicity

There are two opposing philosophies on how to draw inference from empirical data:

- Find the model that fits the data best.
- Keep it simple. Don't choose a complicated model over a simpler model if the simpler model suffice (Occam's razor).

Ideas like maximum-likelihood and empirical loss minimization work by the first principle.

Model selection/test theory work by the second principle to compensate for the fact that an optimization procedure generally overfits to the given data set.

Model or Predictor Assessment

Another problem of importance is

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

The training error

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

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

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

is the expected EPE.

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

Figure 7.1 – The Bias-Variance Tradeoff

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

The Bias-Variance Tradeoff for nearest neighbors

Consider the case with $Y = f(X) + \epsilon$ where X and ϵ are independent, $E(\epsilon) = 0$ and $V(\epsilon) = \sigma^2$.

With $\hat{f}_k$ the k-nearest neighbor regressor the expected test error, conditionally on the training data $X_1 = x_1, \ldots, X_N = x_N$, in x_0 is

$$\begin{aligned} E((Y - \hat{f}_k(x_0))^2 \mid X = x_0) &= E((Y - f(x_0))^2 \mid X = x_0) \\ &\quad + E((f(x_0) - E(\hat{f}_k(x_0)))^2 \mid X = x_0) \\ &\quad + E((\hat{f}_k(x_0) - E(\hat{f}_k(x_0)))^2 \mid X = x_0) \\ &= \sigma^2 + \underbrace{\left[f(x_0) - \frac{1}{k} \sum_{l \in N_k(x_0)} f(x_l) \right]^2}_{\text{Squared bias}} + \underbrace{\frac{\sigma^2}{k}}_{\textit{variance}} \end{aligned}$$

Small choices of k (complex model) will give a large variance and generally a smaller bias, and vice versa for large choices of k (simple model).

Figure 2.4 – Bias-Variance Tradeoff for k-nearest neighbors

The test error here is an estimate of $\text{EPE}(\hat{f})$ for the estimated predictor based on an independent dataset.

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, for the purpose of estimating parameters of secondary importance?

Figure 7.2 – Space of Models

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

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

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 expected 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

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

Mallows' C_p

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

is an equivalent of Mallows' C_p statistic – 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 $\mathbf{S}$ is a linear smoother, that is, $\hat{\mathbf{f}} = \mathbf{S}\mathbf{y}$, one can generalize C_p as

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

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

$$\hat{\sigma}^2 = \frac{1}{N - \mathrm{trace}(2\mathbf{S}_0 - \mathbf{S}_0^2)}||\mathbf{y} - \mathbf{S}_0\mathbf{y}||^2.$$

for a "low-bias" smoother $\mathbf{S}_0$.

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

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

As a function of d the classical definition of C_p,

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

is a monotonely increasing function of our C_p.

Generalization Error

Instead of the expected in-sample error we can consider the expected 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!

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 $L : F \times \mathcal{A} \to [0, \infty)$. 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

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

Expected prediction error equals the expectation of (conditional) cross entropies.

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

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 expected in-sample error

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

Then one derives (difficult) the approximation

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

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 $\bar{\text{err}}$ 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^*.

Practical BIC

With the same framework as for AIC

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

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.

Cross-Validation

Let $\kappa : \{1, \ldots, N\} \to \{1, \ldots, K\}$ and denote by $\hat{f}^{-k}$ for $k = 1, \ldots, K$ the estimator of f based on the data (x_i, y_i) with $\kappa(i) \neq k$.

The (x_i, y_i) with $\kappa(i) = k$ work as a test dataset for $\hat{f}^{-k}$ and

$$\hat{\mathrm{PE}}(\hat{f}^{-k}) = \frac{1}{N_k} \sum_{i:\kappa(i)=k} L(y_i, \hat{f}^{-k}(x_i))$$

with $N_k = |\{i | \kappa(i) = k\}|$

The K-fold κ-cross-validation estimator of Err is the weighted average

$$\begin{aligned} CV_\kappa &= \sum_{k=1}^{K} \frac{N_k}{N} \hat{\mathrm{PE}}(\hat{f}^{-k}) \\ &= \frac{1}{N} \sum_{i=1}^{N} L(y_i, \hat{f}^{-\kappa(i)}(x_i)) \end{aligned}$$

Figure 7.8 – Err as a Function of N

We should write $\text{Err} = \text{Err}(N)$ as a function of the sample size. If $\hat{f}_N, f \in \mathcal{F}$ and f minimizes EPE then $\text{EPE}(\hat{f}_N) \geq \text{EPE}(f)$ and

$$\text{Err}(N) = E(L(Y, \hat{f}_N(X))) \geq \text{EPE}(f)$$

If we have a consistent estimator; $\hat{f}_N \to f$, then

$$\text{Err}(N) \to \text{EPE}(f).$$

Cross-Validation

Among several models we will choose the model with smallest CV_κ. How to choose κ? How to choose K?

We aim for $N_1 = \ldots = N_K$ in which case

$$E(CV_\kappa) = \text{Err}(N - N_1).$$

With a steep learning curve at N we need N_1 to be small or we underestimate Err.

Extreme case; N-fold or leave-one-out cross-validation with $\kappa(i) = i$ leads to an almost unbiased estimator of $\text{Err}(N)$, but the strong correlation of the $\hat{\text{EPE}}(\hat{f}^{-i})$'s works in the direction of given a larger variance. Recommendations are that 5- or 10-fold CV is a good compromise between bias and variance.

The Wrong and The Right Way to Cross-Validate

- Mess with the data to find variables/methods that seem to be useful.
- Estimate parameters using the selected variables/methods and use cross-validation to choose tuning parameters.

WRONG

Don't mess with the data before the cross-validation.

Cross-Validation must be out side of all modeling steps, including filtering or variable selection steps.

Estimates of Expected Prediction Error

If $\hat{f}$ is estimated based on a data set, we can only get an estimate of $\text{EPE}(\hat{f})$ by an independent test set $(x_1, y_1), \ldots, (x_B, y_B)$ as

$$\hat{\text{EPE}}(\hat{f}) = \frac{1}{B} \sum_{b=1}^{B} L(y_b, \hat{f}(x_b)).$$

Cross-validation provide estimates $\hat{\text{Err}}$ of the expected generalization error.

- $\text{EPE}(\hat{f})$ is a random variable with mean Err.
- $\hat{\text{Err}}$ is a random variable with mean Err.

Can $\hat{\text{Err}}$ be regarded as an approximation/estimate of $\text{EPE}(\hat{f})$?

Figure 7.15 – The Relation Between $\hat{\text{Err}}$ and $\text{EPE}(\hat{f})$

Classification and The Confusion Matrix

For a classifier with two groups we can decompose the errors:

	Predicted y	
Observed y	1	0
1	$\Pr(Y = 1, f(X) = 1)$	$\Pr(Y = 1, f(X) = 0)$
0	$\Pr(Y = 0, f(X) = 1)$	$\Pr(Y = 0, f(X) = 0)$

This is the confusion matrix and

$$\text{EPE}(f) = \Pr(Y = 0, f(X) = 1) + \Pr(Y = 1, f(X) = 0).$$

As with $\text{EPE}(\hat{f})$ the confusion matrix can only be estimated using an independent test dataset. "Estimates" based on e.g. cross-validation are estimates of $E(\Pr(Y = k, \hat{f}(X) = l))$.

EXTRA: Linear Smoother Bias-Variance Decomposition

Assumptions: $\hat{\mathbf{f}} = \mathbf{S}\mathbf{Y}$ and conditionally on $\mathbf{X}$ the Y_i's are uncorrelated with common variance σ^2.
Then with $\mathbf{f} = E(\mathbf{Y}|\mathbf{X}) = E(\mathbf{Y}^{\text{new}}|\mathbf{X})$ and $\mathbf{Y}^{\text{new}}$ independent of $\mathbf{Y}$

$$\begin{aligned} E(||\mathbf{Y}^{\text{new}} - \hat{\mathbf{f}}||^2|\mathbf{X}) &= E(||\mathbf{Y}^{\text{new}} - \mathbf{S}\mathbf{Y}||^2|\mathbf{X}) \\ &= E(||\mathbf{Y}^{\text{new}} - \mathbf{f}||^2|\mathbf{X}) + ||\mathbf{f} - \mathbf{S}\mathbf{f}||^2 \\ &\quad + E(||\mathbf{S}(\mathbf{f} - \mathbf{Y})||^2|\mathbf{X}) \\ &= N\sigma^2 + \underbrace{||(I - \mathbf{S})\mathbf{f}||^2}_{\text{Bias}(\lambda)^2} + \sigma^2 \text{trace}(\mathbf{S}^2) \\ &= \sigma^2(N + \text{trace}(\mathbf{S}^2)) + \text{Bias}(\lambda)^2 \end{aligned}$$

where we use that $E(\hat{\mathbf{f}}|\mathbf{X}) = E(\mathbf{S}\mathbf{Y}|\mathbf{X}) = \mathbf{S}\mathbf{f}$.

EXTRA: Estimation of σ^2 using low bias estimates

Take

$$\text{RSS}(\hat{\mathbf{f}}) = \sum_{i=1}^{N} (y_i - \hat{\mathbf{f}}_i)^2$$

is a natural estimator of $E(||\mathbf{Y} - \hat{\mathbf{f}}||^2 | \mathbf{X})$. Its mean is then

$$\sigma^2(N - (\text{trace}(2\mathbf{S} - \mathbf{S}^2)) + \text{Bias}(\lambda)^2.$$

Choosing a low-bias – that is trace$(2\mathbf{S} - \mathbf{S}^2$ is large – model, we expect Bias$(\lambda)^2$ to be negligible and we estimate σ^2 as

$$\hat{\sigma}^2 = \frac{1}{N - \text{trace}(2\mathbf{S} - \mathbf{S}^2)} \text{RSS}(\hat{\mathbf{f}}).$$

From this point of view

$$\text{trace}(2\mathbf{S} - \mathbf{S}^2)$$

can be justified as the effective degrees of freedom.