Linear Discriminant Analysis

Chapter 4 – Part III

Linear Discriminant Analysis

- Why not Logistic Regression?
- Using Bayes' Theorem for Classification
- Linear Discriminant Analysis
    - Single predictor case

Why not Logistic Regression?

- Logistic regression models the class probability $P(Y=k \mid X=x)$
    -

- Alternative:
    -

    - **Linear/Quadratic Discriminant Analysis** --

Why not Logistic Regression?

- Why not just use logistic regression?

1. When groups are well-separated, parameter estimates for logistic regression are highly unstable.
2. For small sample sizes, discriminant analysis can be more stable if groups are close to normal.
3. Discriminant analysis is more natural when we consider more than two classes.

Bayes' Theorem for Classification

- We used the terms earlier, Bayes Rule and Bayes Classifier.
- The theorem by Thomas Bayes, gives a nice relationship between conditional distributions.
- Recall, here *Y* can take on values from 1 to K representing the class.
- Bayes' Theorem:
- Let $\pi_k = P(Y = k)$, which represents the marginal (overall) probability (or proportion) of class *k.*
  - In Bayes' language, this is the **prior probability**.
  - Our guess for the probability prior to looking at the *X* value.

Bayes' Theorem for Classification

- Bayes' Theorem: $P(Y=k \mid X=x) = \dfrac{P(X=x \mid Y=k)P(Y=k)}{P(X=x)}$
- Let $f_k(x) = P(X=x \mid Y=k)$ be the **density function** in group *k*.
- Then $P(X=x) = \sum_{m=1}^{K} \pi_m f_m(x)$ since marginal distribution of *X* is mixture of the individual distributions with proportion π_k for group *k.*
- So
$$p_k(x) = P(Y=k \mid X=x) = \frac{\pi_k f_k(x)}{\sum_{m=1}^{K} \pi_m f_m(x)}$$
  - Also known as **posterior probability**.
  -

Bayes' Theorem for Classification

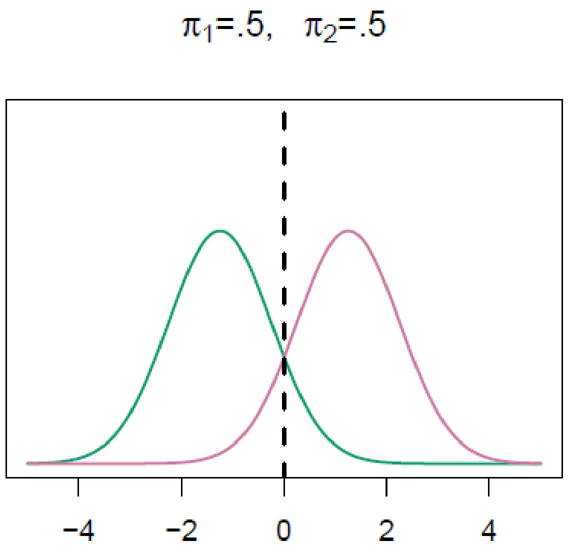
- We implicitly assumed that our predictors were discrete, since we wrote

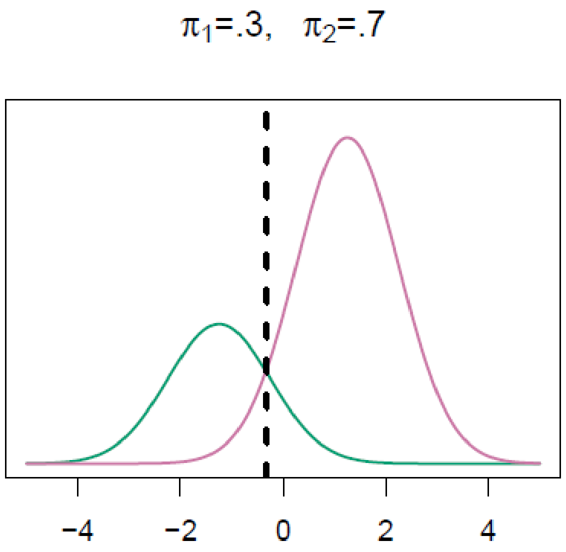
$$f_k(x) = P\left(X = x \mid Y = k\right)$$

- Still valid if predictors are not discrete, i.e. continuous.
  - $f_k(x)$ now probability **density function**, representing probability on infinitesimal neighborhood.

- Plug in estimates of π_k, f_k for each *k*.
  - For π_k , use observed proportion in the group if not known.
  - For f_k, more complicated. Typically need to assume something.
- Classify to class with highest posterior.
  - This is Bayes Rule.

Bayes' Theorem for Classification





Linear Discriminant Analysis when p=1

- Most common assumption is each group is normal (or Gaussian):
- $$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{1}{2\sigma_k^2}(x-\mu_k)^2 \right)$$
- We get:

$$p_k(x) = P\left(Y = k \mid X = x\right) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_k)^2 \right)}{\sum_{m=1}^{K} \pi_m \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_m)^2 \right)}$$

- Looks scary!

Linear Discriminant Analysis when p=1

- We have:

$$p_k(x) = P\left(Y = k \mid X = x\right) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_k)^2\right)}{\sum_{m=1}^{K} \pi_m \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_m)^2\right)}$$

- So, for any given value of $X = x$, we would plug that value in and classify to whichever class gives the largest value.

Linear Discriminant Analysis when p=1

- Rule for each value of $X = x$ is to assign to the class with largest **discriminant function**.

$$\delta_k(x) = x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)$$

-

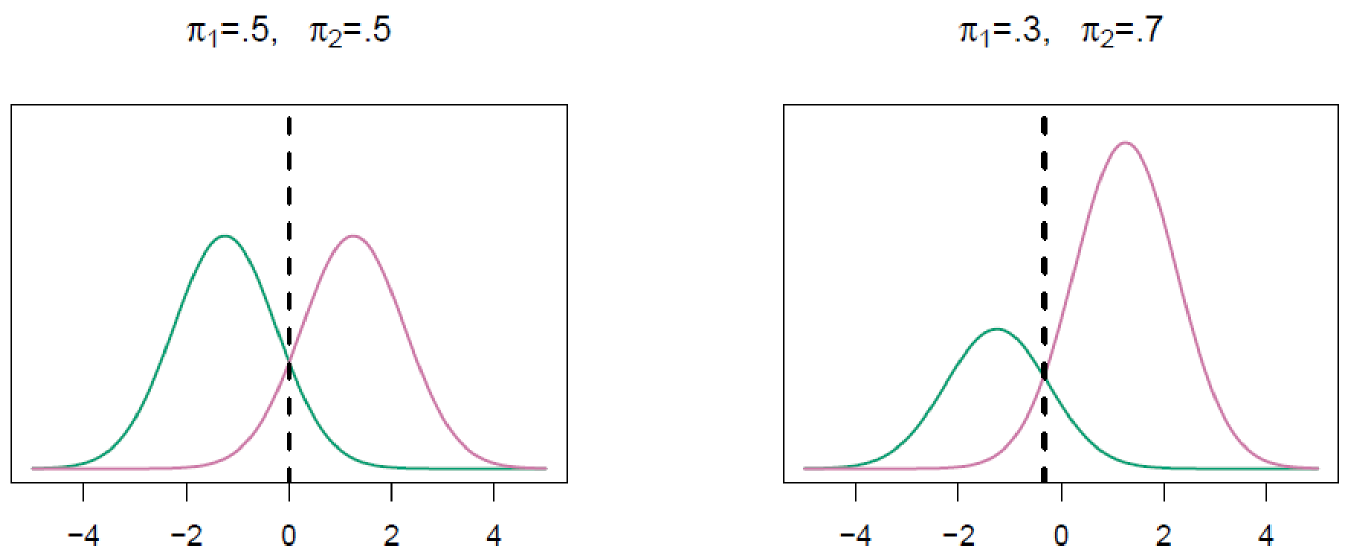
- Hence it is called **linear discriminant analysis**.

Linear Discriminant Analysis when *p*=1

- Example: For two classes, i.e. K = 2 and equal proportions in each group.

- Unequal proportions shifts boundary to classify more into larger class.

Bayes' Theorem for Classification



Estimating the Parameters

- In practice, we need to estimate the parameters:

Linear Discriminant Analysis

- Linear Discriminant Analysis assumes:
  - Each class is normally distributed.
  - Different means.
  - Same variances.
  - Yields a linear function for its decision rule.
- Next time: Generalize to multiple predictors.

More next time!

Linear Discriminant Analysis - Part II

Chapter 4 – Part IV

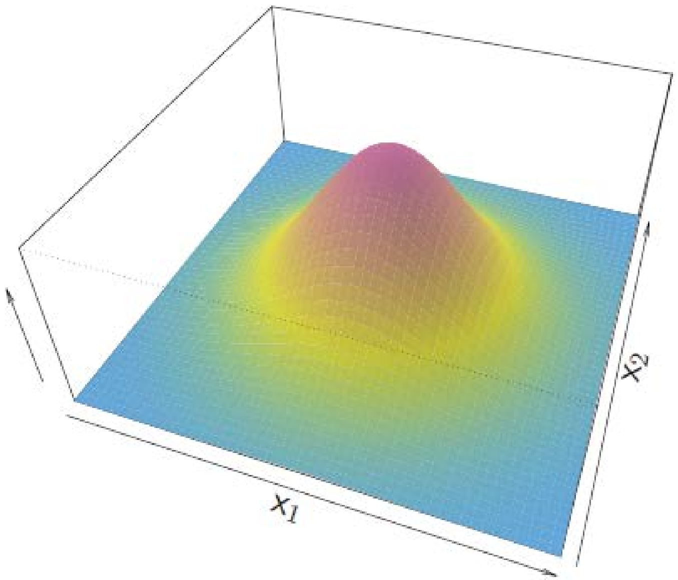
Linear Discriminant Analysis – Part II

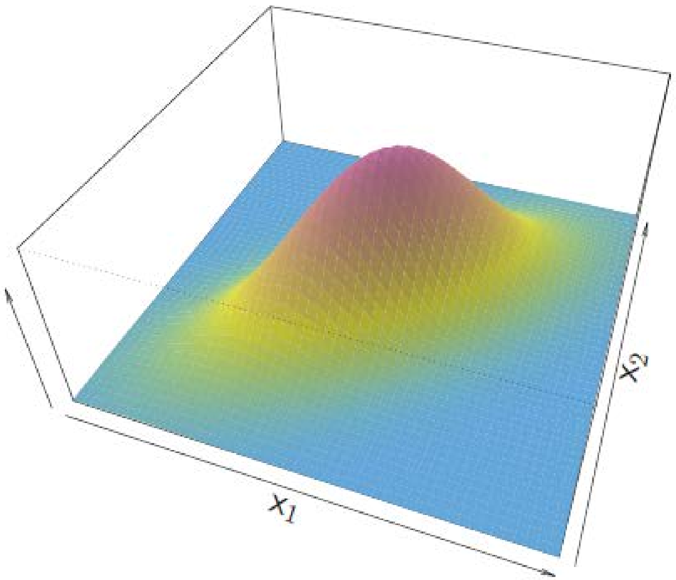
- Linear Discriminant Analysis
  - The multiple predictor case

Multivariate Case

- X is now multivariate.
- For each class, it now has:
    - Mean vector μ_k now p-dimensional vector.
    - Covariance matrix Σ_k is p x p matrix.
        - Diagonal represents variance for each predictor.
        - Off-diagonals are covariances between predictors.
- We assume multivariate normal in each class.

Multivariate Normal Density





Multivariate Normal Density

- The multivariate normal density can be written as:

$$f_k(x) = \frac{1}{(2\pi)^{p/2}\left|\Sigma_k\right|^{1/2}} \exp\left(-\left(x-\mu_k\right)^T \Sigma_k^{-1}\left(x-\mu_k\right)\right)$$

- Value inside exponential called **Quadratic Form.**

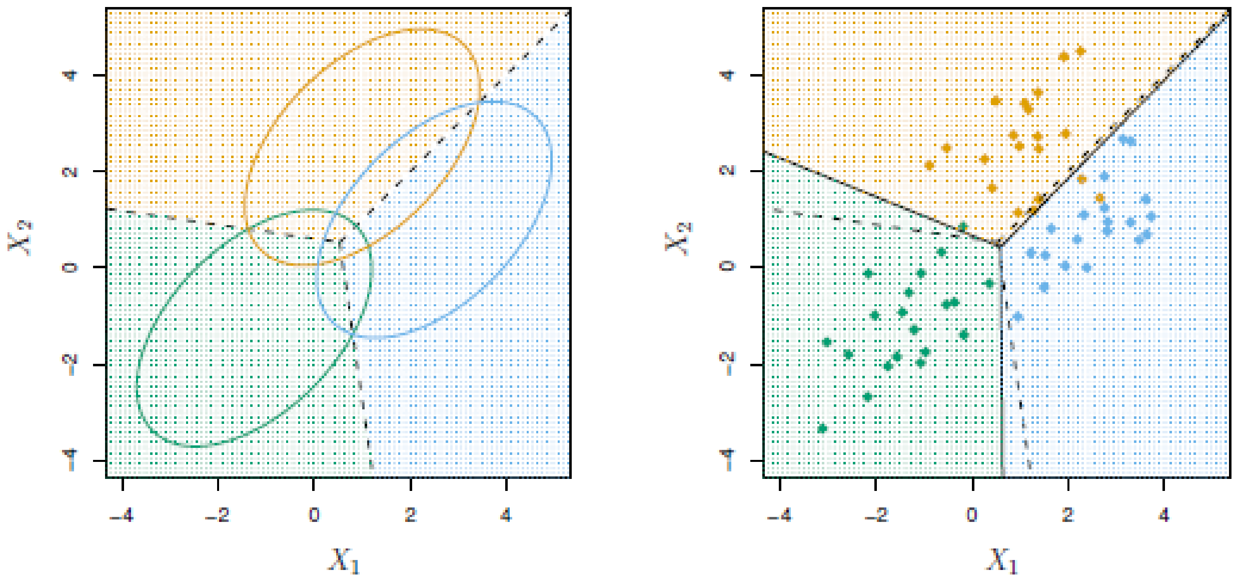
Linear Discriminant Analysis (LDA)

- Assume each class is multivariate normal with same covariance matrix, Σ
- Plugging in as before gives us the discriminant function.
- Looks complicated, but
- Linear Discriminant Analysis (LDA)!

Linear Discriminant Analysis (LDA)

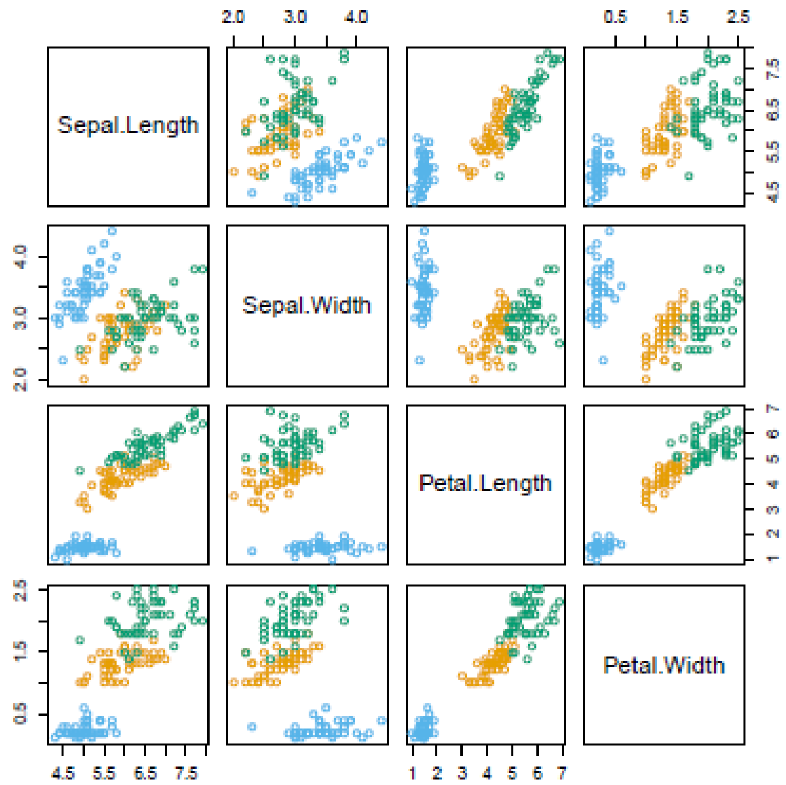
- Consider discussion of leverage statistic in linear regression.
- Measured leverage by expanding ellipses centered at mean of predictors.
-
- Leverage statistic is also known as **Mahalanobis Distance**.

Linear Discriminant Analysis (K = 3, p = 2)



Example: Fisher's Iris Data

- Classic data set.
  - 3 species of Iris (flower).
  - 50 observations per class in training set.



Estimating Probabilities from LDA

- Once we have estimates $\hat{\delta}_k(x)$, we can turn these into estimates for class probabilities:

$$\widehat{\Pr}(Y = k|X = x) = \frac{e^{\hat{\delta}_k(x)}}{\sum_{l=1}^{K} e^{\hat{\delta}_l(x)}}.$$

- So classifying to the largest $\hat{\delta}_k(x)$ amounts to classifying to the class for which $\widehat{\Pr}(Y = k|X = x)$ is largest.

- When $K = 2$, we classify to class 2 if $\widehat{\Pr}(Y = 2|X = x) \geq 0.5$, else to class 1.

Linear Discriminant Analysis

- For the credit card default data, we fit LDA using balance and student status as predictors.
- 10,000 training observations.
- Training error of only 2.75% misclassified. Excellent performance?
- Ideally we have test set. But here overfitting may not be too bad, since we have 10,000 observations and only 2 predictors.
- Next time, we will talk more about error rates, and also discuss Quadratic Discriminant Analysis.

More next time!

Discriminant Analysis and Classification - Continued

Chapter 4 – Part V

Discriminant Analysis and Classification

- Error Rates for Classification
  - The Confusion Matrix
- Extensions to LDA
  - Quadratic Discriminant Analysis
  - Naïve Bayes

Error Rates for Classification

- Error rate for the Default data was 2.75%.
- Sounds good, but:

Confusion Matrix for Binary Classification

		True Default Status		
		No	Yes	Total
Predicted	No	9644	252	9896
Default Status	Yes	23	81	104
	Total	9667	333	10000

- **Confusion Matrix**
  - Rows are predicted class, i.e. predict No Default / Yes Default.
  - Columns are true class, i.e. True Default Status (No/Yes).
- LDA predicts total of 9896 non-default and only 104 in default.

Error Rates for Binary Classification

		True Default Status		
		No	Yes	Total
Predicted	No	9644	252	9896
Default Status	Yes	23	81	104
	Total	9667	333	10000

- Instead of an overall error rate, consider class specific error rates.
- **False Positive Rate (1 – Specificity, Type I Error)** –

- **False Negative Rate (1 – Sensitivity, Type II Error)** –

Error Rates for Binary Classification

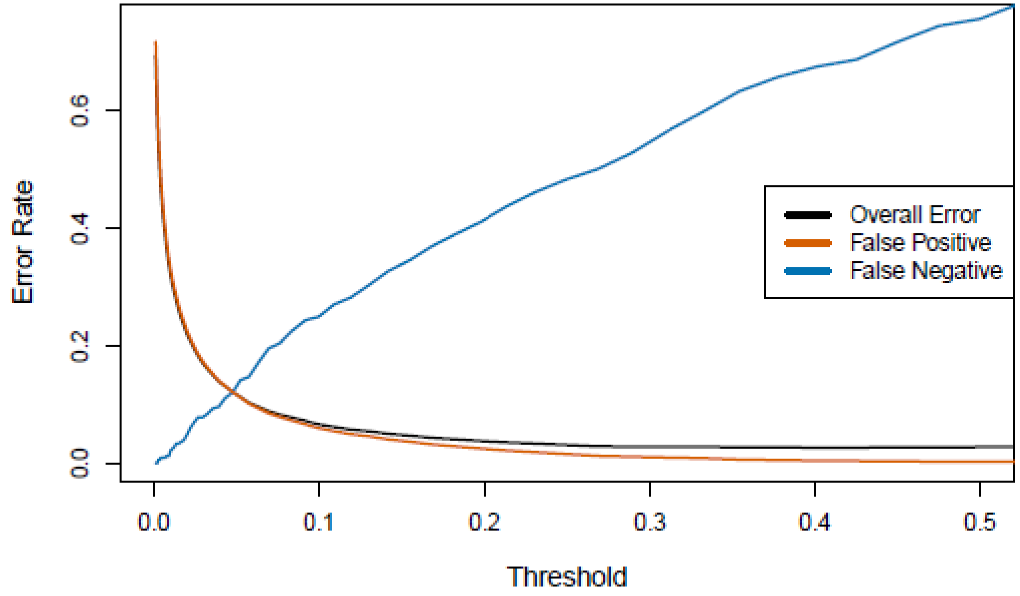
- LDA, and Bayes Rule in general, try to minimize overall error rate.
- Credit card company more concerned with class specific rates.
- Want to identify more than 24.3% of the defaulters.
- Willing to deny credit to non-defaulters if helps identify more defaulters.

Error Rates for Binary Classification

- LDA or other estimate of Bayes Classifier predicts class with largest posterior probability.
- In binary case, classifies to positive if probability is > 0.5.

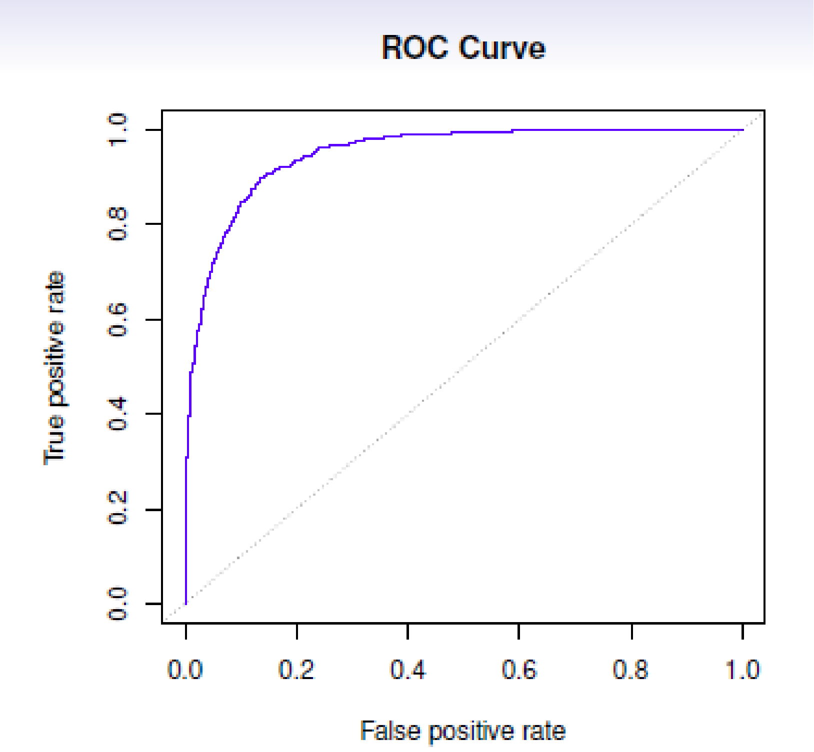
		True default status		
		No	Yes	Total
Predicted default status	No	9,432	138	9,570
	Yes	235	195	430
	Total	9,667	333	10,000

Error Rates for Binary Classification



Receiver Operating Characteristic Curve

- The **receiver operating characteristic curve (ROC curve)**



Other Forms of Discriminant Analysis

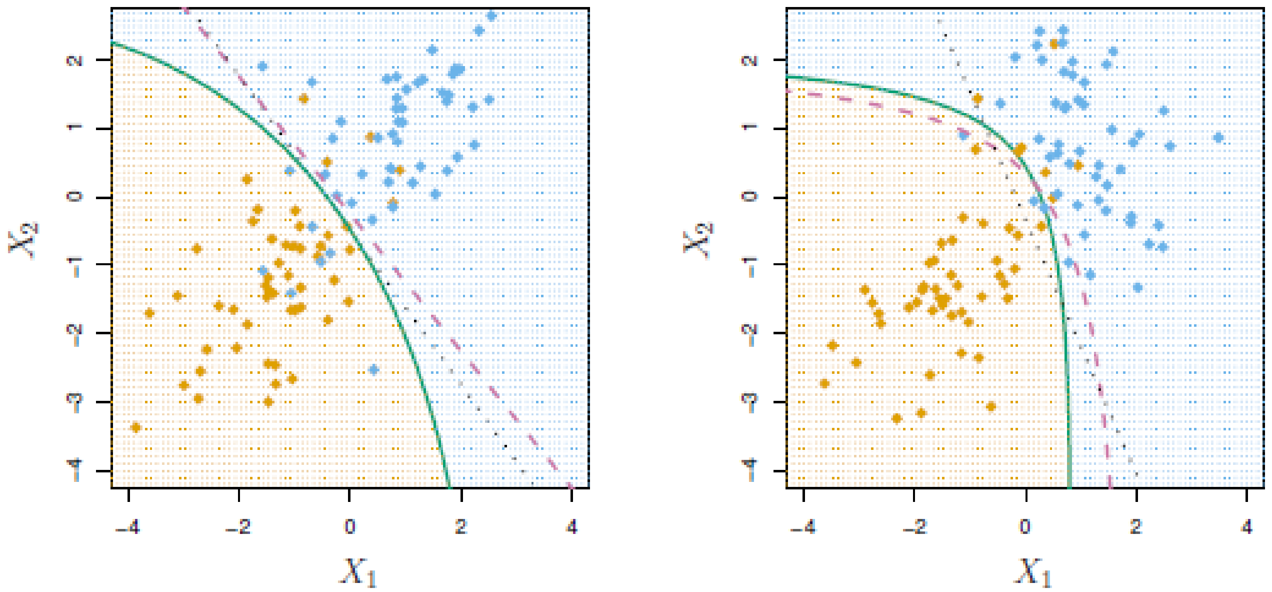
- Recall: posterior probability:

$$\Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}$$

- When $f_k(x)$ are Gaussian densities, with the same covariance matrix $\mathbf{\Sigma}$ in each class, this leads to linear discriminant analysis. By altering the forms for $f_k(x)$, we get different classifiers.

- Gaussian, but different variances in each class, leads to **Quadratic Discriminant Analysis (QDA)**.

Quadratic Discriminant Analysis



$$\delta_k(x) = -\frac{1}{2}(x - \mu_k)^T \mathbf{\Sigma}_k^{-1}(x - \mu_k) + \log \pi_k$$

Because the $\mathbf{\Sigma}_k$ are different, the quadratic terms matter.

Gives quadratic boundaries instead of linear.

LDA vs. QDA

- LDA assumes equal variances in each group.
- QDA is more flexible, allowing for unequal variances.

Naïve Bayes

- LDA
- QDA
- Naïve Bayes

Logistic Regression vs. LDA

- For binary problem, LDA classifies using posterior probabilities.
- Looking back at their form, and taking the log odds gives:

$$\log\left(\frac{p_1(x)}{1-p_1(x)}\right) = \log\left(\frac{p_1(x)}{p_2(x)}\right) = c_0 + c_1 x_1 + \ldots + c_p x_p$$

Logistic Regression vs. LDA

- In logistic regression, we estimate the parameters using maximum likelihood, conditional on the predictors.

- In LDA, stronger assumptions are made.

- Note: using quadratic terms in logistic regression, can have same discussion for logistic regression vs. QDA.

KNN vs. (Logistic Regression and LDA)

- Recall: KNN is fully non-parametric.
  - No assumptions are made about shape of the decision boundary.

- **Advantage:**

- **Disadvantage:**

QDA vs. (Logistic Regression, LDA, and KNN)

- QDA is a compromise between linear, and non-parametric.

- If true decision boundary is:
  - *Linear:*
  - *Moderately non-linear:*
  - *Highly non-linear:*

- Also, less data, the simpler we need to be.

Classification Approaches

- We have completed our discussion of some common approaches for classification.
- Next time, we will discuss some ideas of how to deal with not having external test sets.
- The approaches: Resampling Methods.

More next time!

Resampling Methods

Chapter 5 – Part I

Resampling Methods

- Cross-Validation (CV)
  - The Validation Set Approach
  - Leave-One-Out CV
  - k-Fold CV
  - Bias-Variance Trade-Off for k-Fold CV
  - CV on Classification Problems

What are Resampling Methods?

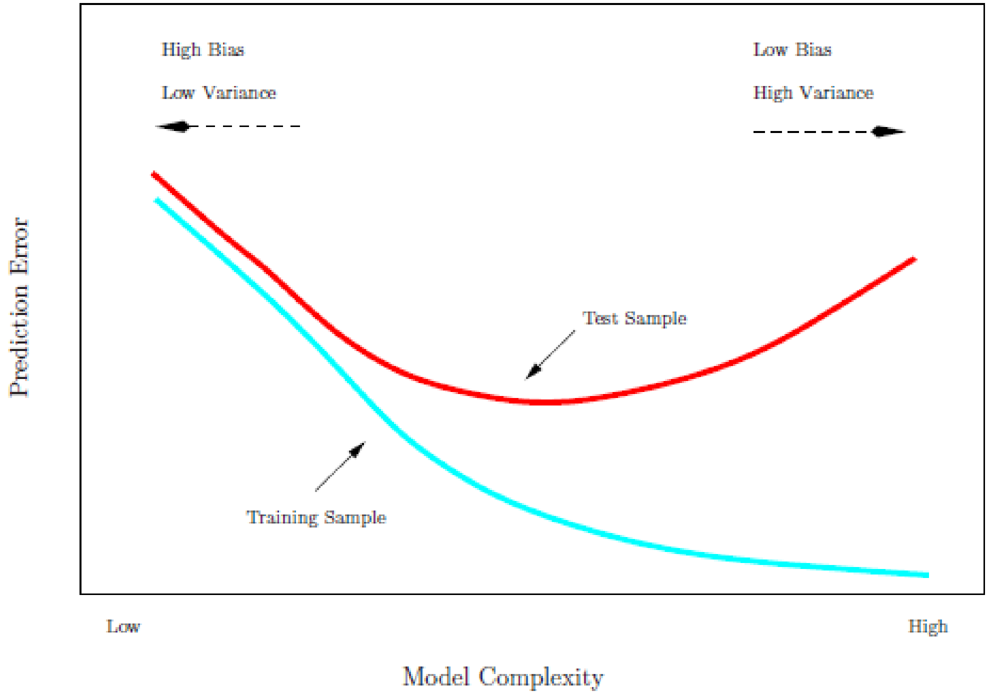
- **Resampling Methods** involve repeatedly drawing samples from the training set.
  - Refit model of interest.
  - Obtain additional information about fitted model.

- Computationally expensive – but we have powerful computers.
- Two methods: Cross-Validation and Bootstrap

Training vs. Test Error

- Recall difference between training and test error.
- **Test Error** is the average error from using the method to predict on a new observation that was not used in training.
- **Training Error** is calculated by applying the method to the training observations.
- Ideally want the test error.

Training vs. Test Error



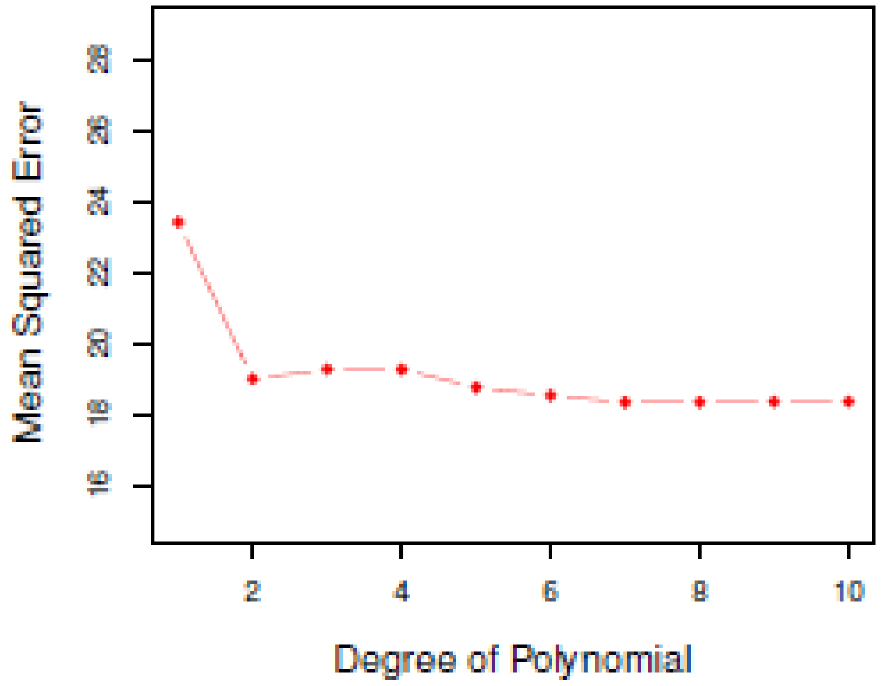
Estimating Test Error

- Best way to estimate test error.
  - Have a large separately designated test set.
  - Often not feasible.
- Alternative Approaches:

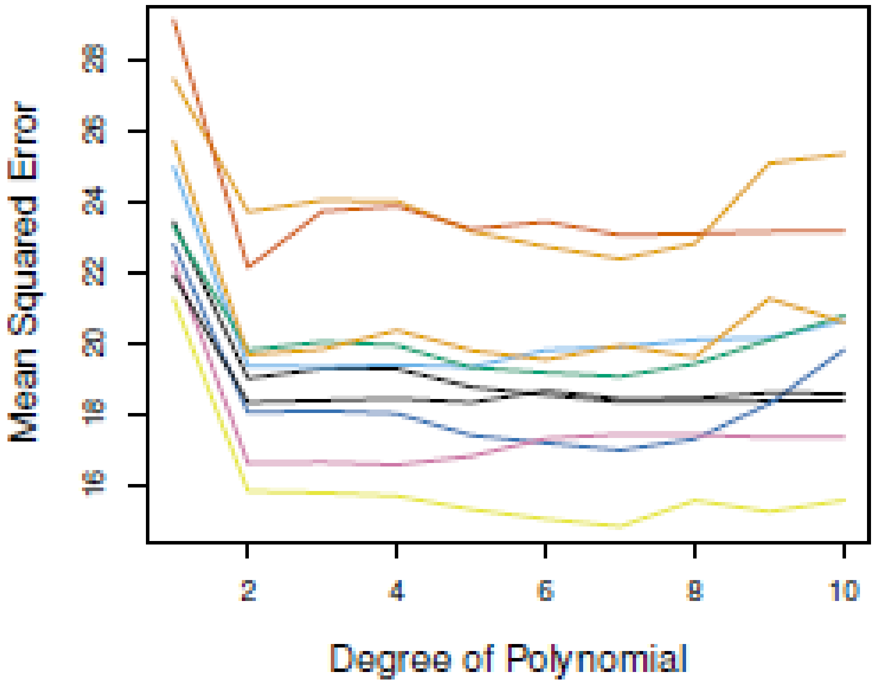
Validation Set Approach

- If we have enough data, we <u>*randomly*</u> divide into a **training set** and a **validation set**.

Example: Auto Data



Example: Auto Data



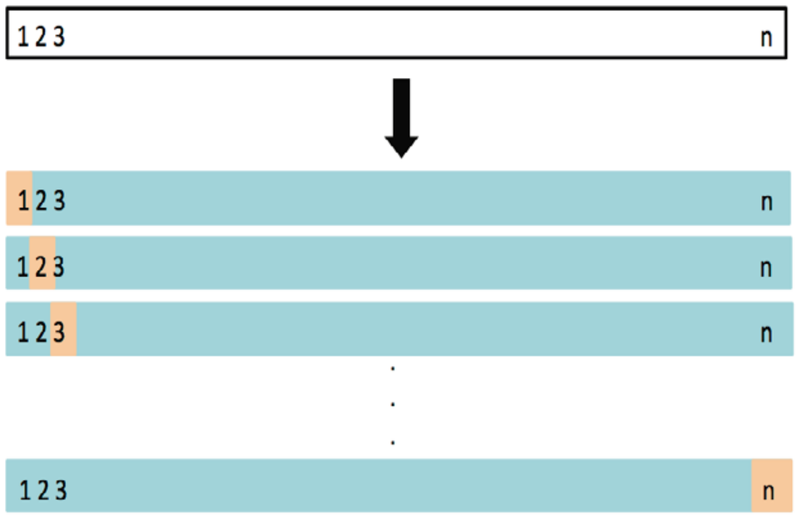
Validation Set Approach

- **Advantages:**

- **Disadvantages:**

Leave-One-Out Cross-Validation (LOOCV)

- **Leave-One-Out Cross-Validation (LOOCV)** attempts to address the drawbacks of validation set approach.
- Instead of splitting the data set into two parts, we:

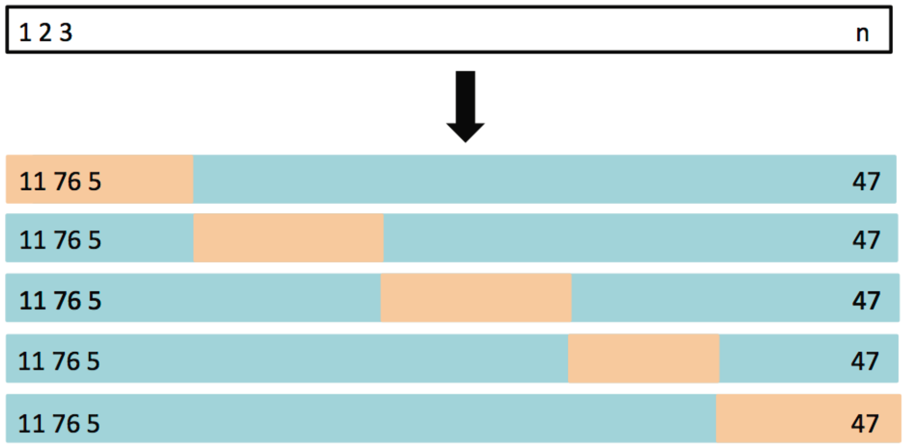


LOOCV vs. Validation Set

- **Advantages:**
  - LOOCV has less bias in estimating test error, since each time we fit the method with a training set of size *n*-1, rather than smaller.
  - Will obtain same result each time, since we did not do any random splitting.
- **Disadvantages:**
  - LOOCV is computationally intensive since every method/model is fit *n* times.
    - Except for using least squares regression!
    - In that special case, we only need to fit model once and we can actually calculate our MSE for all left out observations.
    - Surprisingly (but ONLY for the least squares regression case): $$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{1 - h_i}\right)^2$$
    - In ALL other cases, we need to fit the method *n* times!

K-Fold Cross-Validation (K-Fold CV)

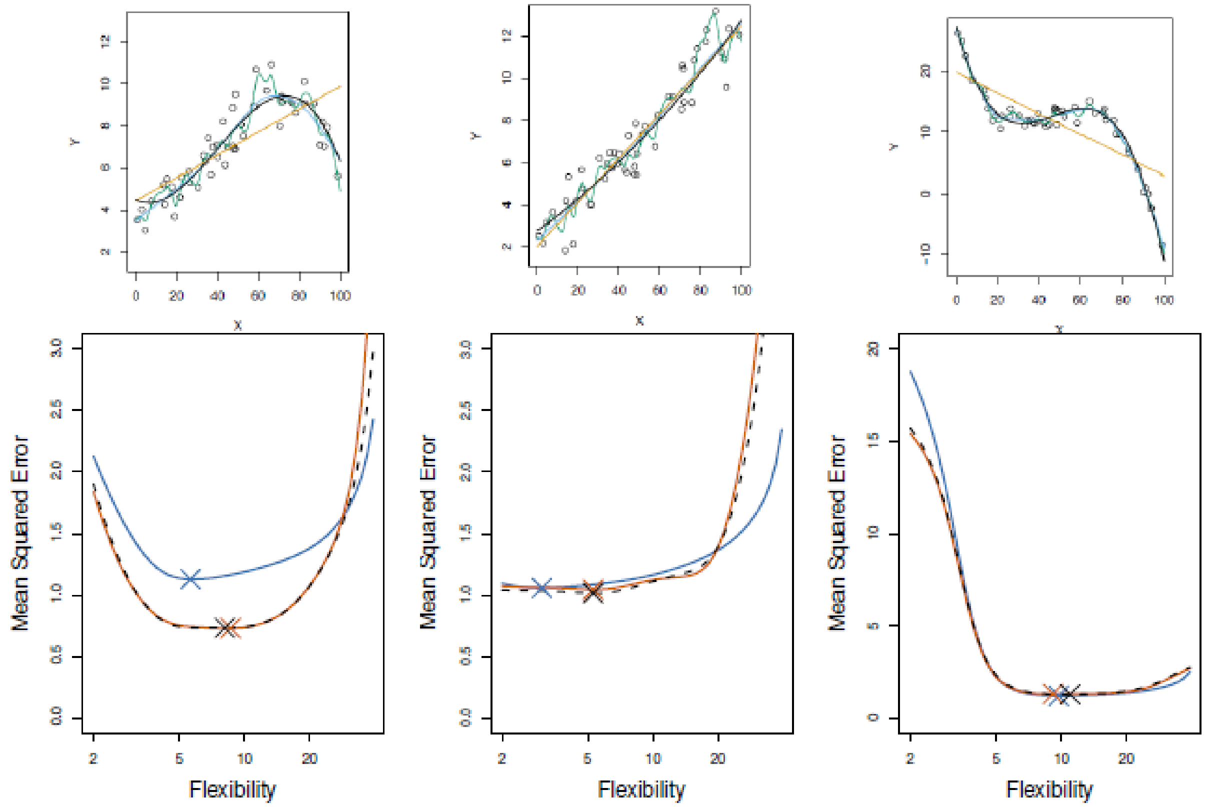
- **K-Fold Cross-Validation (K-Fold CV)** is most widely used approach to estimate test error.
- Idea:
  1. Randomly divide data into K equal parts.
  2. Leave out first part.
  3. Fit on the remaining K-1 parts combined into one set.
  4. Predict on left out part.
  5. Repeat in turn leaving out each part (1, 2, …, K) one part at a time.
  6. Average the K different errors to estimate the test error.



-

Example: K-Fold Cross-Validation

K-Fold Cross-Validation on Simulated Data



Bias-Variance Trade-Off for K-Fold CV

- **Bias in estimation of the test error:** Since each training set only uses a part of the sample, it tends to overestimate test error.

Bias-Variance Trade-Off for K-Fold CV

- **Variance in estimation of the test error:** If we were to have a different sample, how would our estimate of test error change.

Bias-Variance Trade-Off for K-Fold CV

-

- Work well in practice.

- Can reduce variance further by repeating K-Fold a number of times.
    - Split into K folds.
    - Average the errors.
    - Do another random split into K folds.
    - Repeat.
    - Average the averages.

Cross-Validation

- Cross-Validation is one resampling method, i.e. using subsets of the data.
- Typically we use 5 or 10 Fold CV as a way to choose from among different methods, or complexity.
- Next time, will finish our discussion of CV.

More next time!

Cross-Validation – Part II

Chapter 5 – Part II

Cross-Validation – Part II

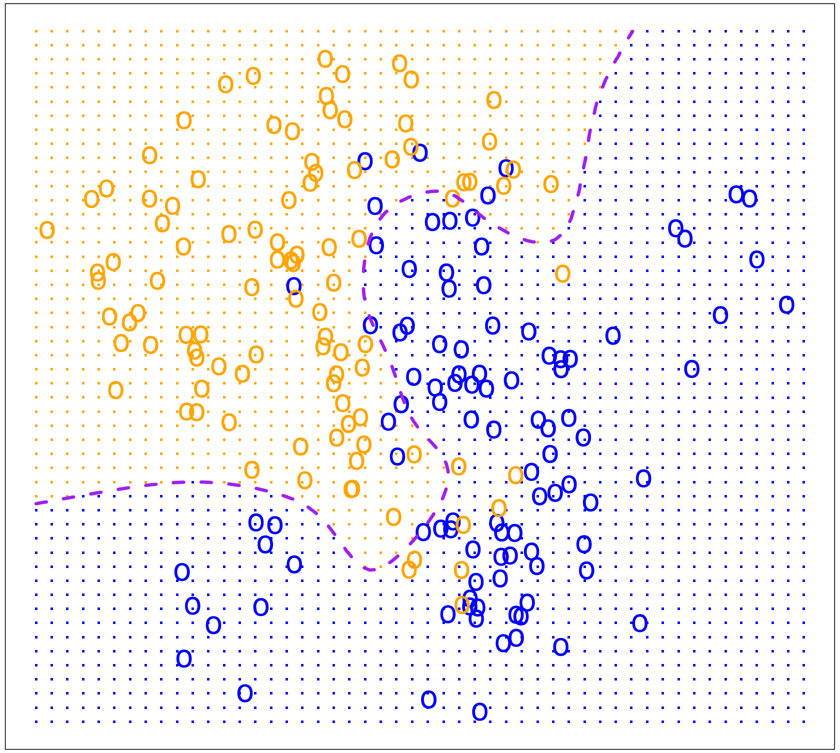
- CV for Classification Problems
- Right and Wrong Ways for CV

Cross-Validation for Classification Problems

- Can use K-Fold CV for classification problems.

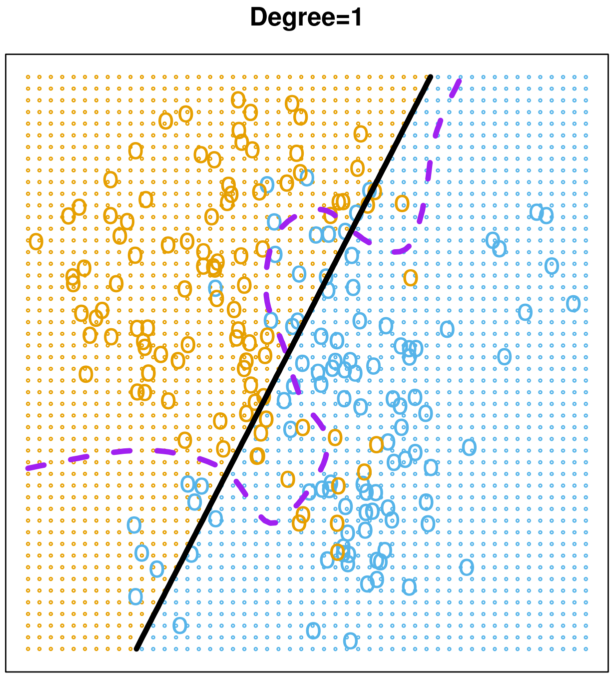
- Same idea.
  - Divide into K parts.
  - Hold out 1 part at a time.
  - Average the error rate across all K left out sets.

Example: Polynomials in Logistic Regression



Example: Polynomials in Logistic Regression

- Left: logistic regression fit.



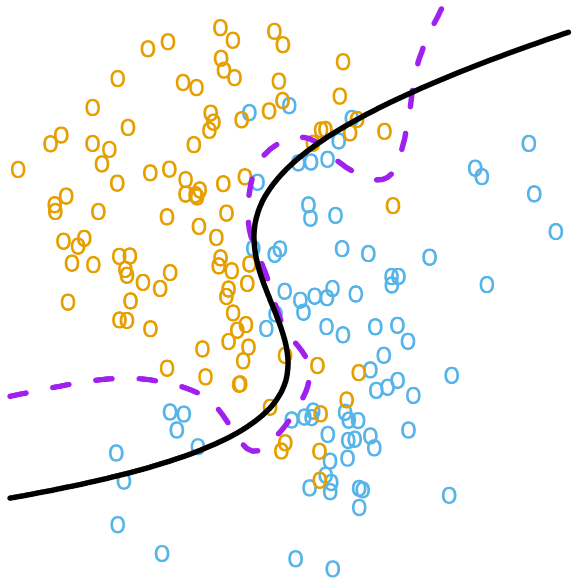
- Right: include quadratic terms in logistic regression.

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

Test Error Rate: 0.201

Test Error Rate: 0.197

Example: Polynomials in Logistic Regression

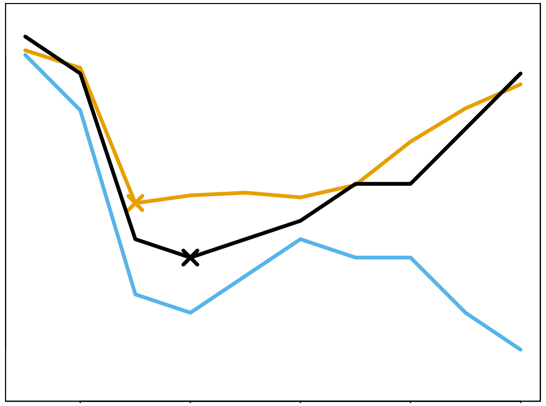


Test Error Rate: 0.160

Test Error Rate: 0.162

Example: Polynomials in Logistic Regression

- In practice, do not know truth.
  - Cannot compute test error.
- Use CV to choose order of polynomial.
- Also use KNN on this data.
  - Use CV to choose K.



Cross-Validation: Right or Wrong Way?

- Consider classification on a binary problem.
- Data consists of p = 5,000 predictors and only n = 50 observations.

Cross-Validation: Right or Wrong Way?

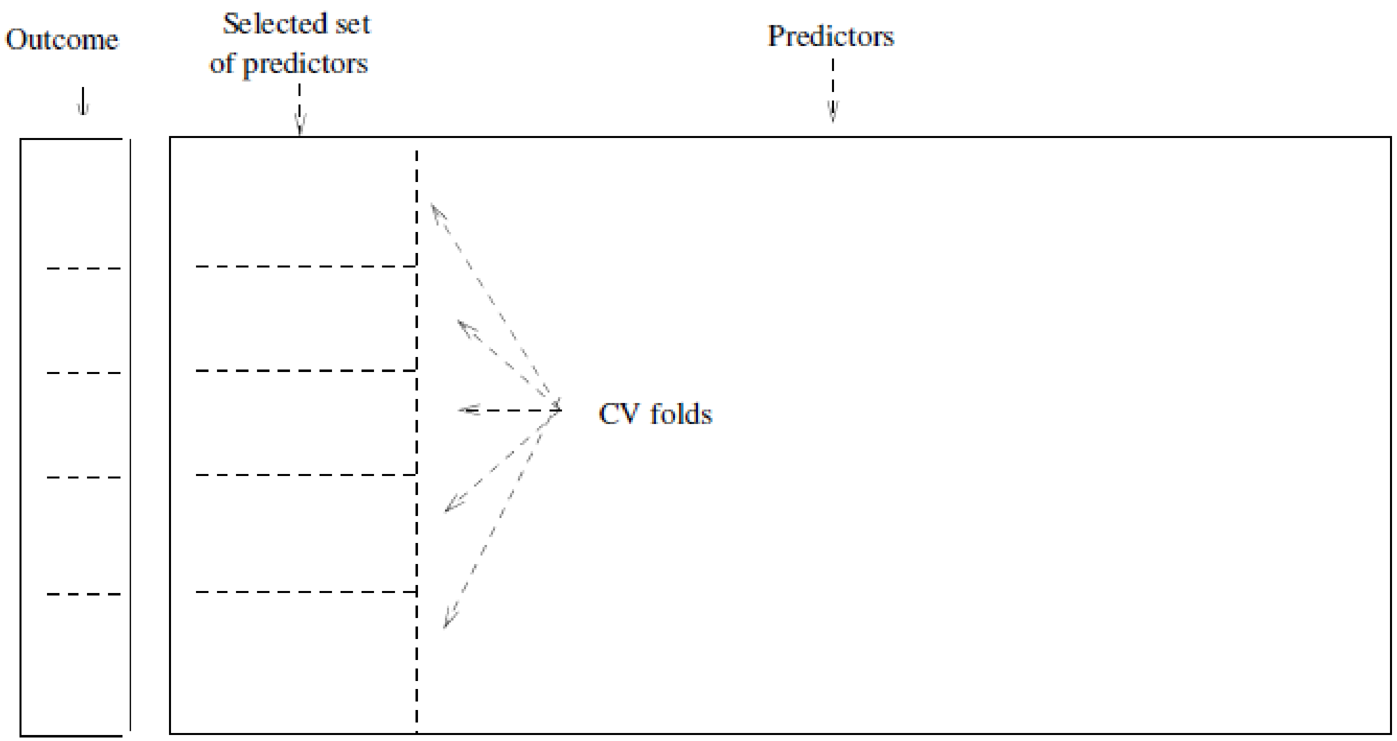
- Problem:

- With multi-step methods, must do CV on the outermost loop!

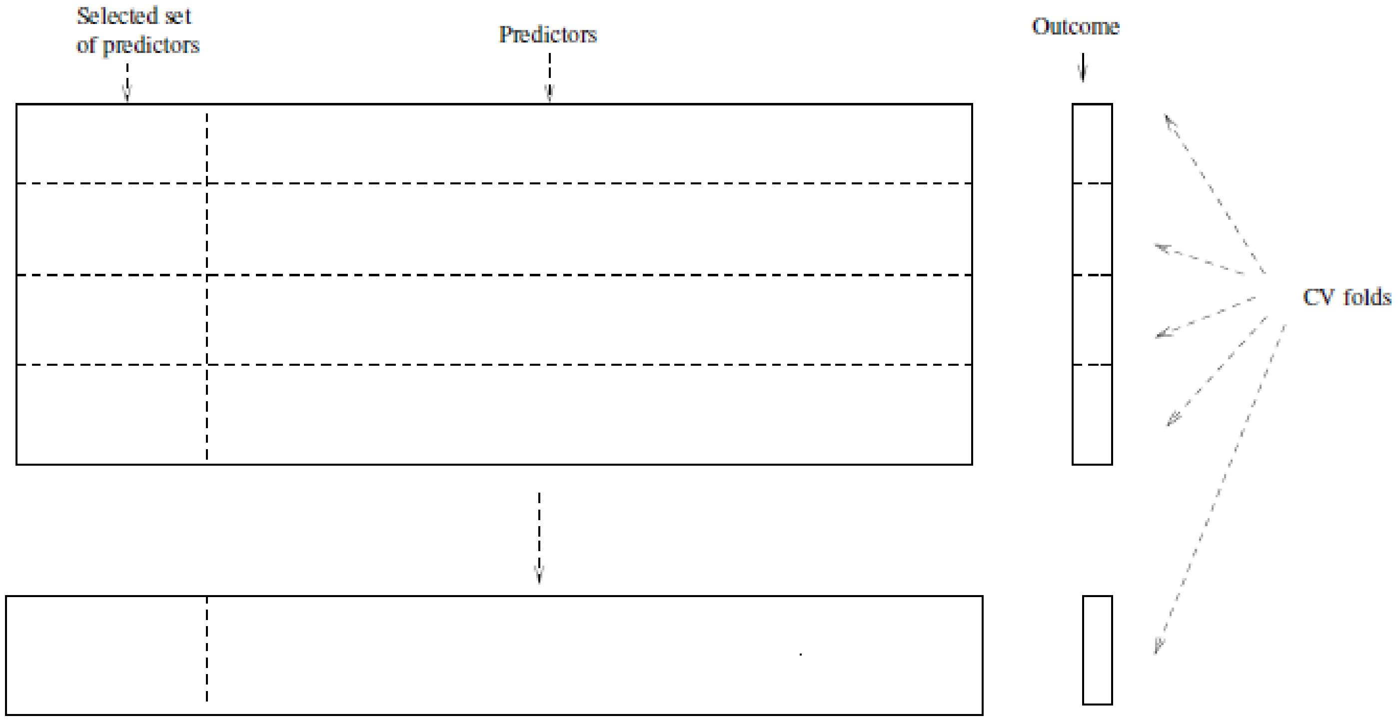
- All steps are part of the method, not just the final fitting procedure.

- Once final method is chosen, then apply method one last time, on FULL DATA.

Cross-Validation: Right or Wrong Way?



Cross-Validation: Right or Wrong Way?



Cross-Validation

- Cross-Validation is most common resampling method for selection among methods.
- Next time, will consider another resampling method, the Bootstrap.

More next time!