NBA 4920/6921 Lecture 8

Hold-out Methods

Murat Unal

Johnson Graduate School of Management

9/23/21

Agenda

- Assignment 1 Survey
- Training vs testing
- Hold-out methods
  - Validation set approach
  - Leave-one-out cross validation
  - K-fold cross validation

Training error vs test error

Recall the distinction between training error and test error:

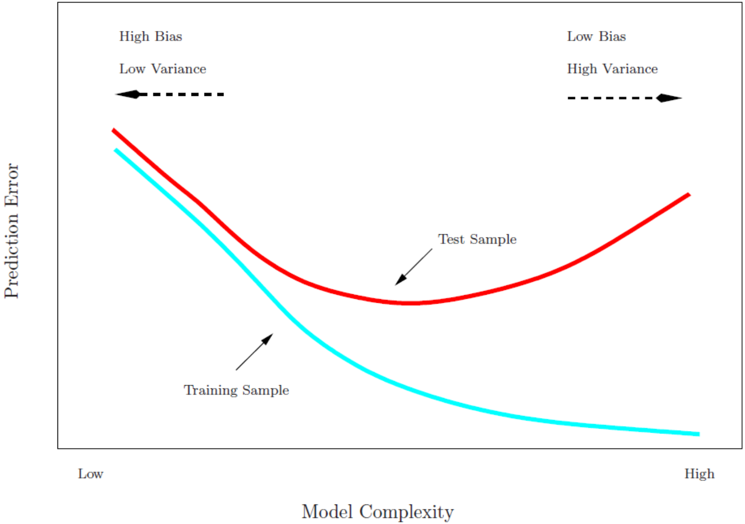
The training error is obtained from training the statistical learning method on the data we have at our disposal, i.e. training data.

It is the result of the training process.

The test error is the average error that results from applying the trained statistical learning method on unseen data, i.e test data.

Ultimately, we are interested in minimizing the test error, but using the training error as a proxy will underestimate the latter.

Training vs test-set performance



Source: ISL

Prediction error estimates

How can we obtain reliable estimates of the test error?

Best solution: we have a large data set and designate 20% as test set, and use it once.

What if this is not feasible?

What if we need to select and train a model?

How can we avoid overfitting our training data during model selection?

Hold-out methods

We can utilize hold-out methods, e.g. cross validation, and use training data to estimate test performance

The idea is to estimate the test error by holding out part of the data set from training, and then applying the trained method on these held out observations

Hold-out methods

This way we can achieve two things:

1. Assess model performance
2. Select the appropriate level of model flexibility

Hold-out methods

1. **The validation set approach**:

   - Hold out subset of the training data
   - **Validate** the trained method on this held out validation set
   - The model does not see the validation set
   - The **validation error** is an estimate of the test error

Validation set approach

Source: Ed Rubin

Validation set approach

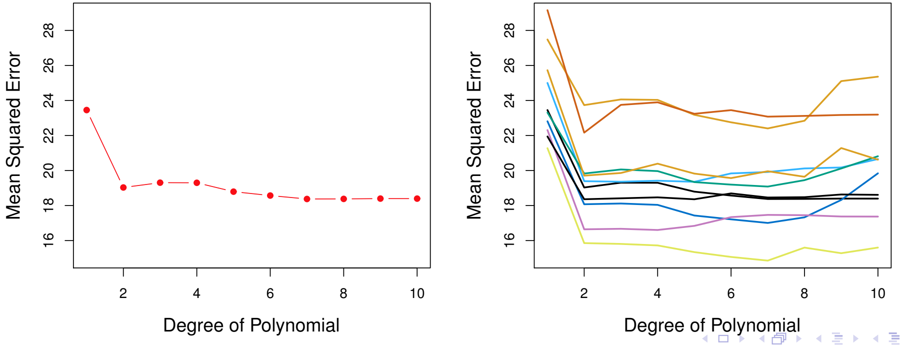
Source: Ed Rubin

Validation set approach

Source: Ed Rubin

Validation set approach

We can apply the approach to compare the performance of linear vs higher-order polynomial terms in regression
Using ten different random splits of the data into training and validation sets, and apply the approach ten times results in ten different curves



Source: ISL

Validation set approach

Drawbacks of the approach:

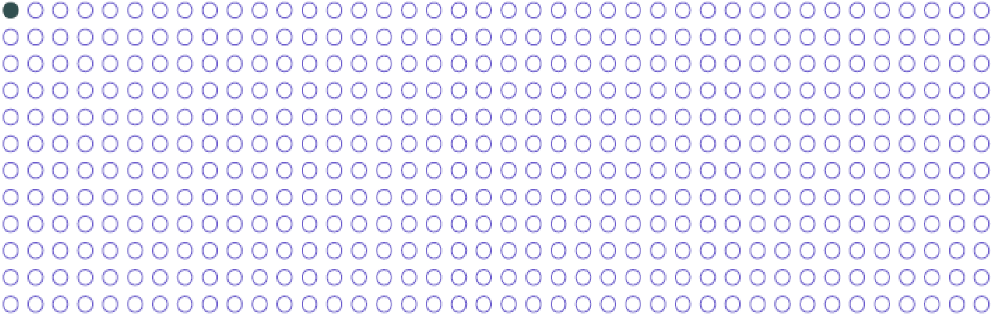
1. High variability in estimating the test error
2. Inefficiency in training due to not including the validation set.
3. Statistical methods tend to perform worse when trained on fewer observations, this can overestimate the test error

2. Leave-one-out cross validation:

- Cross validations uses all of the data for training and therefore remedies the drawbacks of the validation set approach.
- In LOOCV each observation takes a turn as the validation set
- The validation set now is exactly one observation.
- All other observations get to train the model.
- Repeat the validation exercise for every observation.

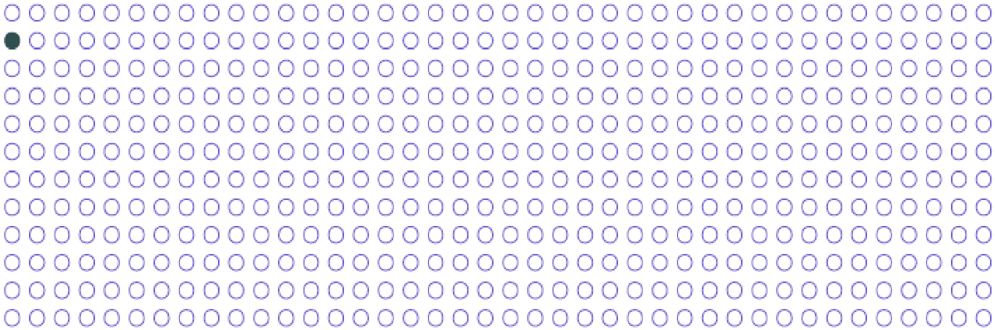
$$CV_n = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Leave-one-out cross validation



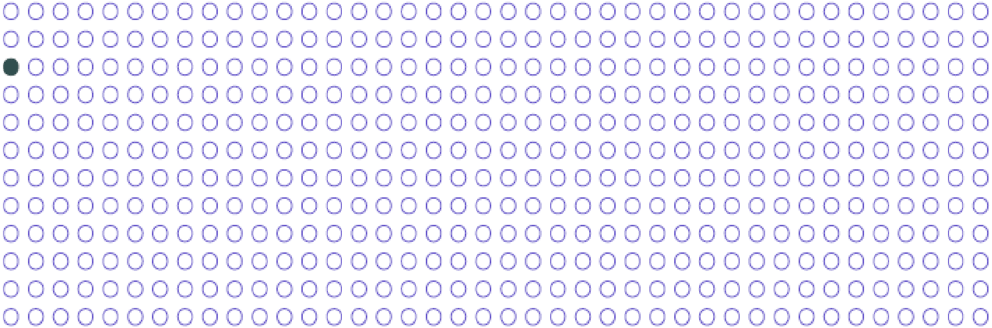
Source: Ed Rubin

Leave-one-out cross validation



Source: Ed Rubin

Leave-one-out cross validation



Source: Ed Rubin

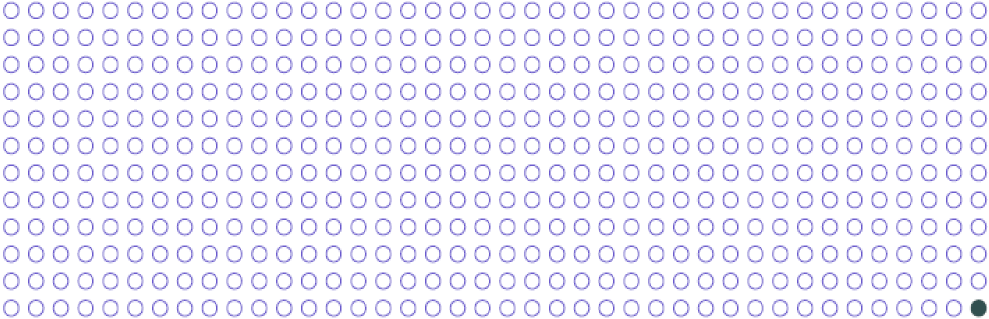
Leave-one-out cross validation

Source: Ed Rubin

Leave-one-out cross validation

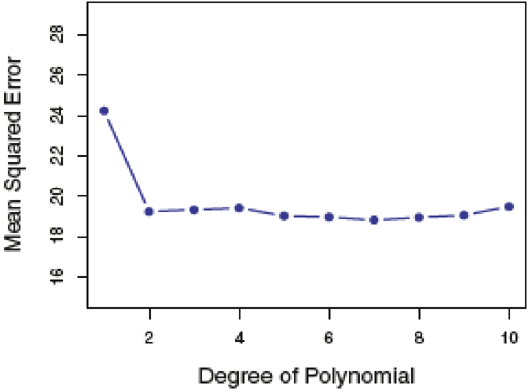
Source: Ed Rubin

Leave-one-out cross validation



Source: Ed Rubin

Leave-one-out cross validation



Source: ISL

Leave-one-out cross validation

Benefits of the approach:

1. Reduces bias by using almost all observations for training.
2. Resolves dependency on a specific validation set.
3. Removes variation. Performing LOOCV multiple times always yield the same results

Drawbacks of the approach:

1. Since the model has to be fit n times, it is expensive to implement.

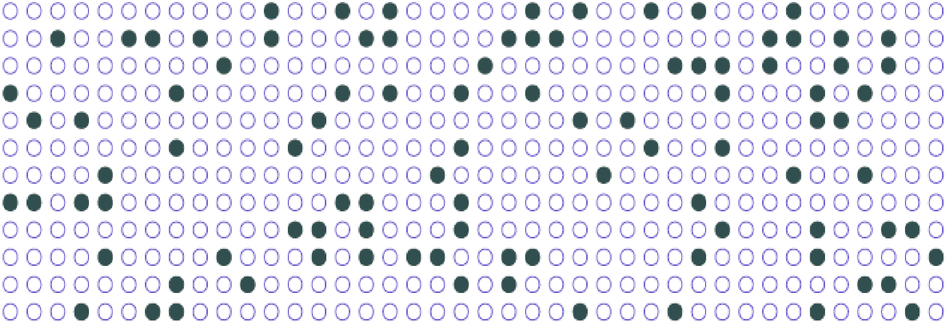
K-fold cross validations

2.K**-fold cross validation**:

1. Divide the data into K equal-sized parts.i.e. folds
2. Leave out fold k and fit the model on the remaining K-1 folds
3. Do this repeatedly for each fold $k = 1, \cdots, K$
4. Obtain test error estimate by averaging the folds' errors

$$CV_k = \frac{1}{k}\sum_{k=1}^{K} MSE_k$$

K-fold cross validation



Source: Ed Rubin

K-fold cross validation

For $K = 5$, start with fold 1 as the validation set, obtain $MSE_{k=1}$

Source: Ed Rubin

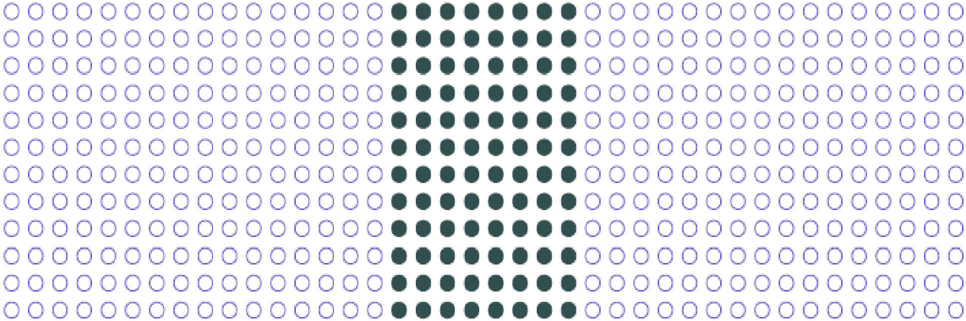
K-fold cross validation

Fold 2 is the validation set, obtain $MSE_{k=2}$

Source: Ed Rubin

K-fold cross validation

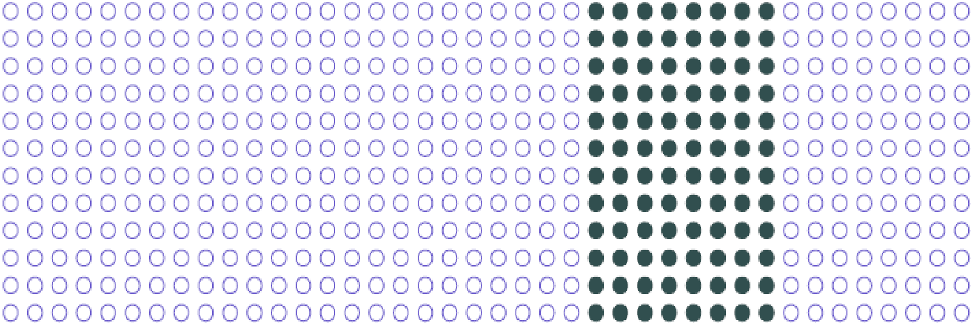
Fold 3 is the validation set, obtain $MSE_{k=3}$



Source: Ed Rubin

K-fold cross validation

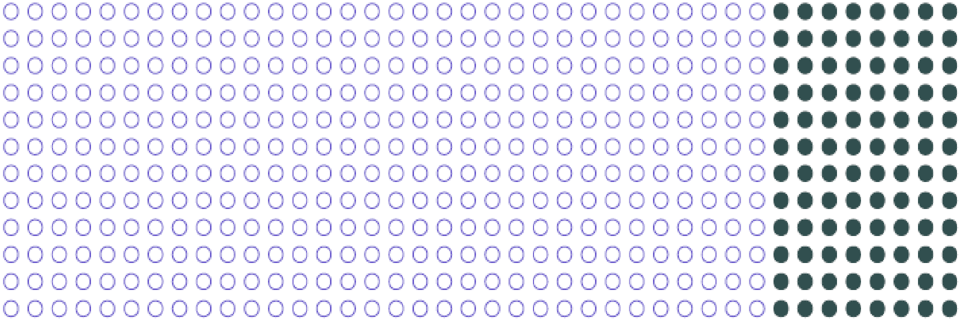
Fold 4 is the validation set, obtain $MSE_{k=4}$



Source: Ed Rubin

K-fold cross validation

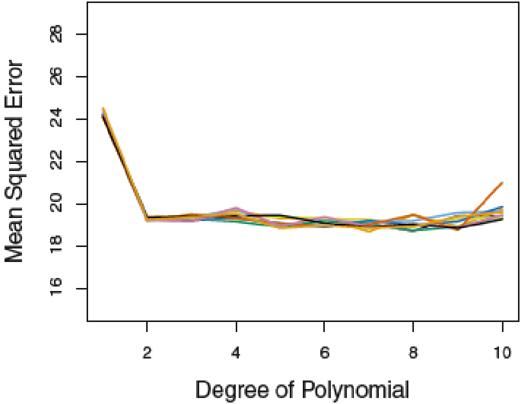
Fold 5 is the validation set, obtain $MSE_{k=5}$



Source: Ed Rubin

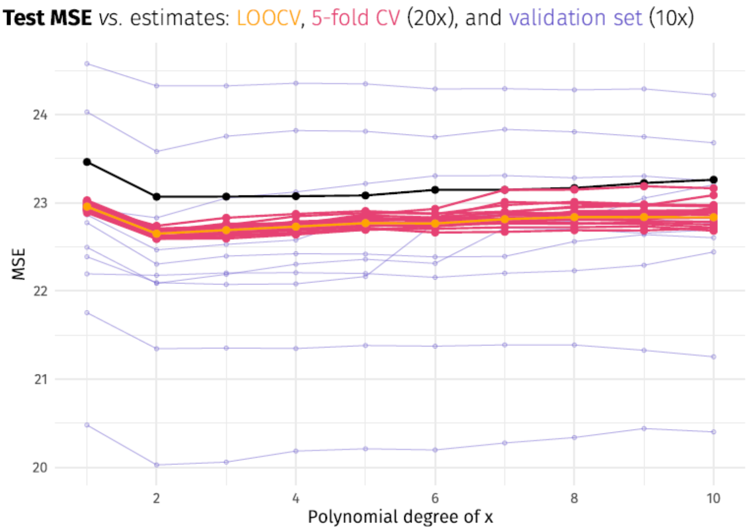
K-fold cross validation

10-fold CV was run nine separate times, each with a different random split of the data into ten parts. 10-fold CV reduces the variability in the test error estimates significantly.



Source: ISL

Comparing hold-out methods



Source: Ed Rubin

Cross validation on classification problems

Cross-validation works just as described, except that rather than using MSE to quantify test error, we instead use the number of misclassified observations.

The LOOCV error rate takes the form

$$CV_n = \frac{1}{n} \sum_{i=1}^{n} 1(y_i \neq \hat{y}_i)$$

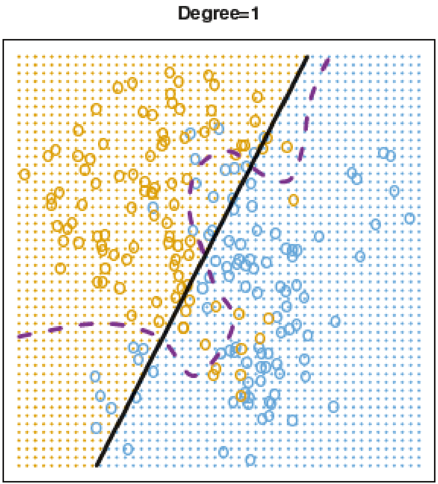
Cross validation on classification problems

Let's fit a logistic regression model to a classification problem

Just like in linear regression, we can fit polynomials if the decision boundary looks non-linear

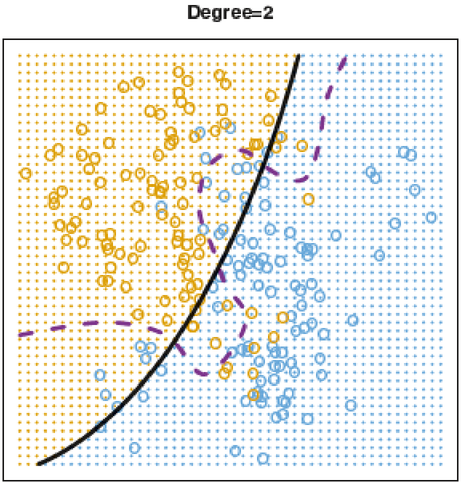
$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_2 + \beta_4 X_2^2$$

Cross validation on classification problems



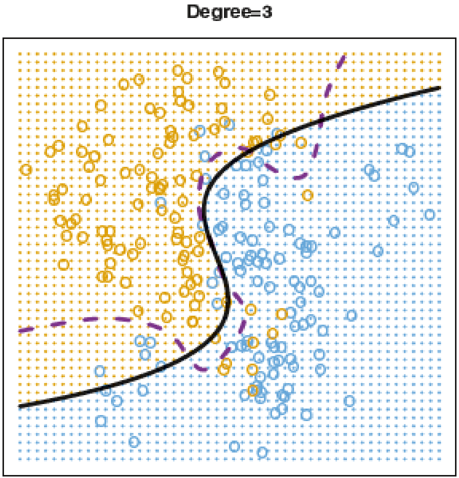
Source: ISL

Cross validation on classification problems



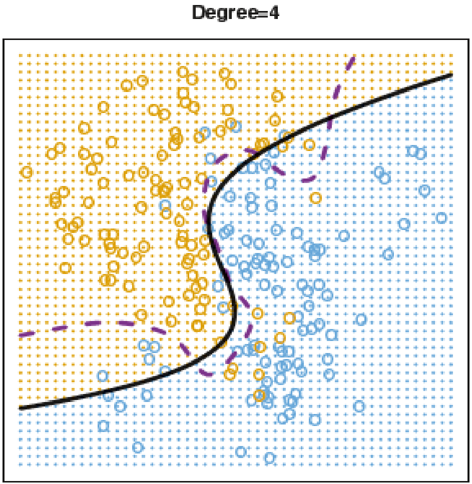
Source: ISL

Cross validation on classification problems



Source: ISL

Cross validation on classification problems



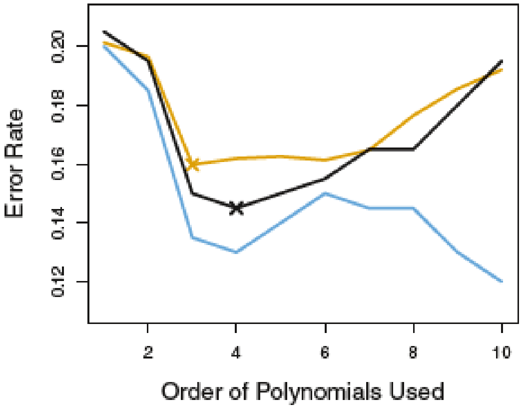
Source: ISL

Cross validation on classification problems

How might we decide between the four logistic regression models displayed in the previous figures?

We can use cross validation in order to make this decision

Cross validation on classification problems



Source: ISL

Test error (brown), training error (blue), and 10-fold CV error (black) on the two-dimensional classification data

Recap

To find the best performing statistical method in our context we need to estimate the test error

We apply hold-out methods to estimate the test error

Divide the data into subsets/folds: training and validation sets

Train the model on the training data and validate/evaluate it on the held-out validation set

References

Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani (2017)
An Introduction to Statistical Learning
Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani (2017).
https://www.statlearning.com/

Ed Rubin (2020)
Economics 524 (424): Prediction and Machine-Learning in Econometrics
Univ, of Oregon.