STA 314: Statistical Methods for Machine Learning I

Lecture 3 - Model selection and cross-validation

Xin Bing

Department of Statistical Sciences
University of Toronto

Review

- We have learned the bias-variance-tradeoff:
  - As the complexity of the fitted model increases, its bias decreases while its variance increases.
  - The variance of any fitted model is roughly proportional to

$$\frac{\text{complexity of } \hat{f}}{n}.$$

  - When the sample size (n) is limited, a fitted model with high complexity performs poorly due to large variance.
  - When n is large enough, a more complex fitted model tends to peform better as they have smaller bias than simpler models.

Review

- We have learned the OLS approach:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\alpha}\|_2^2.$$

- We have learned the statistical properties of $\hat{\boldsymbol{\beta}}$ under the linear model

$$Y = \beta_0 + X_1\beta_1 + \cdots + X_p\beta_p + \epsilon.$$

  - Unbiasedness
  - Estimation error (ℓ_2)
  - Inference (confidence intervals, hypothesis testing).

Why consider alternatives to the OLS estimator?

Alternative fitting procedures to OLS could yield **better prediction accuracy** and **model interpretability**.

- Prediction / Estimation: the OLS estimator

  $$\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

  has large variance when p is large. Especially, if $p > n$, then OLS estimator is not unique and its variance is infinite.

- Interpretability: By removing irrelevant features – that is, by setting some coefficient estimates to zero – we can obtain a model that is more parsimonious hence more interpretable.

What are the alternatives?

- **Subset Selection**. We identify a subset of the p predictors that we believe to be related to the response. We then fit a model using the OLS approach on the identified set of predictors.
- **Shrinkage**. We fit a model involving all p predictors, but the estimated coefficients are shrunken towards zero relative to the OLS estimator. This shrinkage (also known as regularization) has the effect of reducing variance. Some could also perform variable selection.
- **Dimension Reduction**. We project the p predictors into a M-dimensional subspace, where $M < p$. This is achieved by computing M different linear combinations, or projections, of the original predictors. Then the resulting M projections are used as new predictors to fit a linear regression model by OLS.

How to choose the optimal one among a set of models?

Example

$$\begin{aligned} \text{Model 1:} \qquad & Y = \alpha_0 + \alpha_1 X_1 + \epsilon \\ \text{Model 2:} \qquad & Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon \end{aligned}$$

lead to different predictors at $X = x = (x_1, x_2)$

$$\hat{f}_1(x) = \hat{\alpha}_0 + \hat{\alpha}_1 x_1 \qquad \text{v.s.} \qquad \hat{f}_2(x) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2.$$

Ideally, we choose the one that has a **smaller expected MSE**.

How to compare expected MSEs?

- When we have $\mathcal{D}_{test}$, we compare the test MSE errors directly

$$\frac{1}{m}\sum_{i=1}^{m}\left(y_i^{(T)} - \hat{\alpha}_0 - \hat{\alpha}_1 x_{i1}^{(T)}\right)^2$$

$$\text{v.s.} \quad \frac{1}{m}\sum_{i=1}^{m}\left(y_i^{(T)} - \hat{\beta}_0 - \hat{\beta}_1 x_{i1}^{(T)} - \hat{\beta}_2 x_{i2}^{(T)}\right)^2$$

- What if we don't have $\mathcal{D}_{test}$?

Model selection

There are two common approaches for model selection when we don't have $\mathcal{D}_{test}$:

- We can avoid estimating the expected MSE by making an adjustment to the training error to account for the model complexity:
  - Mallow's C_p
  - AIC
  - BIC
  - adjusted R^2

- We can directly estimate the expected MSE by manually creating a "test set" using data-splitting techniques:
  - validation set approach
  - cross-validation approach

Avoid estimating the expected MSE: C_p, AIC, BIC, and adjusted R^2

- These techniques adjust the training error for the model complexity.
- They are only used to select among a set of parametric models with different numbers of predictors.

Cont'd: Avoid estimating the expected MSE

For any given fitted model $\hat{f}$, let $\hat{f}(x_i)$ be the fitted value for the i the observation. For instance, for a fitted linear model with p predictors,

$$\hat{f}(x_i) = x_i^\top \hat{\boldsymbol{\beta}} = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip}.$$

Recall that

$$e_i = y_i - \hat{f}(x_i)$$

is the ith residual. The residual sum of squares (RSS) is defined as

$$RSS(\hat{f}) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - \hat{f}(x_i) \right)^2 .$$

Yes, this is indeed the training MSE of $\hat{f}$! It gets smaller as p increases.

Mallow's C_p

Let p be the total # of parameters in the model

$$C_p(\hat{f}) = \frac{1}{n}\left(RSS(\hat{f}) + 2p\sigma^2\right).$$

When σ^2 is unknown, one use a consistent estimator $\hat{\sigma}^2$.

- Essentially, the C_p adds a penalty $2p\hat{\sigma}^2$ to the training MSE to adjust for the fact that the training error is always in favor of more complex models.
- C_p tends to take on a small value for models with a low test error, so when determining which of a set of models is best, we choose the model with the lowest C_p value.
- C_p is mainly for (linear) fitted models (such as via OLS) in regression problems

AIC and BIC

Let $\hat{f}$ be the fitted model obtained from the MLE approach.
Let $L(\hat{f})$ be the maximized value of the likelihood function for $\hat{f}$.

- **AIC**:

  $$AIC(\hat{f}) = -2\log L(\hat{f}) + 2p,$$

  In the linear model with $\epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$, $AIC(\hat{f})$ is proportional to $C_p(\hat{f})$, selecting the same model.

- **BIC**:

  $$BIC(\hat{f}) = -2\log L(\hat{f}) + (\log n)p,$$

  BIC places a heavier penalty $(\log n)p$ on models with many predictors, and hence results selecting smaller-size models than AIC and C_p.

AIC and BIC

- For both AIC and BIC, we select the best model that has the lowest value.

- To compute AIC and BIC, we need to specify the likelihood, i.e. the distribution of $Y \mid X$, and to compute the maximum likelihood estimator.

- AIC and BIC can also be used for selecting parametric models in classification problems.

Adjusted R^2

Recall that the total sum of squares (TSS) is defined as

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2, \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

Recall that

$$R^2(\hat{f}) = \frac{TSS - RSS(\hat{f})}{TSS} = 1 - \frac{RSS(\hat{f})}{TSS}.$$

By contrast,

$$Adjusted\ R^2(\hat{f}) = 1 - \frac{RSS(\hat{f})/(n-p-1)}{TSS/(n-1)}.$$

Remark. Unlike the R^2 statistic, the adjusted R^2 statistic pays a price for the inclusion of unnecessary variables in the model.

Adjusted R^2 vs R^2

$$Adjusted\ R^2(\hat{f}) = 1 - \frac{RSS(\hat{f})/(n-p-1)}{TSS/(n-1)}.$$

- Maximizing the adjusted R^2 is equivalent to minimizing $RSS/(n-p-1)$. While RSS always decreases as the number of variables in the model increases, $RSS/(n-p-1)$ may increase or decrease, due to the presence of p in the denominator.
- Unlike C_p, AIC, and BIC, for which a **smaller** value indicates a model with lower test error, a **larger** value of adjusted R^2 indicates a model with smaller test error.

Example (Revisited)

$$\begin{aligned} \text{Model 1:} \qquad & Y = \alpha_0 + \alpha_1 X_1 + \epsilon \\ \text{Model 2:} \qquad & Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon \end{aligned}$$

lead to different predictors at $X = x = (x_1, x_2)$

$$\hat{f}_1(x) = \hat{\alpha}_0 + \hat{\alpha}_1 x_1 \qquad \text{v.s.} \qquad \hat{f}_2(x) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2.$$

Their residual sum of squares (RSS) can be computed as:

$$RSS(\hat{f}_1) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{\alpha}_0 - \hat{\alpha}_1 x_{i1})^2$$

$$RSS(\hat{f}_2) = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2}\right)^2$$

Mallow's C_p, AIC, BIC and adjusted R^2

- Mallow's C_p

$$
\begin{aligned}
C_p(\hat{f}_1) &= \frac{1}{n}\left(RSS(\hat{f}_1) + (2 \times 1)\sigma^2\right) \\
C_p(\hat{f}_2) &= \frac{1}{n}\left(RSS(\hat{f}_2) + (2 \times 2)\sigma^2\right).
\end{aligned}
$$

- Adjusted R^2

$$
\begin{aligned}
Adjusted\ R^2(\hat{f}_1) &= 1 - \frac{RSS(\hat{f}_1)/(n-2)}{TSS/(n-1)} \\
Adjusted\ R^2(\hat{f}_2) &= 1 - \frac{RSS(\hat{f}_2)/(n-3)}{TSS/(n-1)}.
\end{aligned}
$$

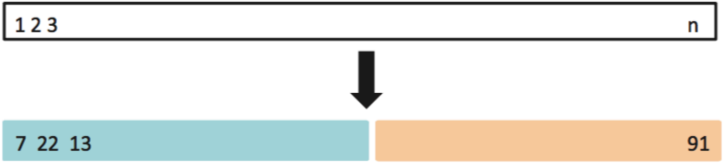
Direct estimation of the expected MSE: data-splitting techniques

We *randomly* split the available data to create a validation set that functions as a test set.

- Validation set approach: one-time data splitting
- Cross-validation approach: multiple-time data splitting

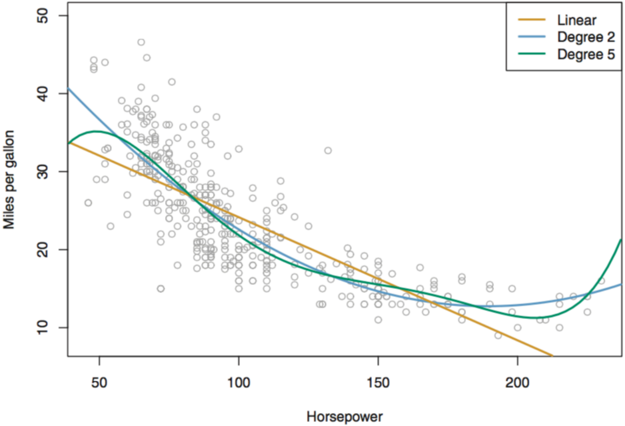
Validation set approach

- *Randomly* divide the available set of samples into two parts: a **training set** and a **validation** (hold-out) set.
    - What is the proportion? Depends.
- The model is fitted on the training set, and the fitted model is evaluated by the validation set.
- The resulting validation-set MSE provides an estimate of the expected MSE.



Example: Auto Data

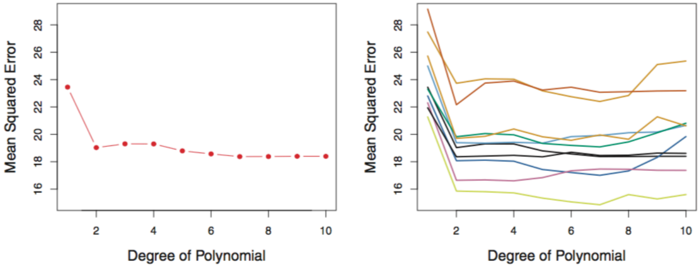
In Lecture 2, we find there appears to be a non-linear relationship between **mpg** and **horsepower**.



Whether a cubic or higher-order predictor provides a better fit?

Example: Auto Data – Compare linear vs higher-order polynomial terms in a linear regression.

We randomly split the 392 observations into two sets, a training set containing 196 of the data points, and a validation set containing the remaining 196 data.



- Left: Validation error estimates for a single split into training and validation data sets.
- Right: Validation method repeated 10 times with each time using a different random split of the observations into a training set and a validation set.
- We can see the one-time data splitting is not stable

Drawbacks of Validation Set Approach

- The validation estimate of the test error can be highly unstable, depending on which observations are included in the training set and which are in the validation set.
- Only a subset of the observations – those in the training set rather than in the validation set – are used to fit the model.
  The resulting estimate or classifier is worse!
- How to remedy these drawbacks?

Leave-One-Out Cross-Validation (LOOCV)

- First split the data into two parts by leaving out the **first** observation:
  - a validation set: (x_1, y_1)
  - a training set: the remaining observations $(x_2, y_2), ..., (x_n, y_n)$
  - using the training set, we fit the model $\hat{f}_1$ and predict y_1 as $\hat{f}_1(x_1)$ using the value x_1. The test error could be approximated by

$$MSE_1 = (y_1 - \hat{f}_1(x_1))^2.$$

  - not good enough!

Leave-One-Out Cross-Validation

- Repeat the procedure by leaving out the **second** observation:
  - a validation set: (x_2, y_2),
  - a training set: the remaining observations $(x_1, y_1), (x_3, y_3), ..., (x_n, y_n)$
  - using the training set, we fit the model $\hat{f}_2$ and predict y_2 as $\hat{f}_2(x_2)$ using the value x_2. Compute

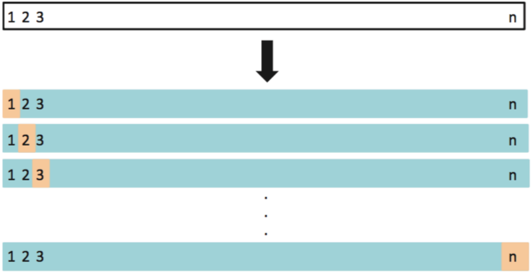
$$MSE_2 = (y_2 - \hat{f}_2(x_2))^2.$$

- Repeating the approach n times by leaving out **each** observation to obtain $MSE_1, ..., MSE_n$.

- The LOOCV estimate for the test MSE is the average of these n test error estimates:

$$CV_{(n)} = \frac{1}{n} \sum_{i=1}^{n} MSE_i.$$

Leave-One-Out Cross-Validation



Validation data sets in beige, and training sets in cyan.

LOOCV vs Validation Set Approach

LOOCV has the following advantage over the validation set approach.

- The training set of LOOCV is almost the same as the entire data set. The fitted model is almost as good as that based on the entire data set.
- The validation approach yields different results when applied repeatedly, because the training/validation set is randomly divided. LOOCV has no randomness in the splitting.

However, LOOCV can be computationally expensive. (In linear model, the computation can be simplified, the formula is shown in page 202 of the textbook).

k-Fold Cross-Validation

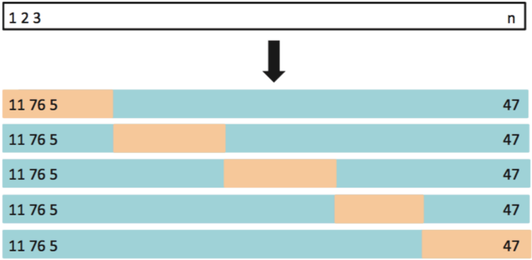
- **k-fold CV** is to randomly divide the data into k (roughly) equal-sized groups or folds.
- The first fold is treated as a validation set, and the method is fit on the remaining $k-1$ folds. We compute the mean squared error, MSE_1, for the observations in the first fold.
- Then we repeat the procedure to fold 2, fold 3,..., fold k, and get MSE_2, MSE_3,...,MSE_k.
- The k-fold CV estimate is computed by averaging these values,

$$CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k} MSE_i.$$

Remark

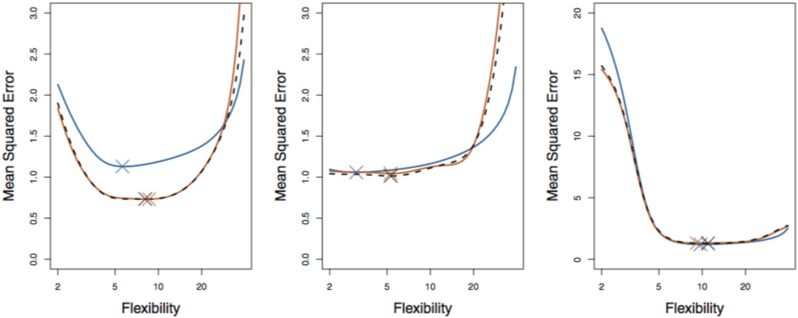
- LOOCV is a special case of n-fold CV.
- 5-fold or 10-fold is commonly used in practice.

k-Fold Cross-Validation



Validation data sets in beige, and training sets in blue.

k-Fold Cross-Validation



True test MSE (in blue), the LOOCV estimate (black dashed line), and the 10-fold CV estimate (in orange) for three simulated data sets.

Cross-Validation on Classification Problems

- Cross-validation also works for classification problems.
- For LOOCV, we split the data in the same way as before. We compute the error on the validation set

$$Err_1 = 1\left\{y_1 \neq \hat{f}_1(x_1)\right\}.$$

- Then we repeat the procedure n times, and get Err_2, Err_3,...,Err_n.
- The LOOCV estimate is computed by averaging these values,

$$CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n} Err_n.$$

Cross-Validation is sometimes tricky!

Independence between the fitted model and the validation set is the key! Meaning that you should **NOT** use the validation set to fit your model.

Example (A tricky one)

Consider a simple two-step approach applied to some data $\mathcal{D}^{train}$.

- Starting with 5000 predictors and 100 samples, find the 10 predictors having the largest correlation with the outcome.
- We then apply the OLS using only these 10 predictors.

How do we estimate the expected MSE of the fitted model from this approach?

Discussion / recommendation on these two approaches

- The data-splitting technique has two advantages relative to AIC, BIC, C_p, and adjusted R^2:
  - it provides a direct estimate of the test error
  - It can also be used in a wider range of model selection tasks, even in cases where it is hard to pinpoint the model degrees of freedom (e.g. the number of predictors in the model) or hard to estimate the error variance.

- The data-splitting technique also has a couple of drawbacks comparing to the other approach:
  - it requires a relatively large sample size
  - it is difficult to have guarantees for the model selected by using CV.
  - when the distribution is specified and the error of variance can be consistently estimated, the first approach is preferred.

What's next?

Apply these two techniques
for model selection under linear models.