

# STATS 211: Statistical Methods II

## Lecture 2: Model assessment and selection

Babak Shahbaba

UCI, Winter 2012

## Modeling objectives

- In general, we build regression models either for hypothesis testing or prediction.
- When our objective is to perform hypothesis testing, the choice of model (e.g., the covariates included in the model) are dictated by our domain knowledge.
- For example, if we are investigating the relationship between mother's age and child's birthweight, we might want to include mother's smoking status as a covariate if it is known to be an important factor affecting birthweight.
- This lecture deals with problems where our objective is prediction.
- Theoretically, we can search among all possible models (e.g., with different covariates) and choose the one with the best prediction accuracy (i.e., model selection).
- When we find the best model, we usually need to estimate its prediction accuracy for future data (i.e., model assessment).

## Evaluating predictive models

- As mentioned above, if our objective is to use our model to predict the value of the response variable for future observations, we should evaluate our model based on its prediction power defined based on some appropriate loss function.
- For regression models, we typically use the squared error loss function:

$$L(y, \hat{y}) = (y - \hat{y})^2$$

- So we can define our goal as finding the model with the lowest expected prediction error  $EPE = E[L(y, \hat{y})]$ .

## Evaluating predictive models

- For the squared error loss, we have

$$\begin{aligned} EPE &= E(y^2) + E(\hat{y}^2) - 2E(y\hat{y}) \\ &= \text{Var}(y) + E(y)^2 + \text{Var}(\hat{y}) + E(\hat{y})^2 - 2E(y)E(\hat{y}) \\ &= \text{Var}(y) + (E(y) - E(\hat{y}))^2 + \text{Var}(\hat{y}) \end{aligned}$$

- Note that future observations  $y$  are independent of  $\hat{y}$ .
- Here, the first term,  $\text{Var}(y)$ , reflects the random variation of the response variable regardless of what model we use.
- The middle term is the squared of bias for the estimator  $\hat{y}$ .
- The last term, is the variance of the estimator  $\hat{y}$ .
- Therefore, we can write  $EPE = \text{Var}(y) + \text{Bias}^2(\hat{y}) + \text{Var}(\hat{y})$ .
- Note that, only the last two terms depend on our model for  $\hat{y}$ ; so we should try to minimize these two terms.

## Evaluating predictive models

- In general, there is a tradeoff between bias and variance: complex models tend to have lower bias and higher variance, whereas simple models tend to have higher bias and lower variance.
- In standard linear regression models, complexity increases as we increase the number of covariates.
- We can try to find the right balance between variance and bias, for example, by finding the right number of covariates that should be included in the model.
- Alternatively, we can use better models, whose performance is not severely affected by the large number of covariates.
- In what follows, we will discuss several strategies for evaluating the performance of predictive models.

## Data splitting strategy

- For a given model, we are interested in estimating the expected prediction error for future observations.
- When we have enough data, one possible strategy is to treat a part of the observed data as future observations; that is, we do not use them in our model, and we pretend that the values of the response variable for these observations are unknown.
- To this end, we use a subset of the data for building a model, and use a different subset (not used for our model) for estimating model performance for future observations.
- In this case, the first subset is regarded as the *training* set and the second subset is regarded as the *test* set. The two sets are mutually exclusive similar to real situations, where future observations are different from the ones used for building our model.

## Data splitting strategy

- Many model building strategies involve fine tuning a set of parameters, whose values affect model performance. Choosing the values of these parameters is a part of *model selection*: each value corresponds to one possible model, and we attempt to find the best model.
- For example, the tuning parameter could be a vector of  $p$  binary indicators  $r_1, \dots, r_p$  such that if  $r_j = 1$ , we include the  $j^{th}$  covariate in our model.
- We usually choose an appropriate values for such parameters based on the performance of the model on the third subset 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.

## Data splitting strategy

- This way, we use the training set to build several alternative models, choose their tuning parameters based on their performance on the validation set, and after choosing the tuning parameters, we evaluate their performance on the test set.
- For example, we can use 60% of the data for training, 20% for validation, and 20% for testing.

## Cross-validation

- This strategy of splitting data into three subsets works well when the sample size is large. When the sample size is small, we typically use other strategies.
- One possibility is to use the data splitting approach along with some sample re-use strategy such as *cross-validation*.

## Cross-validation

- In cross validation, we divide the data set into  $K$  subsets of almost equal size.
- For  $k = 1, \dots, K$ , we use all the subsets except the  $k^{th}$  subset to train the model.
- We then evaluate the performance of the model (e.g., using the squared error loss) based on the  $k^{th}$  subset.
- This way, we have  $K$  measures for model performance.
- We can use these the average of these  $K$  measures as our estimate of the expected loss (e.g., expected prediction error).
- It is also common to provide the corresponding standard error for this estimator.
- When  $K = N$ , the procedure is call the *leave-one-out* approach.

## Model selection as a decision problem

- Model comparison is more appropriately discussed as a decision problems.
- This is specially true in the Bayesian paradigm.
- Decision theory, in general, provides a mathematical framework for making decisions under uncertainty.
- 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

- 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

- We define *Utility* as a function  $u = U(v, a)$  that maps the product of outcome space and action space to a real number  $u \in \mathcal{R}$  representing how much we gain if we choose action  $a$  and the outcome is  $v$ .
- Alternatively, we might choose a loss function instead of utility (e.g., negative of utility) representing our loss when we choose action  $a$  and the outcome is  $v$ .
- For the model selection problem, the loss function,  $L(v, a)$  can be defined as follows:
- $L(M_0, M_0) = L(M_1, M_1) = 0, L(M_0, M_1) = L(M_1, M_0) = 1$ .
- This is known as the 0 – 1 loss function.

## Model selection as a decision problem

- Now, assume that we have observed data  $y$ .
- Using this data, we want to choose between one of the two possible models.
- The tool for making decision is called *decision rule*, and it's denoted as  $\delta(y)$ . Note that  $\delta$  is function of data (i.e.,  $y$ ) only.

## Posterior risk

- In general, the posterior risk for a decision rule is

$$r(\delta|y) = \int_{\mathcal{V}} L(v, \delta(y)) P(v|y) dv$$

- Note that we replaced the action  $a$  with the decision rule  $\delta(y)$  since our action now depends on our decision rule which itself depends on the observed data.

## Formal Bayes rule

- *The expected loss principle*: In deciding between different rules, choose the one with the smallest posterior risk. That is, take the action according to the rule with the smallest posterior expectation of loss function.
- The resulting rule is called a *formal Bayes rule*.
- $\delta_0(y)$  is a formal Bayes rule if  $r(\delta_0|y) < \infty$  for all  $y$  and  $r(\delta_0|y) \leq r(\delta|y)$  for all  $y$  and  $\delta$ .
- In theory, this is all we need to know for all sorts of decision problems (e.g., model selection, prediction, point estimation, and hypothesis testing).
- For example, if we use a simple 0-1 loss function for the model selection problem, 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  $P(v|y)$ .

## Posterior odds

- Suppose that we believe the model probabilities are  $P(M_0)$  and  $P(M_1)$  *a priori*.
- As mentioned above, with a simple 0-1 loss function, we choose the model with a higher posterior probability.
- 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 in general setting though) 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.

## Bayes factor

- Using the BF has some difficulties. For example, it is typically not possible to use improper priors.
- Other alternatives such as fractional Bayes Factor (O'Hagan 1995) are more appropriate (this is beyond the scope of this course, but you can refer to the paper by O'Hagan: "Fractional Bayes factors for model comparison", JRSS, 1995, 56, 99-118).

## 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

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

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

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

where  $\theta$  are the model parameters.

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

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

- Therefore,

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

## Akaike's information criterion

- As mentioned above, when a good loss function is not readily available, using a simple 0-1 loss function (i.e., 0 if we identify the correct model, and 1 if we fail to identify the correct model), simplifies the decision rule such that  $M_1$  is accepted over its corresponding alternative  $M_0$  if  $M_1$  has a higher posterior probability compared to  $M_0$ .
- In this case, higher posterior probability correspond to lower posterior risk, and therefore, our decision rule is consistent with the *expected loss* principle: “in deciding between different rules, choose the one with the smallest posterior risk”.
- It turns out (as discussed in Appendix B in Gelman et. al., 2002), the model with the highest posterior probability would have the lowest KL (Kullback-Leibler) information, and as the result the lowest expected deviance.

## Akaike's information criterion

- Akaike defined the Akaike's information criterion (AIC) based on KL divergence between the true probability model,  $Q(y)$ , and the assumed probability model,  $P(y|\theta)$ , which depends on parameters  $\theta$ ,

$$\begin{aligned} D(Q||P) &= \int Q(y) \log \frac{Q(y)}{P(y|\theta)} dy \\ &= \int \log[Q(y)] Q(y) dy - \int \log[P(y|\theta)] Q(y) dy \end{aligned}$$

## Akaike's information criterion

- Note that only the second term involves our model,  $P(x|\theta)$ . Therefore, it is the only part that should be considered in model selection.
- Also, note that this part is the expected log-likelihood with respect to the true model.
- Therefore, we want to increase the expectation of log-likelihood,  $\log[P(y|\theta)]$ . In other words, we would like to reduce the expectation of deviance,  $-2\log[P(y|\theta)]$ .

## Akaike's information criterion

- In general, we do not know the true model, so we do not know the expected log-likelihood or deviance.
- We can use the maximized log-likelihood,  $L(y, \hat{\theta})$ , given the observed data  $y$ , as our estimate.
- However, as shown by Akaike, this estimate is biased (optimistic).
- Akaike also showed the amount of biased is,  $k$ , the number of model parameters. Therefore, he proposed the following measurement for model comparison:

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

- The second term in the above estimate can be considered as a penalty for model complexity.
- Among different models, we choose the one with the lowest AIC.