36-724 Spring 2006: Assessing Model Fit

Brian Junker

March 28, 2006

- Expected Prediction Error
- Optimism in Training Error and AIC
- Model Selection and BIC
- Fit - Complexity Tradeoff
- Vapnik-Chernovenkis Dimension
- CV vs AIC vs BIC vs SRM??
There is usually some explicit dependence on parameters \( \theta(X) \), so that 
\[
\hat{f}(X) = f(X; \hat{\theta}(X)) \text{ or } \hat{p}(y|x) = p(y|x; \hat{\theta}(x)).
\]

We may be interested in

- **Model estimation**: Choosing \( \hat{\theta}(X) \) by minimizing loss (e.g. squared-error or log-likelihood).

- **Model selection**: Choosing among models \( f(X; \hat{\theta}(X)) \) (or \( p(y|x; \hat{\theta}(x)) \)).

Especially for model selection, there may be some further feature of the model, say “complexity”, indexed by an additional parameter \( \alpha \):

\[
\begin{align*}
f(X; \hat{\theta}(X)) &= f_\alpha(X; \hat{\theta}_\alpha(X)) \text{ or } \\
p(y|x; \hat{\theta}(x)) &= p_\alpha(y|x; \hat{\theta}_\alpha(x))
\end{align*}
\]

so that model selection often amounts to choosing \( \alpha \).
Optimism in Training Error and AIC

We can evaluate fit or predictive/classification success in one of several ways, e.g.

- **Training error**: $\bar{\text{err}} = \frac{1}{N} \sum_{i=1}^{N} L(y_i, \hat{f}(x_i))$;
- **In-sample error**: $\text{Err}_{\text{in}} = \frac{1}{N} \sum_{i=1}^{N} E_{y_{\text{true}}} E_{y_{\text{new}}} L(Y_{i_{\text{new}}}, \hat{f}(x_i))$;
- **Extra-sample error**: $\text{Err} = E[L(Y, \hat{f}(X))];$

We typically expect these to be increasing in size:

- $\bar{\text{err}}$ capitalizes on chance by using the same $y_i$’s for training and for testing;
- $\text{Err}_{\text{in}}$ considers new $y_i$’s for testing but uses the same design points $x_i$ for training and testing;
- $\text{Err}$ considers both new $y_i$’s and new $x_i$’s for testing.

First we consider in-sample error.

The **optimism** of training error relative to in-sample error,

$$\text{op} = \text{Err}_{\text{in}} - E_{y_{\text{true}}} \bar{\text{err}}$$

is typically positive. In fact, for squared-error loss, 0-1 loss, and other loss functions, one can show

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

and in particular for linear smoothers $\hat{y} = S\lambda y$,

$$\sum_{i=1}^{N} \text{Cov} (\hat{y}_i, y) = d\sigma^2, \quad d = \text{trace}(S\lambda)$$

Thus

$$\text{Err}_{\text{in}} = E_{y_{\text{true}}} \bar{\text{err}} + \text{op}$$

$$= E_{y_{\text{true}}} \bar{\text{err}} + \frac{2}{N} \sum_{i=1}^{N} \text{Cov} (\hat{y}_i, y_i) \quad \text{(for many loss functions)}$$

$$= E_{y_{\text{true}}} \bar{\text{err}} + 2d\sigma^2/N \quad \text{(for linear smoothers)}$$
Model Selection and BIC

We saw earlier in the course that a Bayesian approach to model selection would embed the problem in a single hierarchical model

- Level 1: \( y \sim p(y|\theta, M_\ell) \), \( \ell = 1, 2, \ldots, M \)
- Level 2: \( \theta \sim p(\theta|M_\ell) \), \( \ell = 1, 2, \ldots, M \)
- Level 3: \( p(M_\ell) = p_\ell \), \( \sum_\ell p_\ell = 1 \)

Then,

\[
p(y|M_\ell) = \int p(y|\theta, M_\ell) p(\theta|M_\ell) d\theta
\]

\[
p(M_m|y) = \frac{p(y,M_m)}{p(y)} = \frac{p(y|M_m)p(M_m)}{\sum_\ell p(y|M_\ell)p(M_\ell)}
\]

If we take \( p_\ell = p(M_\ell) \equiv 1/M \), and we use the Laplace approximation to the posterior under each model

\[-2 \log p(y|M_\ell) \approx BIC_\ell = -2 \log p(y|\hat{\theta}, M_\ell) + d_\ell \log(N)\]

then

\[
p(M_m|y) \approx \frac{e^{-\frac{1}{2}BIC_m}}{\sum_\ell e^{-\frac{1}{2}BIC_\ell}}
\]

Relationship with AIC

The general form of in-sample error estimates is thus

\[\bar{\text{Err}}_\text{in} = \bar{\text{err}} + \delta p\]

which leads immediately to measures like Mallows’ \( C_p \)

\[C_p = \bar{\text{err}} + 2 \frac{d}{N} \hat{\sigma}_e^2\]

this is of the same form as Akaike’s Information Criterion

\[AIC = -\frac{2}{N} \log P(Y|\hat{\theta}) + 2 \frac{d}{N}\]

which we motivated earlier as a Taylor approximation to \( \bar{\text{Err}}_\text{in} = \frac{1}{N} E_{\text{train}} E_{\text{pred}} L(Y_{\text{new}}, \hat{\theta}) \), in the case of log-likelihood loss

\[L(Y_{\text{new}}, \theta) = -2 \log P(Y|\theta) = -2 \sum_{i=1}^{N} p(y_i|\theta)\]

Thus the AIC penalty can be viewed as an estimate of the optimism in training-sample error measures.
**Vapnik-Chernovenkis Dimension**

Vapnik-Chernovenkis (VC) dimension is a way of measuring complexity that does not depend on locally linear structure.

Let $f(x, \alpha)$ be an indexed set of indicator functions (for 2-group classification).

**Definition** The set of points $(x_1, \ldots, x_n)$ can be *shattered* by $f(x, \alpha)$ if, for *every* set of binary labels $y_1, \ldots, y_n$, there is a value $\alpha$ so that $f(x_j, \alpha) = y_j$.

**Definition** The *VC dimension* of the family $f(x, \alpha)$ is the largest number of points that can be arranged so that $f(x, \alpha)$ can shatter them.

If $g(x, \alpha)$ is a real-valued function, the VC dimension of $g(x, \alpha)$ is the VC dimension of $f(x, \alpha) = 1_{[g(x,\alpha)-\beta>0]}$. 

---

**Fit - Complexity Tradeoff**

As we have observed before, BIC is basically AIC with a larger penalty ($d \log(N)$ vs. $2d$), but the motivations for the penalties are rather different:

- In AIC, $2d$ is an estimate of the optimism in the training error;
- In BIC, $d \log(N)$ is an estimate of the curvature of the posterior near the mode $\hat{\theta}$.

Both AIC and BIC are of the form

$$\text{Index} = \text{Fit} + \text{Complexity}$$

analogous to $\text{MSE} = \text{Bias}^2 + \text{Variance}$.

Our estimates of complexity depend on our measure of model dimension $d$:

- $d = \text{trace } S_\lambda$
- $d = E[D|y] - D(E[\theta|y]), D = -2 \log p(y|\theta)$

are closely related (Spiegelhalter et al., *JRSSB*, 2002; Hodges and Sargent, *Bmka*, 2001) and work well only for locally linear models.
What does VC dimension measure?

- In nice cases it is the number of parameters, e.g.
  \[ f(x, \alpha) = 1_{(\alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 > 0)}, \quad (x_1, x_2)^T \in \mathbb{R}^2 \]
  has VC dimension \( h = p + 1 \) (can you prove it?).

- The example of \( \sin(\alpha \cdot x) \) above shows that a model with one parameter can have very high VC dimension.

- It is also possible to create models with 3 or 4 or 7 or more parameters that have VC dimension only 1 or 2 (how?).

So VC dimension is related to the number of parameters, but it is not really the same.
**Structural Risk Minimization**

VC-dimension can be used to look at the optimism of $\overline{err}$ with respect to extra-sample error:

Let $h$ be the VC dimension of the binary classifier $f(x, \alpha)$.

Vapnik showed that for the binary classification problem with 0/1 loss,

$$P_T \left[ \text{Err} \leq \overline{err} + \sqrt{\frac{h(\log(2N/h) + 1) - \log(\eta/4)}{N}} \right] \geq 1 - \eta$$

where $N = |T|$ is the number of training examples. (somewhat tighter bounds for classification, and for regression, are given in HTF).

The *structural risk minimization approach* fits a sequence of nested models with increasing VC dimension $h_1, h_2, \ldots$, and then chooses the model that minimizes the bound

$$\overline{err} + \sqrt{\frac{h(\log(2N/h) + 1) - \log(\eta/4)}{N}}$$

---

**CV vs AIC vs BIC vs SRM??**

- VC dimension is defined for a wider class of models than $d = \text{trace} S$, but unfortunately it can be very hard to compute!
- All three (AIC, BIC, SRM) only require you to compute the training error $\overline{err}$ and not full cross-validation (CV) error.
- BIC is more conservative than AIC.
- SRM is “wildly conservative”, but has been successful with a variety of machine learning approaches like decision trees, neural net, support vector machines, etc.
- As with Mallows' $C_p$, asymptotically it can be shown that AIC and leave-one-out ($N$-fold) CV should give the same results.
- Asymptotically, BIC has been shown to coincide with carefully chosen $K$-fold CV.
- Asymptotically BIC should pick the model that generated the data (assuming there is one!); so BIC is better if you want the best *structure* rather than the best *predictor*.
- There are alternatives to the above (like fully Bayesian model selection or model averaging!).

(Based on comments of Andrew Moore [see http://www.autonlab.org/tutorials])