Statistical Learning for Data Science

Lecture 16

唐晓颖

电子与电气工程系

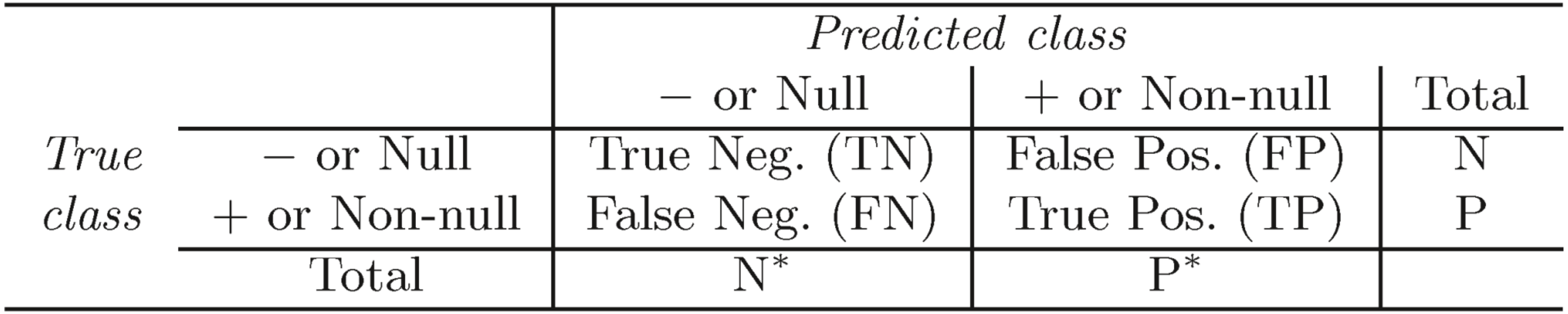
南方科技大学

May 8, 2023

Linear Discriminant Analysis (LDA)

- Measures for classification

Possible results when applying a classifier or diagnostic test to a population

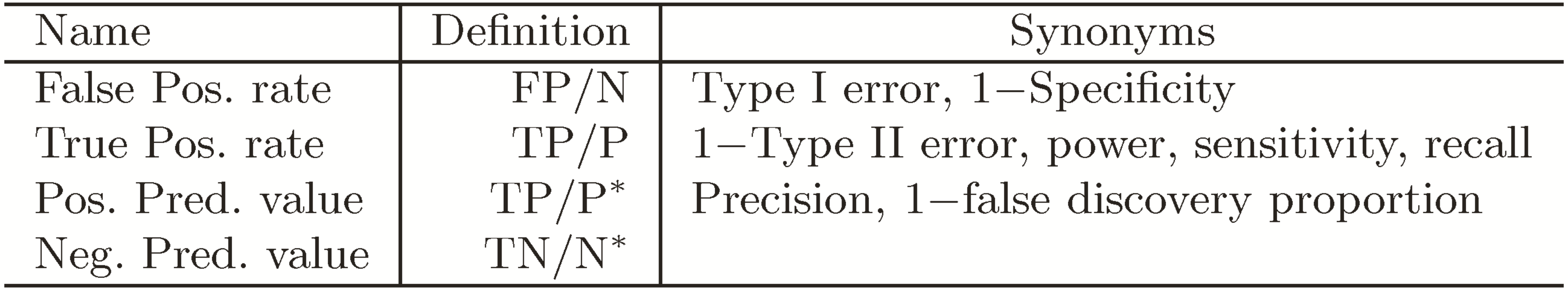


		Predicted class		
		− or Null	+ or Non-null	Total
True class	− or Null	True Neg. (TN)	False Pos. (FP)	N
	+ or Non-null	False Neg. (FN)	True Pos. (TP)	P
	Total	N*	P*	

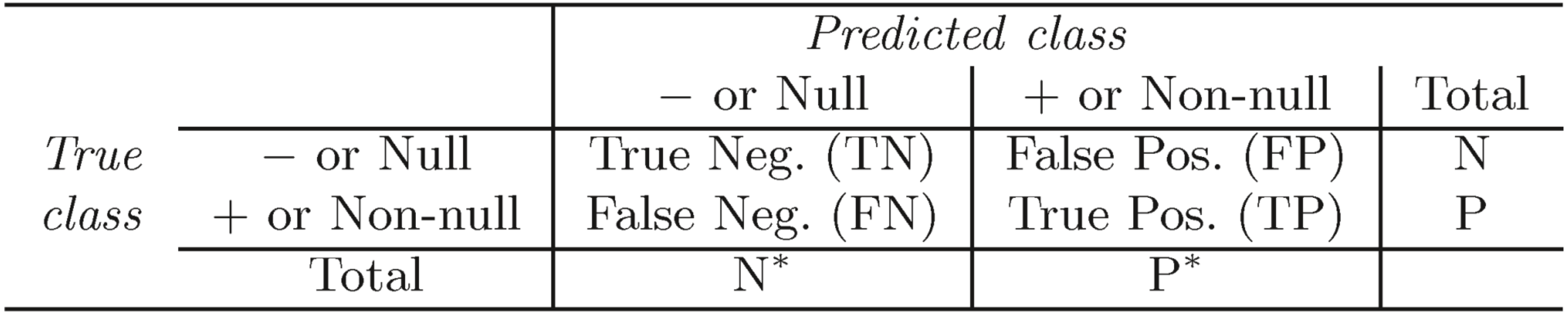
Linear Discriminant Analysis (LDA)

- Measures for classification

Important measures for classification and diagnostic testing.



Name	Definition	Synonyms
False Pos. rate	FP/N	Type I error, 1−Specificity
True Pos. rate	TP/P	1−Type II error, power, sensitivity, recall
Pos. Pred. value	TP/P^*	Precision, 1−false discovery proportion
Neg. Pred. value	TN/N^*	



		Predicted class		
		− or Null	+ or Non-null	Total
True class	− or Null	True Neg. (TN)	False Pos. (FP)	N
	+ or Non-null	False Neg. (FN)	True Pos. (TP)	P
	Total	N^*	P^*	

Linear Discriminant Analysis (LDA)

- Adjusting the LDA threshold

The classification of a new observation is based on its linear discriminant score, which is the weighted sum of its predictor variables. The classification threshold for LDA is typically set to 0.5

In some cases, adjusting the LDA threshold can improve the classification performance.

The threshold can be adjusted by changing the cut-off value used to assign observations to the different groups.

Useful when the costs of misclassification of different groups are different or when the classes are imbalanced.

Linear Discriminant Analysis (LDA)

- Adjusting the LDA threshold

Example: A credit card company might particularly wish to avoid incorrectly classifying an individual who will default, whereas incorrectly classifying an individual who will not default, though still to be avoided, is less problematic.

It is possible to modify LDA in order to develop a classifier that better meets the credit card company's needs.

Originally

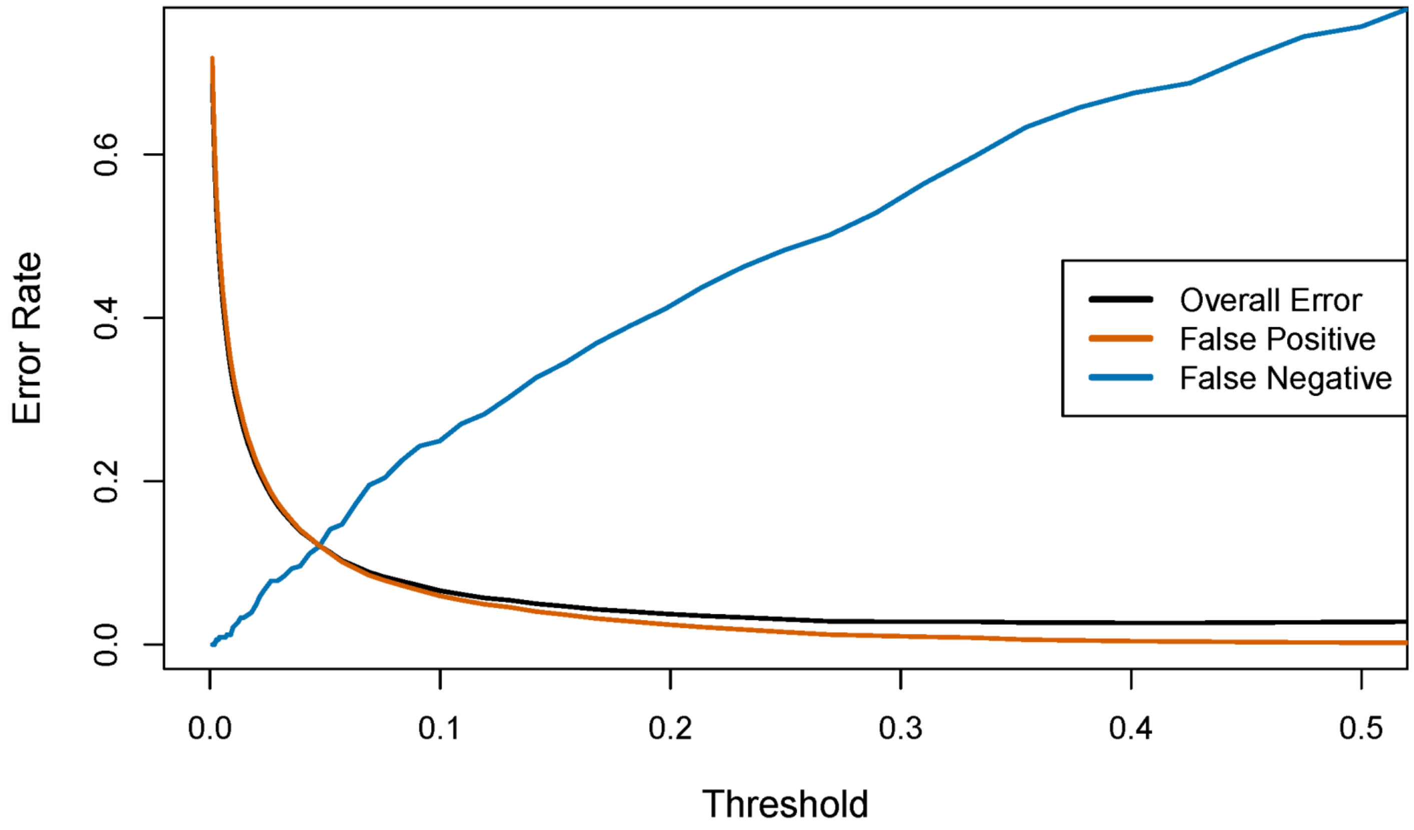
$$\text{default=Yes, if } \Pr(\text{default=Yes} \mid X = x) > 0.5$$

Modified

$$\text{default=Yes, if } \Pr(\text{default=Yes} \mid X = x) > 0.2$$

Linear Discriminant Analysis (LDA)

- Adjusting the LDA threshold



In order to reduce the false negative rate, we may want to reduce the threshold to 0.1 or less.

Linear Discriminant Analysis (LDA)

- The Receive Operating Characteristics (ROC) curve

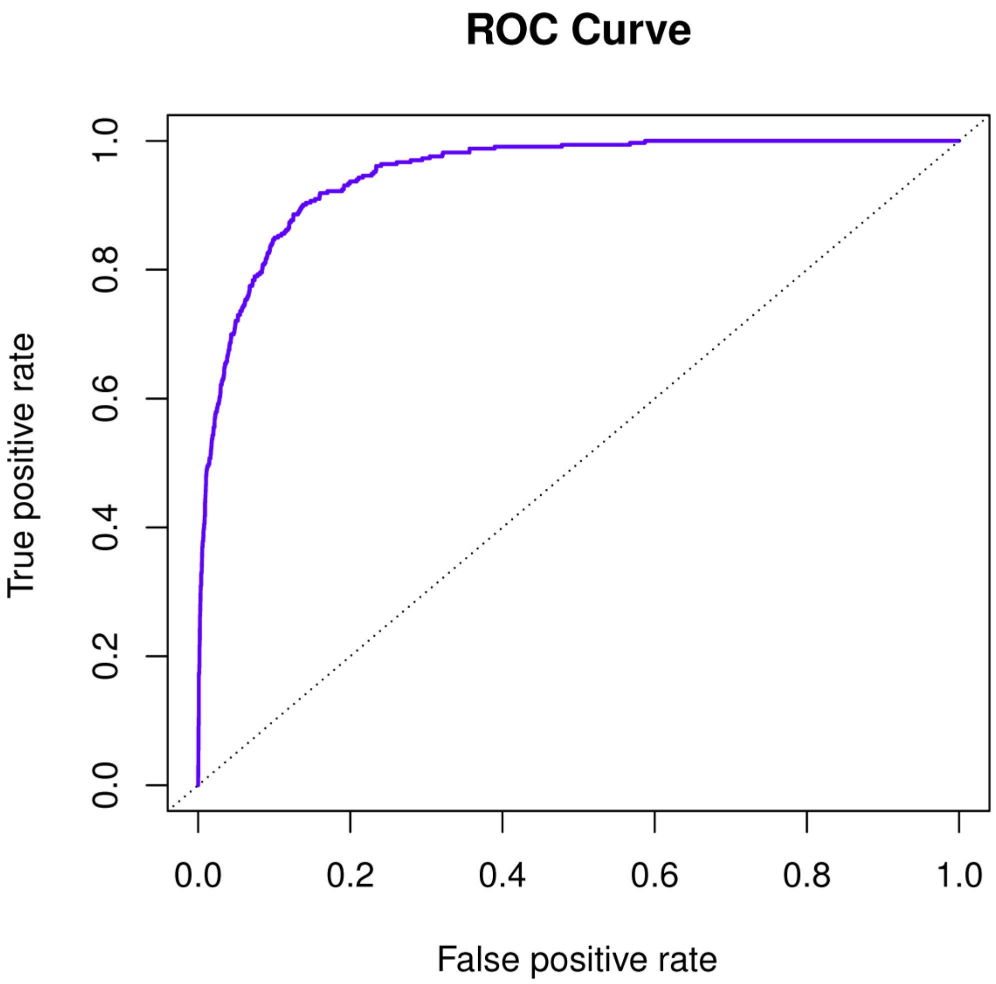
  - The ROC curve is a popular graph for simultaneously displaying the two types of errors for all possible thresholds.

  - The overall performance of a classifier, summarized over all possible thresholds, is given by the area under the curve (AUC). An ideal ROC curve will hug the top left corner, so the larger the AUC the better the classifier.

  - ROC curves are useful for comparing different classifiers, since they take into account all possible thresholds.

Linear Discriminant Analysis (LDA)

- The Receive Operating Characteristics (ROC) curve



AUC = 0.95

The dotted line represents the "no information" classifier; this is what we would expect if student status and credit card balance are not associated with probability of default.

Linear Discriminant Analysis (LDA)

- Quadratic Discriminant Analysis

QDA assumes that each class has its own covariance matrix! That is, it assumes that an observation from the *k*th class is of the form

$$X \sim N(\mu_k, \Sigma_k)$$

$$f(x) = \frac{1}{(2\pi)^{p/2} \left|\Sigma\right|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu) \right)$$

Σ_k : a covariance matrix for the *k*th class

$$\delta_k(x) = -\frac{1}{2}\log\left|\Sigma_k\right| - \frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k) + \log \pi_k$$

$$= \underbrace{-\frac{1}{2}x^T \Sigma_k^{-1} x + x^T \Sigma_k^{-1} \mu_k}_{\text{A quadratic function of x. This is where QDA gets its name.}} - \frac{1}{2}\mu_k{}^T \Sigma_k^{-1} \mu_k - \frac{1}{2}\log\left|\Sigma_k\right| + \log \pi_k$$

Linear Discriminant Analysis (LDA)

- Quadratic Discriminant Analysis

LDA or QDA?

LDA is a *much less flexible* classifier than QDA (there are much more parameters to estimate in QDA than LDA), and so has substantially lower variance. However, if LDA’s assumption that the K classes share a common covariance matrix is badly off, then LDA can suffer from high bias.

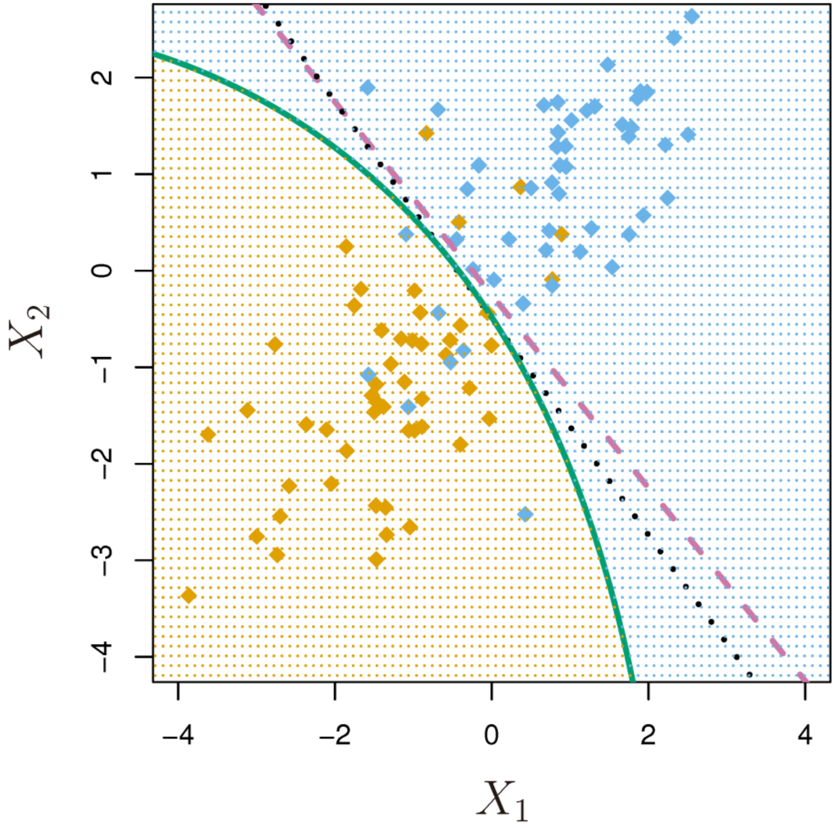
Roughly speaking, LDA tends to be a better bet than QDA if there are *relatively few training observations* and so reducing variance is crucial. In contrast, QDA is recommended if the training set is very large, so that the variance of the classifier is not a major concern, or if the assumption of a common covariance matrix for the K classes is clearly untenable.

Linear Discriminant Analysis (LDA)

- Quadratic Discriminant Analysis

LDA or QDA?

$$\Sigma_1 = \Sigma_2$$



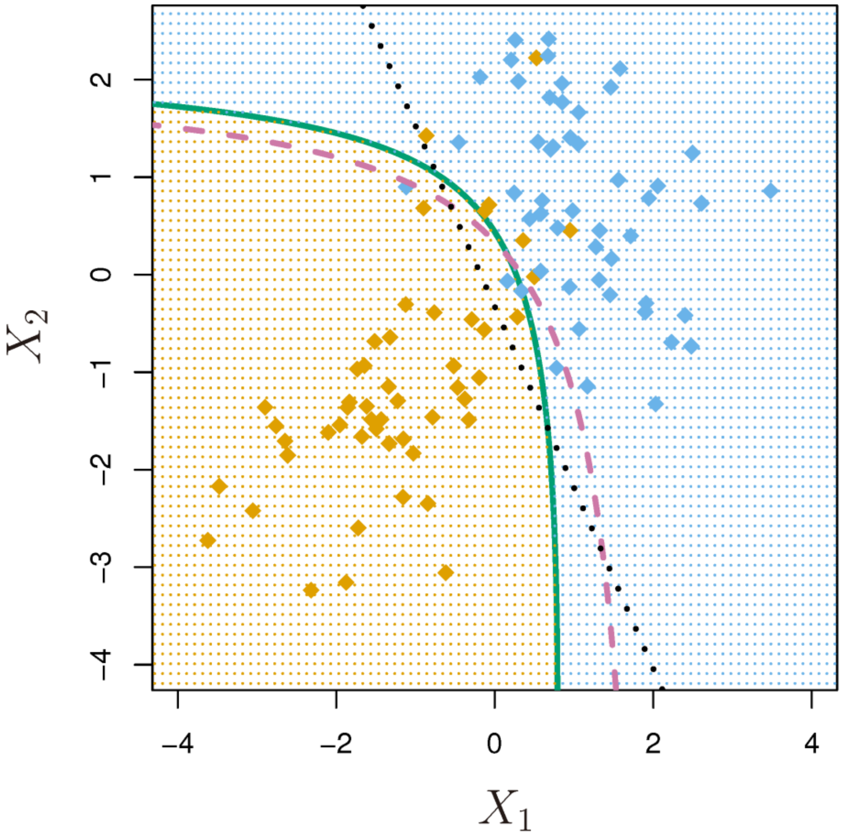
Purple dashed – The Bayes
Black dotted – LDA
Green solid – QDA

Linear Discriminant Analysis (LDA)

- Quadratic Discriminant Analysis

LDA or QDA?

$\Sigma_1 \neq \Sigma_2$



Purple dashed – The Bayes
Black dotted – LDA
Green solid – QDA

Linear Discriminant Analysis (LDA)

- Naïve Bayes

Naïve Bayes assumes conditional independence among predictors within each class.

It makes the "naive" assumption that all features in the input data are independent of each other.

$$
\begin{aligned}
f_k(x) &= \Pr(X = x \mid Y = k) \\
&= \prod_{j=1}^{p} \Pr(X_j = x_j \mid Y = k) \\
&= \prod_{j=1}^{p} f_{jk}(x_j)
\end{aligned}
$$

A Comparison of Classification Methods

- Naïve Bayes

For Gaussian Naïve Bayes (**each covariance matrix is diagonal**)

$$\delta_k(x) \propto \log\left(\pi_k f_k(x)\right) = -\frac{1}{2}\sum_{j=1}^{p}\frac{(x_j-\mu_{kj})^2}{\sigma_{kj}^2} + \log \pi_k$$

$$\begin{aligned} f_k(x) &= \Pr(X = x \mid Y = k) \\ &= \prod_{j=1}^{p} \Pr(X_j = x_j \mid Y = k) \\ &= \prod_{j=1}^{p} f_{jk}(x_j) \end{aligned}$$

Despite strong assumptions, Naïve Bayes often produces very good classification results.

A Comparison of Classification Methods

- Logistic regression versus LDA

For a two-class setting with p=1 predictor

In LDA

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_k)^2\right)}{\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_l)^2\right)}$$

$$\log\left(\frac{p_1(x)}{p_2(x)}\right) = \log\left(\frac{p_1(x)}{1-p_1(x)}\right)$$

$$= \log \pi_1 - \frac{1}{2\sigma^2}(x-\mu_1)^2 - \log \pi_2 + \frac{1}{2\sigma^2}(x-\mu_2)^2$$

$$= \frac{\mu_1 - \mu_2}{\sigma^2} x - \frac{\mu_1^2 - \mu_2^2}{2\sigma^2} + \log \frac{\pi_1}{\pi_2}$$ ⟶ same form as logistic regression

A Comparison of Classification Methods

- Logistic regression versus LDA

Both logistic regression and LDA produce linear decision boundaries. The only difference between the two approaches lies in the fact that the parameters in logistic regression are estimated using maximum likelihood, whereas those in LDA are computed using the estimated mean and variance from a normal distribution.

Logistic regression can also fit quadratic decision boundaries like QDA, by explicitly including quadratic terms in the model.

A Comparison of Classification Methods

- KNN versus QDA versus Logistic regression/LDA

KNN is a completely non-parametric approach: no assumptions are made about the shape of the decision boundary. We can expect KNN to dominate LDA and logistic regression when the decision boundary is highly non-linear.

KNN does not tell us which predictors are important.

QDA serves as a compromise between the non-parametric KNN and the linear LDA/logistic regression.

Summary

- Logistic regression is very popular for classification, especially when K=2.

- LDA is useful when n is small, or the classes are well separated, and the Gaussian assumptions are reasonable. Also, when K>2.

- Naïve Bayes is useful when p is very large.

Resampling Methods

- Involve repeatedly drawing samples from a training set and refitting a model of interest on each sample in order to obtain additional information about the fitted model.

  E.g., *To estimate the variability of a linear regression fit, we can repeatedly draw different samples from the training data, fit a linear regression to each sample, and then examine the extent to which the resulting fits differ.*

- Computational expensive, because they involve fitting the same statistical method multiple times using different subsets of the training data.

Resampling Methods

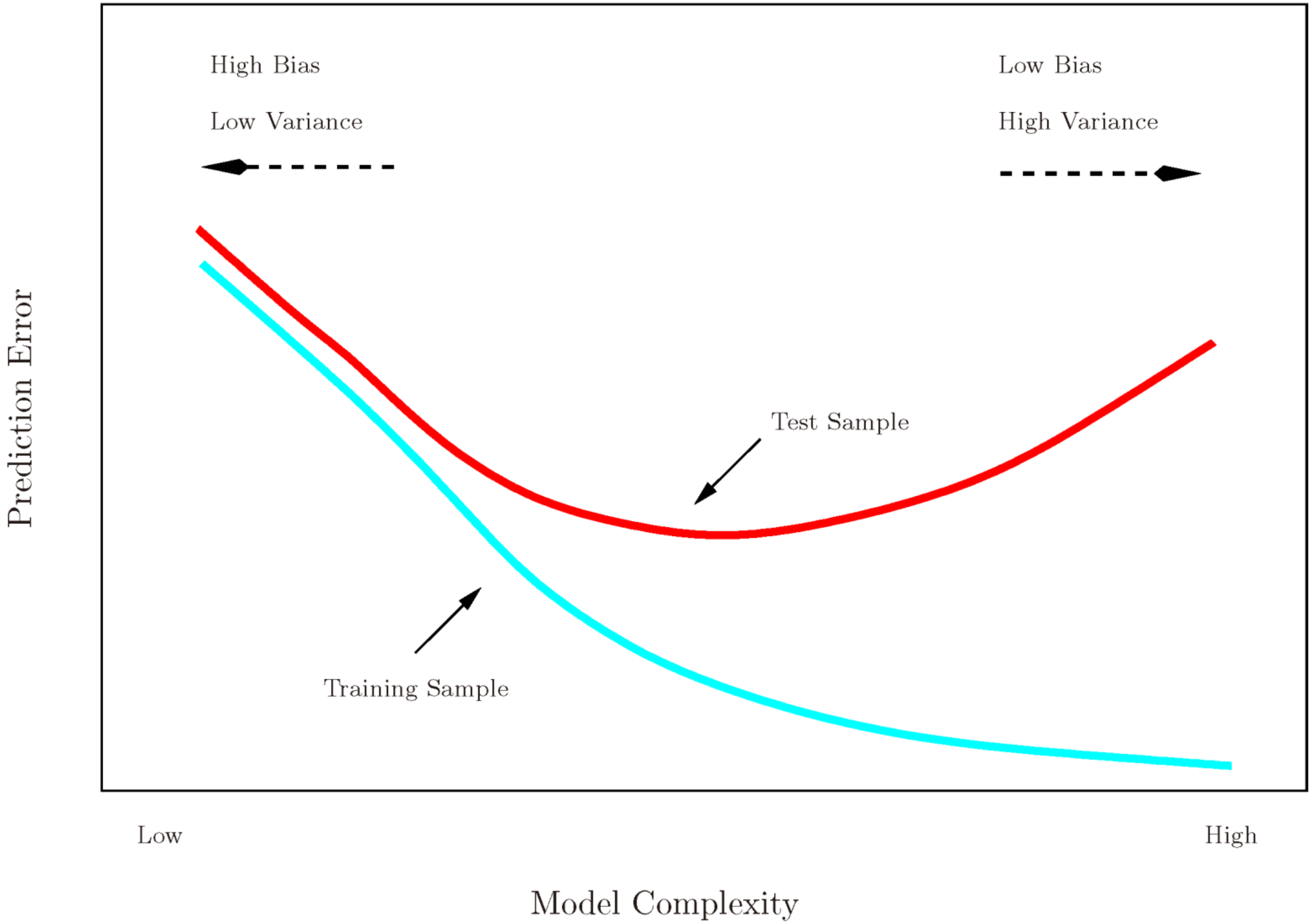
- Cross-Validation
- Bootstrap
- Permutation Test

Cross-Validation

- Test error vs. Training error
  - Test error is the average error that results from using a statistical learning method to predict the response on a new observation – that is, a measurement that was not used in training the method.
  - In contrast, the training error can be easily calculated by applying the statistical learning method to the observations used in its training.
  - The training error rate is often quite different from the test error rate, and in particular the former can ***dramatically underestimate*** the latter.

Cross-Validation

- Test error vs. Training error



Cross-Validation

- How to estimate the test error rate?
  - Best solution: a large designated test set (often not available).
  - Make a mathematical adjustment to the training error rate in order to estimate the test error rate.
  - Hold out a subset of the training observations from the fitting process, and then apply the statistical learning method to those held out observations.

Cross-Validation

- How to estimate the test error rate?
  - Best solution: a large designated test set (often not available).
  - Make a mathematical adjustment to the training error rate in order to estimate the test error rate.
  - Hold out a subset of the training observations from the fitting process, and then apply the statistical learning method to those held out observations. (The validation set approach)

Cross-Validation

- The validation set approach

  - **Randomly** divide the available set of observations into two parts, a *training set* and a *validation set* or *hold-out set*.

  - The model is fit on the training set, and the fitted model is used to predict the responses for the observations in the validation set.

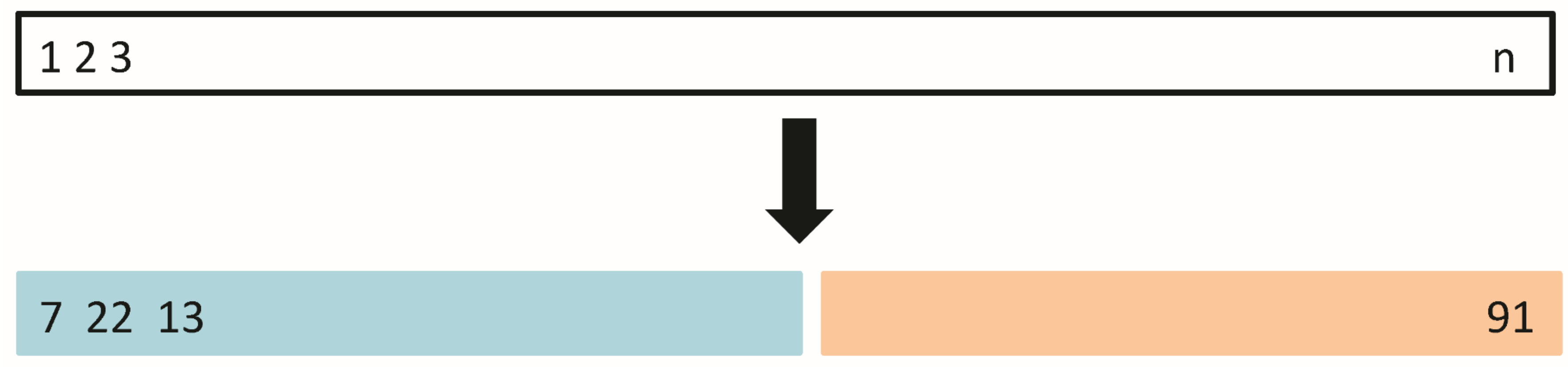
  - The resulting validation set error rate – typically assessed using MSE in the case of a quantitative response and misclassification rate in the case of a qualitative response – provides an estimate of the test error rate.

Cross-Validation

- The validation set approach
  - **Notice for validation set:**
    - Should be representative of the data you want to make predictions on.
    - Should be separate from the training set.
    - Should be large enough to provide a reliable estimate of the model's performance, but not so large that it reduces the amount of data available for training the model.
    - Should be used only for evaluating the model's performance and adjusting its hyperparameters.
    - Should be done randomly to ensure that the splits are unbiased.

Cross-Validation

- The validation set approach

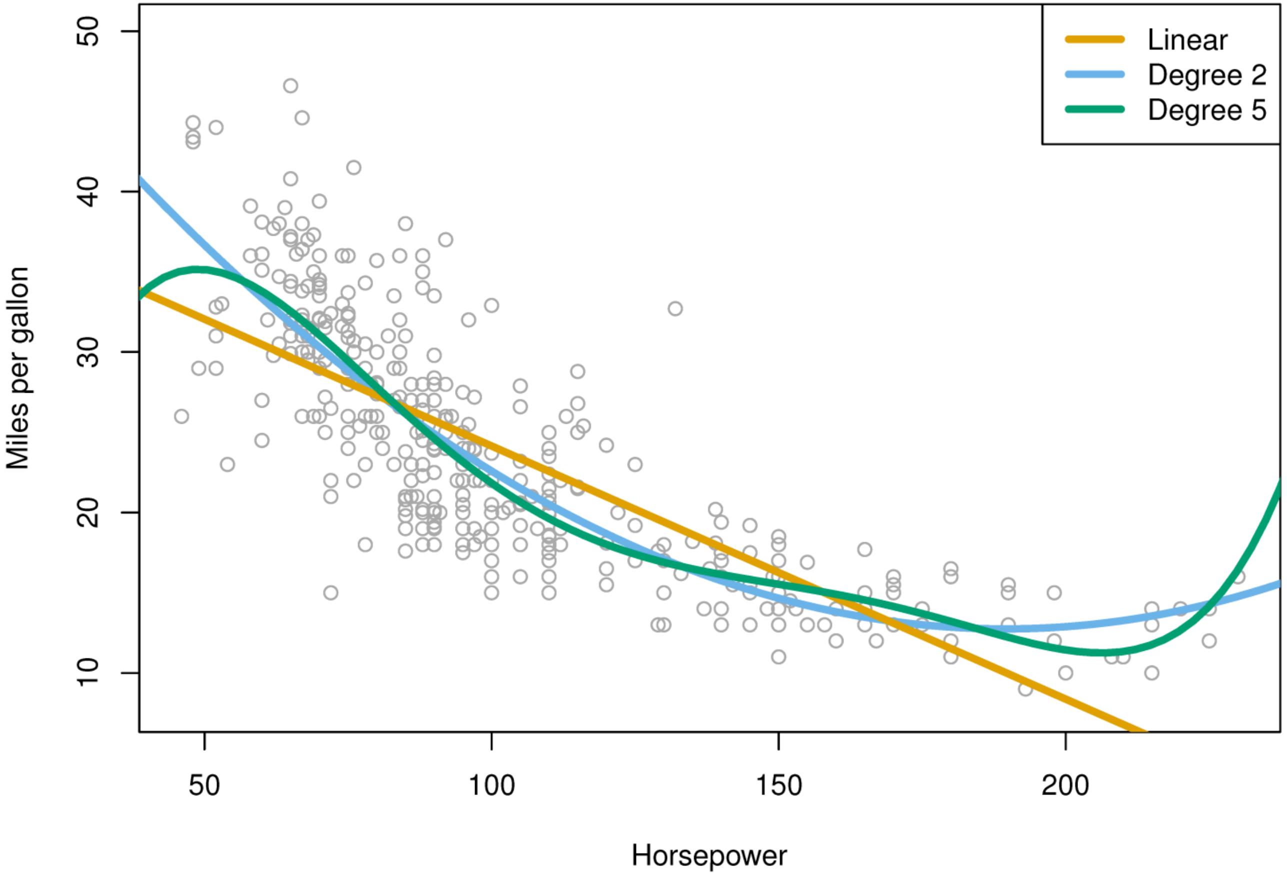


A set of n observations are randomly split into a training set (shown in blue) and a validation set (shown in beige). The statistical learning method is fit on the training set, and its performance is evaluated on the validation set.

Cross-Validation

- The validation set approach

Example: Automobile Data



We may want to compare linear vs high-order polynomial terms in a linear regression.

Cross-Validation

- The validation set approach

  Example: Automobile Data

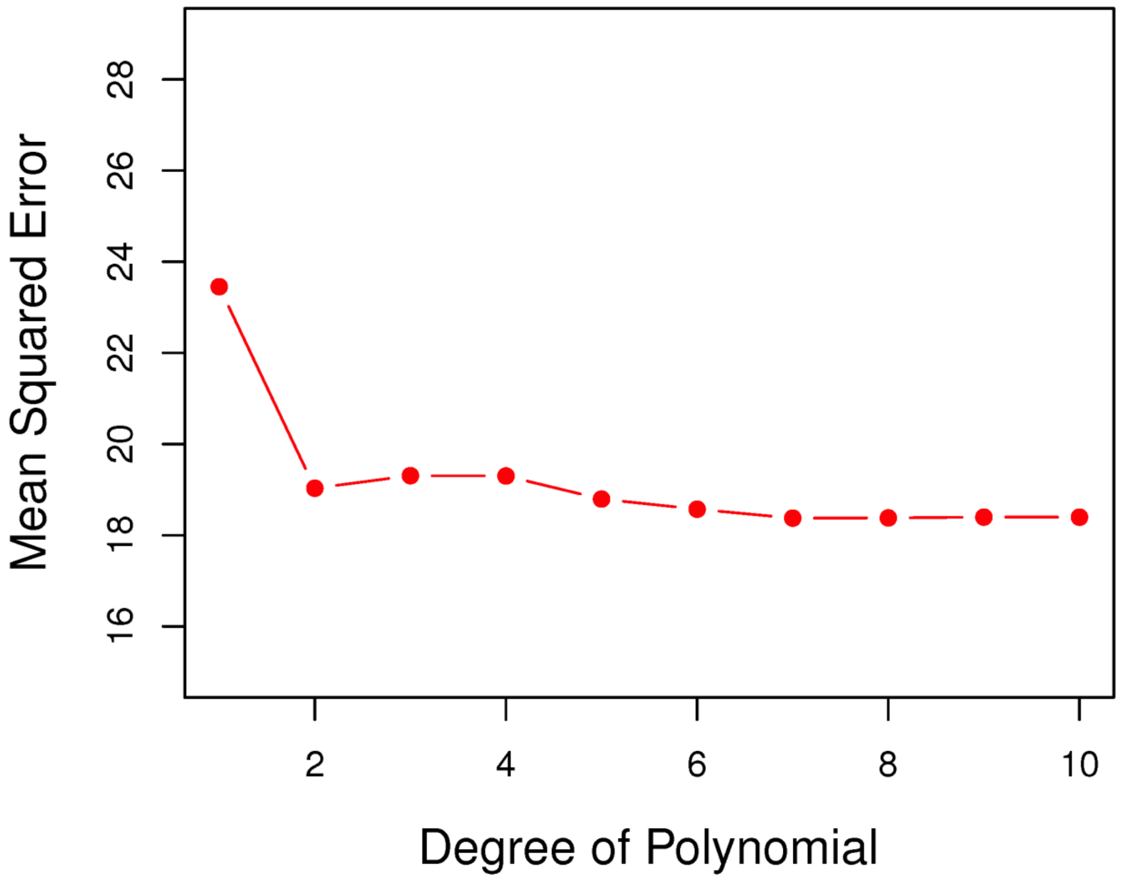
  - We randomly split the 392 observations into two sets, a training set containing 196 of the data points, and a validation set containing the remaining 196 observations.

  - We fit various regression models on the training sample and evaluate their performance on the validation sample, using MSE as a measure of validation set error.

Cross-Validation

- The validation set approach

Example: Automobile Data

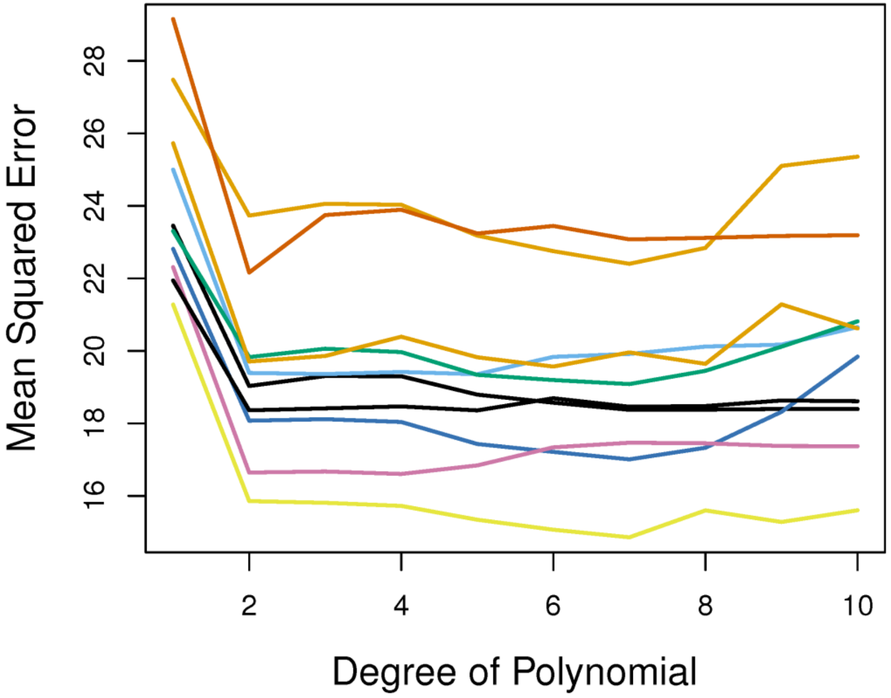


Cross-Validation

- The validation set approach

Example: Automobile Data

If we repeat the process of randomly splitting the sample set into two parts, we will get a somewhat different estimate for the test MSE.

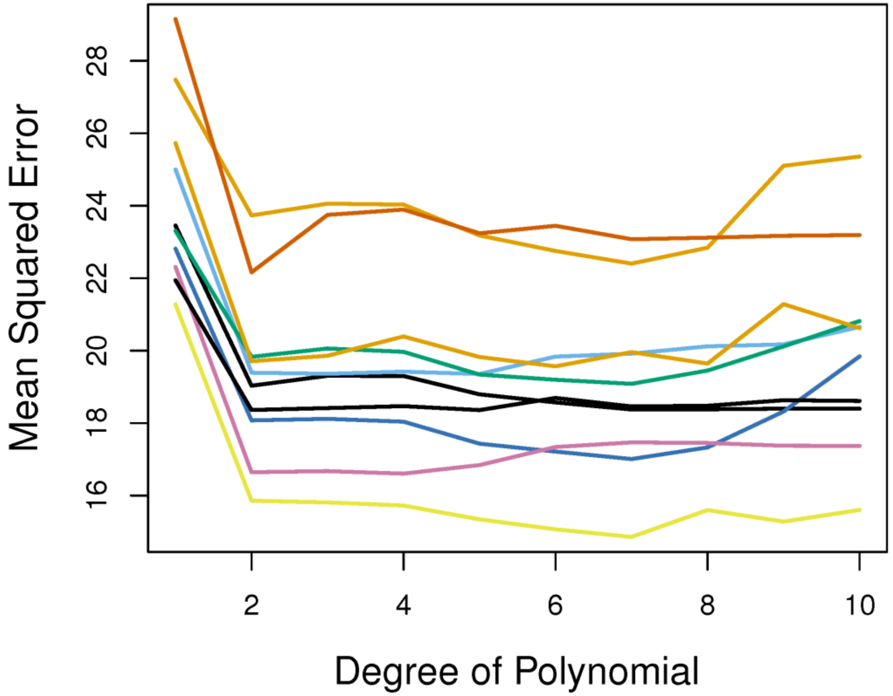


10 different validation set MSE curves produced using 10 different random splits of the observations into training and validation sets.

Cross-Validation

- The validation set approach

Example: Automobile Data

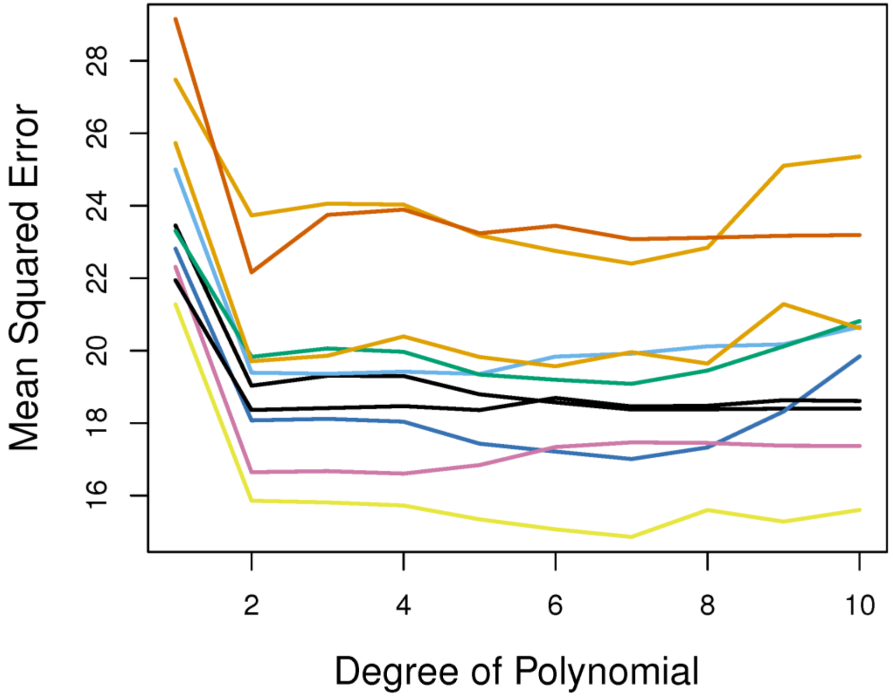


- ❖ All 10 curves indicate that the model with a quadratic term has a dramatically smaller validation set MSE than the model with only a linear term.
- ❖ All 10 curves indicate that there is not much benefit in including cubic or higher-order polynomial terms in the model.

Cross-Validation

- The validation set approach

Example: Automobile Data



- ❖ Each of the 10 curves results in a different test MSE estimate for each of the 10 regression models considered.
- ❖ There is no consensus among the curves as to which model results in the smallest validation set MSE.
- ❖ Based on the variability among these curves, all that we can conclude with confidence is that the linear fit is not good.

Cross-Validation

- The validation set approach

Drawbacks

- The validation estimate of the test error rate can be highly variable, depending on precisely which observations are included in the validation set.

- In the validation approach, only a subset of the observations are used to fit the model. Since statistical methods tend to perform worse when trained on fewer observations, the validation set error rate may tend to **overestimate** the test error rate for the model fit on the entire data set.

Cross-Validation

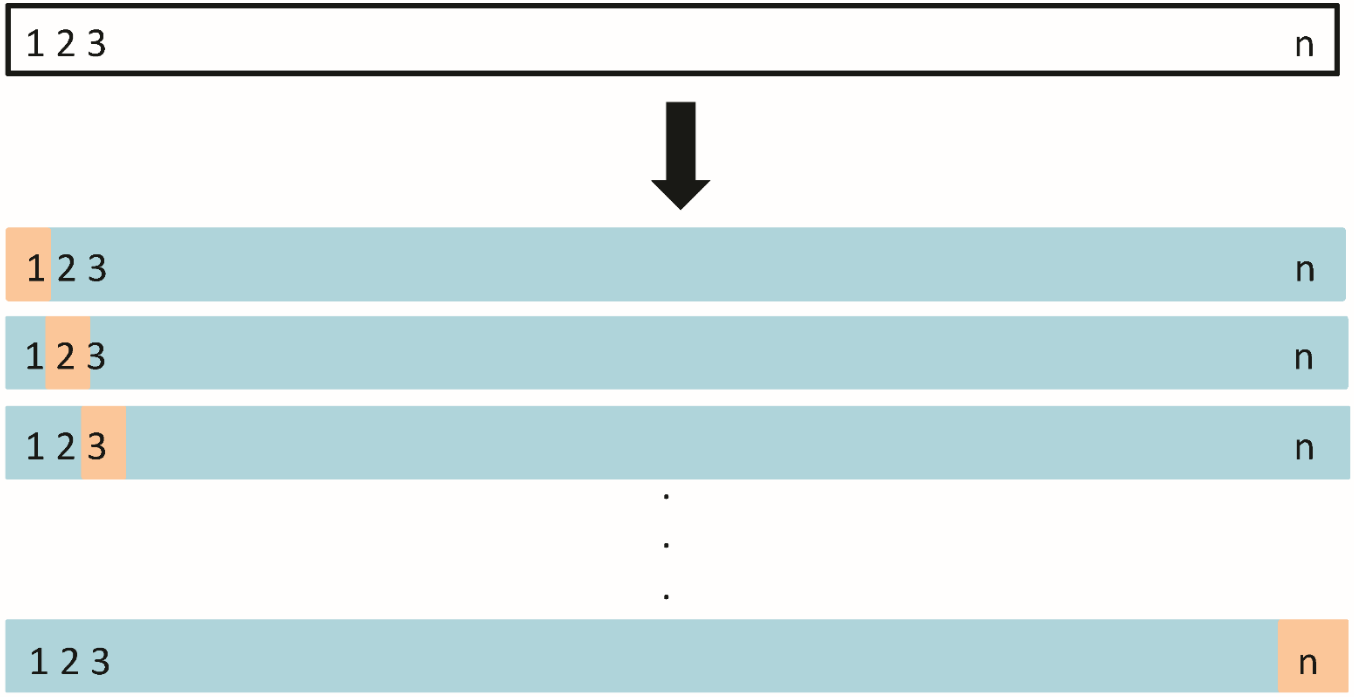
- Leave-One-Out Cross-Validation (LOOCV)
  - In LOOCV, a single observation (x_1, y_1) is used for the validation set, and the remaining observations $\{(x_2, y_2), \ldots, (x_n, y_n)\}$ make up the training set.
  - The statistical learning method is fit on the $n-1$ training observations, and a prediction $\hat{y}_1$ is made for the excluded observation, using its value x_1 . We have $MSE_1 = (y_1 - \hat{y}_1)^2$.
  - Repeat the procedure by respectively selecting $(x_2, y_2), (x_3, y_3), \ldots, (x_n, y_n)$ to be the validation data, and the other n-1 to be the training data, yielding $MSE_2, MSE_3, \ldots, MSE_n$.

Cross-Validation

- Leave-One-Out Cross-Validation (LOOCV)

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

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



Cross-Validation

- Leave-One-Out Cross-Validation (LOOCV)

Advantages of LOOCV over the validation set approach

- It has far less bias. In each loop, the training set contains observations as many as those in the entire data set rather than only half the size of the original data set. Consequently, the LOOCV approach tends not to overestimate the test error rate as much as the validation set approach does.

- The validation approach will yield different results when applied repeatedly (due to randomness in the training/validation set splits). Performing LOOCV multiple times will always yield the same results: there is no randomness in the training/validation set splits.

- It uses all of the available data for both training and testing, which can lead to more accurate estimates of model performance.

Cross-Validation

- Leave-One-Out Cross-Validation (LOOCV)

  Notice:

  - LOOCV can be computationally expensive, especially for large datasets, because it requires training and testing the model for each data point.

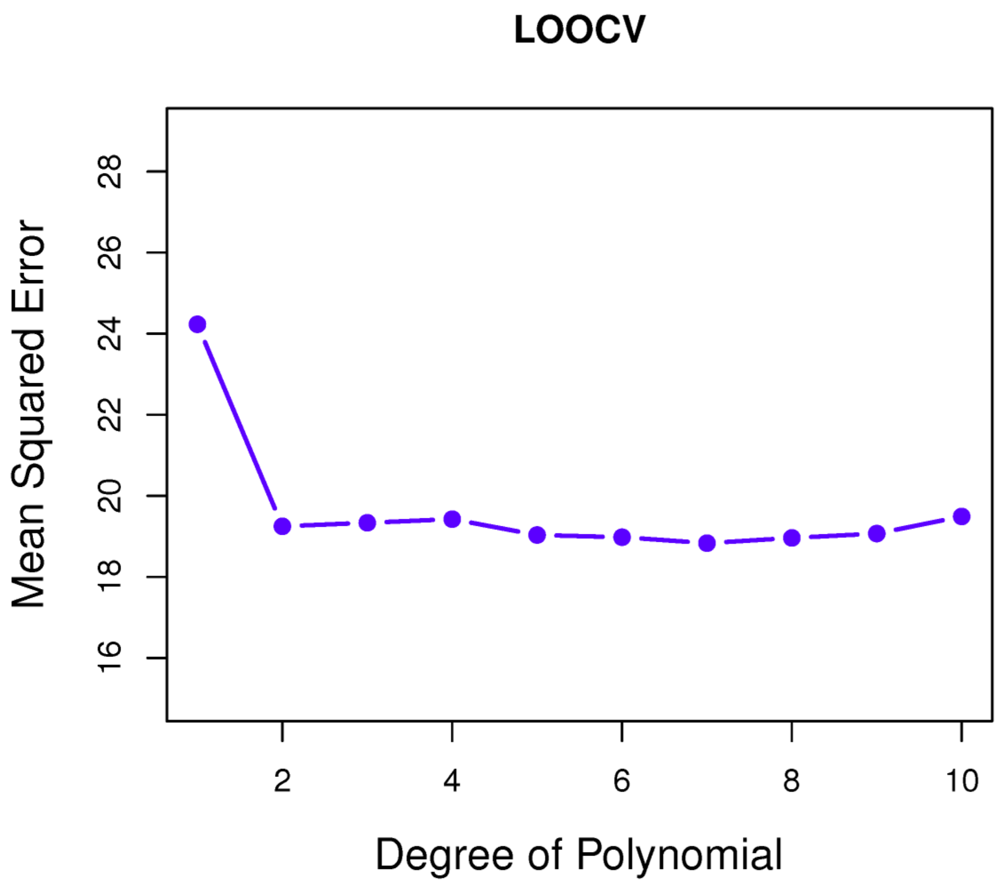
  - The LOOCV estimate of the model's performance is unbiased, but it may have high variance. It may be useful to compute the confidence interval of the LOOCV estimate to get a better idea of the uncertainty in the estimate.

  - LOOCV is a good choice when the dataset is small or when it's important to use all available data for both training and testing. If the dataset is large, other cross-validation techniques, such as k-fold cross-validation, may be more efficient.

Cross-Validation

- Leave-One-Out Cross-Validation (LOOCV)

Example: Automobile Data



Cross-Validation

- Leave-One-Out Cross-Validation (LOOCV)

Note: LOOCV has the potential to be expensive to implement, since the model has to be fit for n times. This can be very time consuming if n is large, and if each individual model is slow to fit.

With **least squares linear or polynomial regression**, an amazing shortcut makes the cost of LOOCV the same as that of a single model fit!

$$CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\hat{y}_i}{1-h_i}\right)^2$$

→ Does not hold in general !

$\hat{y}_i$ -- the *i*th fitted value from the original least squares fit

h_i -- the *i*th leverage statistic

Cross-Validation

- Leave-One-Out Cross-Validation (LOOCV)

  Comment: LOOCV is a very general method, and can be used with any kind of predictive modeling. For example, we could use it with logistic regression or linear discriminant analysis, or any of the methods discussed in later chapters.

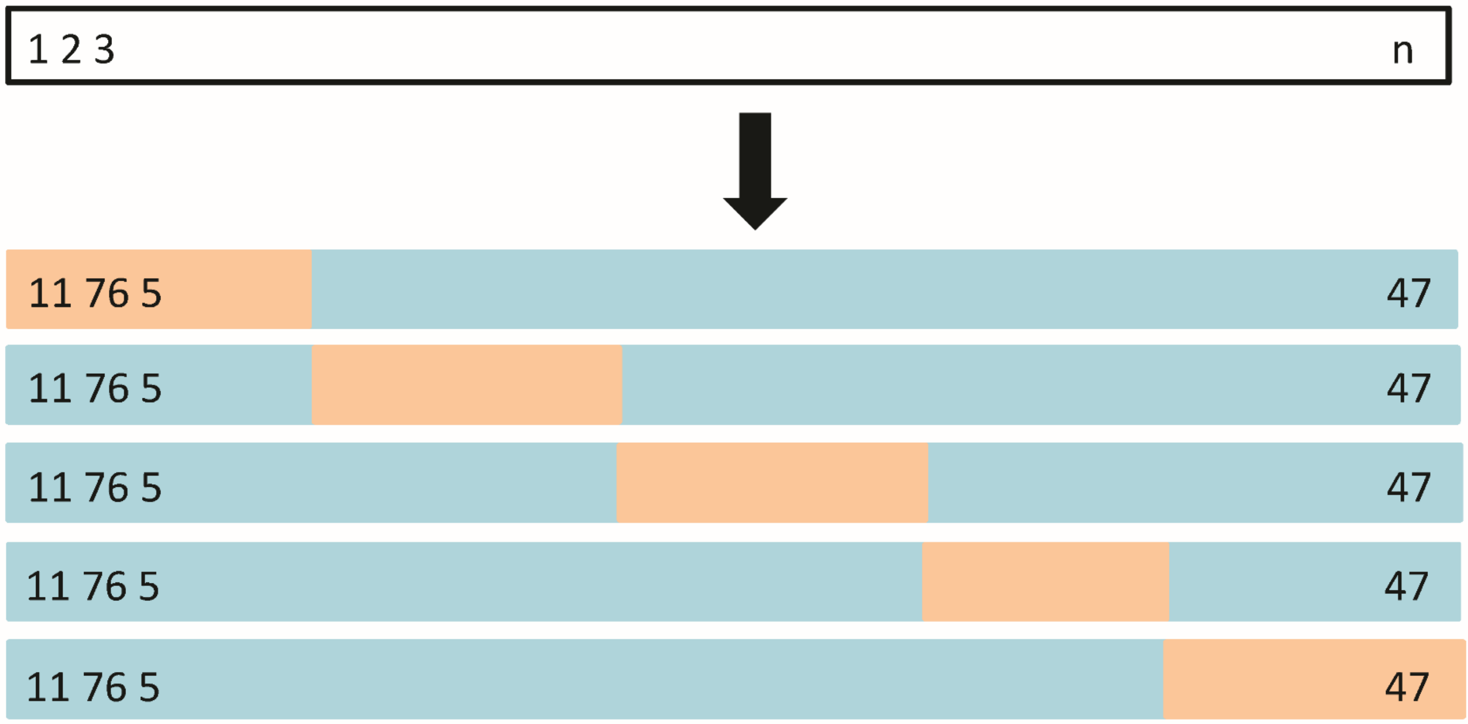
Cross-Validation

- k-Fold Cross-Validation
  - Randomly divide the set of observations into k groups, or folds, of approximately equal size.
  - The first fold is treated as a validation set, and the method is fit on the remaining k-1 folds. MSE_1 is then computed on the observations in the held-out fold.
  - Repeat the procedure k times; each time, a different group of observations is treated as a validation set, resulting in $MSE_1, MSE_2, \ldots, MSE_k$.

Cross-Validation

- k-Fold Cross-Validation

The k-fold CV estimate for the test MSE is the average of these k test error estimates: $CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k} MSE_i$



Cross-Validation

- k-Fold Cross-Validation
  - LOOCV is a special case of k-fold CV in which k=n.
  - Typically, k=5 or k=10.
  - Performing LOOCV may pose computational problems, especially if n is extremely large. In contrast, performing 5-fold or 10-fold is much more feasible.