STATS 235: Modern Data Analysis
Model Assessment and Selection

Babak Shahbaba

Department of Statistics, UCI

Modeling objectives

- We now focus on supervised learning methods.
- We discuss statistical model where the main objective is predicting the values of a response variable for future observations (or in general those that we haven't seen yet).
- Theoretically, we can search among all possible models (e.g., with different predictors) and choose the one with the best prediction accuracy (i.e., model selection).
- When we find the best model, we need to estimate its prediction accuracy for future data (i.e., model assessment).

Prediction error

- We typically evaluate predictive models based on their prediction error presented as the expectation of an assumed loss function, L,

$$\text{Err} = E[L(y, \hat{y})]$$

  where y is the observed value of the response variable, and $\hat{y}$

- For regression models, we usually set $L(y, \hat{y}) = (y - \hat{y})^2$

- For classification models, we usually set $L(y, \hat{y}) = 1$ when $y \neq \hat{y}$, and zero otherwise; this is known the 0-1 loss function

Prediction error

- We use the observed data to estimate error
- Building a predictive model based on the observed data and evaluating it based on the same data will provide optimistic estimates of prediction error
- The optimistic prediction error on the the training data set itself is called the "apparent error"
- To avoid this issue, we usually use an independent test set to estimate prediction error
- If the sample size is large enough, we can divide the observed data into two independent training and test sets

Cross-validation

- When the sample size is relatively small, we recycle the data using K-fold cross-validation (CV)
  - Split the data into K roughly equal parts
  - For $k = 1, \ldots, K$, treat the kth part as the test set to evaluate the model trained on the remaining $K - 1$ parts
  - Obtain the CV estimate of prediction error as

$$\widehat{\mathrm{Err}}_{CV} = \frac{1}{n} \sum_{i=1}^{n} L(y_i, \hat{y}_i)$$

    where $\hat{y}_i$ is the prediction using the parts that do not include the ith observation

- Setting $K = n$ is called "leave-one-out"

Training, validation, and test

- Many model building strategies involve fine tuning a set of parameters, whose values affect model performance.
- For example, the tuning parameter could be a vector of p binary indicators $r_1, \ldots, r_p$ such that if $r_j = 1$, we include the j^{th} predictor in our model.
- We usually choose an appropriate values for such parameters based on the performance of the model on a third dataset (usually a subset of the training set) called the *validation* set.
- After we decided the values of these tuning parameters, and fix the model, we evaluate its performance on the test set.

Model selection as a decision problem

- As we discussed before, model comparison is more appropriately discussed as a decision problems.
- This is specially true in the Bayesian paradigm.
- In this setting, our decision to accept model M_1 over the alternative model M_0 depends not only on the posterior probability of M_1 and M_0, but also on the assumed loss function for such decision.

Model selection as a decision problem

- Recall that we use $\mathcal{V}$ to denote the set of all possible values, v, we need to predict. We refer to $\mathcal{V}$ as the *outcome space*.
- When choosing between two different models, $\mathcal{V} = \{M_0, M_1\}$.
- We present the set of all possible actions, a, as $\mathcal{A}$. We refer to $\mathcal{A}$ as the *action space*, which in our case is related to the act of selecting a model.
- When choosing between two models, $\mathcal{A} = \{M_0, M_1\}$.

Model selection as a decision problem

- For model selection problems, we could use the $0-1$ loss function:
- $L(M_0, M_0) = L(M_1, M_1) = 0, L(M_0, M_1) = L(M_1, M_0) = 1.$
- In this case, the formal Bayes rule based on choosing the model with a smaller posterior risk is the same as choosing the model with a higher posterior probability.

Posterior odds

- Suppose that we believe the model probabilities are $P(M_0)$ and $P(M_1)$ *a priori*.
- We could compare posterior probabilities by presenting them in the form of a posterior odds $P(M_0|y)/P(M_1|y)$ as follows:

$$\frac{P(M_0|y)}{P(M_1|y)} = \frac{P(M_0)P(y|M_0)/P(y)}{P(M_1)P(y|M_1)/P(y)} = \frac{P(M_0)P(y|M_0)}{P(M_1)P(y|M_1)}$$

- That is, the posterior odds is the prior odds, $P(M_0)/P(M_1)$, multiplied by the likelihood ratio, $P(y|M_0)/P(y|M_1)$.

Bayes factor

- Traditionally, statisticians avoid expressing prior odds in favor of one of the alternatives (especially if we are not making a decision, rather, we are reporting our findings).

- Therefore, $P(M_0)/P(M_1) = 1$ so we rely only on

$$P(y|M_0)/P(y|M_1)$$

  which is known as Bayes factor (BF).

- This is analogous (but not the same) to the likelihood ratio test that is commonly used in the frequentist framework.

- Jeffreys (1961) provided interpretive ranges for the BF analogous to what frequentists use for p-values.

Bayesian information criterion

- A related, yet simpler, approach for model selection is based on Bayesian information criterion (BIC).
- Using Laplace's approximation (see Ripley, 1996, page 64), we have

$$\log[P(y|M)] \approx \log[P(y|\hat{\theta}, M)] - \frac{k}{2} \log n$$

where, $\hat{\theta}$ is the maximum likelihood estimate of the parameters of model M, k is the number of free parameters in the model, and n is the sample size.

- We define Bayesian information criterion (BIC) as follows

$$\text{BIC} \quad = \quad -2 \log[P(y|\hat{\theta}, M)] + k \log n$$

and choose the model with the lowest BIC.

Deviance

- The first term in BIC is known as the *deviance*.
- The deviance is a common measure of discrepancy (i.e., lack of fit) between the data and the model (i.e., the lower deviance, the better the model), and it is defined as follows

$$D(y, \theta) = -2 \log[P(y|\theta)] = -2\ell(\theta, y)$$

- For the normal probability distribution, for example, we have

$$P(y|\mu, \sigma^2) = \exp\{\frac{-(y-\mu)^2}{2\sigma^2} - \frac{\log(2\pi\sigma^2)}{2}\}$$

- Therefore,

$$D(y, \hat{\mu}) = \sum_i \{(y_i - \hat{\mu})^2/(\sigma^2)\}$$

Akaike's information criterion

- The second term in BIC can be considered as a penalty for model complexity.
- Akaike proposed to set the penalty to $2k$ and use the following criterion for model comparison instead:

$$AIC \quad = \quad -2\log[P(y|\hat{\theta}, M)] + 2k$$

- Among different models, we choose the one with the lowest AIC.