Statistical Methods - Model evaluation, comparison and selection

Lecture 9

Mattias Villani

Sveriges Riksbank and Stockholm University

May 24, 2010

Overview Lecture 9

- Approaches to model comparison and evaluation
- Evaluating prediction performance
- Model selection criteria
- Bayesian model inference

Approaches to model comparison

- Given a set of models, how do we compare them?
- Classical testing: likelihood ratio test. Restricted to nested models (some extension to the non-nested case have been proposed). Mostly asymptotic. Pair-wise comparison of models.
- Information criteria. AIC, BIC etc
- Prediction-based comparison. Cross-validation and Bootstrap.
- Bayesian marginal likelihood comparison.

Approaches to model evaluation

- The best model in a class of models can still be bad ...
- How do we know if a model is 'any good'? First question: good for **what**?
  - Prediction (can we accurately predict a future event? Black box model/algorithm may be OK)
  - Interpretation (can we learn something from the model structure? Is it useful for organizing our thinking? May be OK even if predictions are inaccurate)
  - Policy analysis (what happens with the response when we change the conditioning variables? Both prediction performance and interpretation usually matters here)
- If prediction is our concern, how can we estimate the expected performance of the model in a new data set? How well does the model generalize from in-sample to out-of-sample?
- Weak evaluation of the model: If simulated data from your model looks nothing like actual data (maybe even the same data used to fit the model!), then the model is 'no good'.

Evaluating prediction performance

- Components:
  - Target variable: Y.
  - Covariates/Inputs: X.
  - Prediction model $\hat{f}(X)$
  - Training sample $\mathcal{T}$, used to estimate $\hat{f}(X)$
  - Loss function: $L[Y, \hat{f}(X)]$ measuring the discrepancy between true Y and the model's prediction $\hat{f}(X)$. Examples:
    - Squared loss $[Y - \hat{f}(X)]^2$
    - Linear loss $\left|Y - \hat{f}(X)\right|$
    - Lin-lin loss:

$$L[Y, \hat{f}(X)] = \begin{cases} c_1 \cdot \left|Y - \hat{f}(X)\right| & \text{if} \quad Y \leq \hat{f}(x) \\ c_2 \cdot \left|Y - \hat{f}(X)\right| & \text{if} \quad Y > \hat{f}(x) \end{cases}$$

Generalization error - Prediction performance on new data

- **Generalization error**, or test error, is defined as

$$\mathrm{Err}_{\mathcal{T}} = \mathrm{E}\{L[Y, \hat{f}(X)]|\mathcal{T}\}$$

  The expectation is with respect to the joint distribution of (Y, X). The training set $\mathcal{T}$ is fixed, so $\mathrm{Err}_{\mathcal{T}}$ is the prediction error we can expect on new independent data **given** the particular training sample at hand. In a given application, $\mathrm{Err}_{\mathcal{T}}$ is what we care about.

- The **expected prediction error** is defined as

$$\mathrm{Err} = \mathrm{E}(\mathrm{Err}_{\mathcal{T}}),$$

  where the expectation is with respect to the training sample $\mathcal{T}$. Err measures the average prediction performance of the method/model over all possible training samples. This is usually less relevant when we actually have a particular training sample at hand.

- **Training error** is the average loss over the training sample:

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

  $\bar{err}$ can be made arbitrarily small by increasing the model complexity and we typically have $\bar{err} < \mathrm{Err}_{\mathcal{T}}$.
- **In-sample error**

$$\mathrm{Err}_{\mathrm{in}} = \frac{1}{N}\sum_{i=1}^{N} \mathrm{E}_{Y^0} L[Y^0, \hat{f}(x_i)]$$

  where the x_i are the training sample points, but each Y_i^0 is a new response observation drawn from $p(Y^0|X = x_i)$.
- Note that $\mathrm{Err}_{\mathrm{in}}$ conditions on the values of the covariates observed in the training sample, whereas $\mathrm{Err}_{\mathcal{T}}$ does not.
- Note the distinction between in-sample error and training error, it is easy to confuse them!

Optimism of the training error rate

- The **optimism** (in the training error) is defined as

$$\mathrm{op} = \mathrm{Err_{in}} - \overline{\mathrm{err}}$$

and the average optimism is

$$\omega = \mathrm{E}_{\mathbf{y}}(\mathrm{op}),$$

where $E_{\mathbf{y}}$ denote the expectation with respect to the outcomes of the y_i, $i \in \mathcal{T}$, but the x's in the training set are fixed.

- One can show quite generally that

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

Very intuitive: the more effect the observations have on their own fit, the larger the optimism.

Optimism of the training error rate. cont.

- Important relation:

$$\mathrm{E}_{\mathbf{y}}(\mathrm{Err}_{\mathrm{in}}) = E_{\mathbf{y}}(\bar{err}) + \frac{2}{N}\sum_{i=1}^{N} Cov(y_i, \hat{y}_i)$$

which implies that we can estimate the in-sample error by

$$\hat{\mathrm{Err}}_{in} = \bar{err} + \hat{\omega},$$

where $\hat{\omega}$ is an estimate of the average optimism.

Estimating the in-sample error

- Example: linear smoothers: $\hat{\mathbf{y}} = L\mathbf{y}$.

$$\begin{aligned} Cov(\mathbf{y}, \hat{\mathbf{y}}) &= \mathrm{E}[\mathbf{y} - \mathrm{E}(\mathbf{y})][\hat{\mathbf{y}} - \mathrm{E}(\hat{\mathbf{y}})]' \\ &= \mathrm{E}[\mathbf{y} - \mu][L\mathbf{y} - \mathrm{L}—]' \\ &= \mathrm{Cov}(y)L' \end{aligned}$$

so if $Cov(\mathbf{y}) = \sigma_\varepsilon^2 I$, then $\sum_{i=1}^{N} Cov(y_i, \hat{y}_i) = \mathrm{tr}Cov(\mathbf{y}, \hat{\mathbf{y}}) = \sigma_\varepsilon^2 \mathrm{trL} = \sigma_\varepsilon^2 \cdot Df$. Here we have the estimate of the in-sample error

$$\hat{\mathrm{Err}}_{in} = \bar{err} + 2 \cdot \frac{Df}{N}\sigma_\varepsilon^2.$$

- This motivates the following definition of Df for general additive-error models $y = f(x) + \varepsilon$ (which may not be linear smoothers)

$$Df(\hat{y}) = \frac{\mathrm{tr}Cov(\mathbf{y}, \hat{\mathbf{y}})}{\sigma_\varepsilon^2} = \frac{\sum_{i=1}^{N} Cov(y_i, \hat{y}_i)}{\sigma_\varepsilon^2}.$$

Model selection criteria

- The estimate of the in-sample error is similar to the more generally applicable Akaike Information Criterion (AIC)

$$AIC = -2 \cdot loglik + 2 \cdot p,$$

where loglik is the maximum of the likelihood function, and p is the number of effective parameters ($p = Df$ for linear smoothers).

- Note that $\hat{\mathrm{Err}}_{in}$ and AIC are in-sample measures and are poor estimators of the generalization error. They can be used to **compare** models, however.
- But AIC is not a consistent model selection criteria ...
- BIC (Bayesian Information Criteria)

$$BIC = -2loglik + \ln N \cdot p$$

- BIC is consistent and penalizes more complex models more heavily than AIC. BIC can be converted to model probabilities, see next lecture.