Regression

Linear, Logistic and LDA

Linear Regression

Linear Regression - Stepwise

- Is there a relationship between X and Y?
- How Strong is the relationship between X and Y?
  - Calculate correlation
- Are there any control factors?
  - Separate out effect of any controlling factors by running correlation of homogeneous group.
- How accurately can we estimate Y using X?
  - Find the Coefficients (Betas) of each X.

Linear Regression - Stepwise

- How accurately can we predict future Ys?
  - Fitting the existing data vs generalizing the model.
  - Test sensitivity of the model.
- Is the relationship Linear?
  - Scatter plot of X and Y.
- Are X independent? Is there an interaction effect?
  - PCA and multi-collinearity tests.

Terminology

- $Y \approx \beta_0 + \beta_1 X.$
  - $\approx$ is read as “is approximately modeled as”.
  - We are regressing Y on X.
- β_0 & β_1 are referred as intercept and slope. Generally, as coefficients or parameters. These are unknown constants we are finding.
  - $\widehat{\beta_0}\ and\ \widehat{\beta_1}$ are the estimates we produce using data through modeling.
- $\hat{y} = \widehat{\beta_0} + \widehat{\beta_1}x$ where $\hat{y}$ are the predicted values of Y using X = x.
- $\hat{y}_i = \widehat{\beta_0} + \widehat{\beta_1}x_i$ gives us the predicted value of y using x.

Terminology

- $y_i - \hat{y}_i = e_i$ is the ith residual.
- ESS is error sum of squares given by ESS = $e_1^2 + e_2^2 + \cdots + e_n^2$
- We minimize ESS to find optimal coefficients of ($\widehat{\beta_0}\ and\ \widehat{\beta_1}$) Betas.
- Assessing accuracy of coefficient.
  - $Y \ \approx \beta_0 + \beta_1 X + \in$ is the population regression line
  - Where ∈ is the mean zero random error term.
  - Test Hypothesis Ho: There is no relationship between X and Y i.e. $\beta_1 = 0$
  - We use t-test to test accuracy of coefficients.
- Assessing accuracy of the model
  - RSE and R^2
  - RSE is an estimate of the standard deviation of ∈. i.e. average amount that Y deviate from true regression line. RSE is considered as lack of fit of the model. RSE = $\sqrt{\frac{ESS}{n-2}}$

Logistic Regression

Classification: A special class of Regression

- When the predicted variable Y is qualitative, we refer to it as classification.
- The methods used for classification first predict the probability of each class.
- We will discuss logistic regression for classification.

Why Not Linear Regression

- In case predicted variable has three classes, say coded as 1, 2, 3, then the coding implies ordering and different orders will produce different linear models.
- On the other hand if the classes are rank ordered (Ordinal) then the model treats difference in ranks as equidistant.
- In case of binary variables, then linear regression model can be fitted with a few out of range values making it hard to interpret.
  - Interestingly, coefficient of predictor variables derived using least squares in fact estimate the probabilities of the class.

Logistic Regression

- Rather than modeling Y directly, Logistic regression models the probability that Y belongs to a particular category i.e. P(Y = 1|X = x). This probability is abbreviated as P(X = x).
- We can then use P(X = x) > 0.5 to predict Y = 1 for an individual. The choice of 0.5 can be chosen subjectively.
- Logistic Model
  - We model relation between p(X) and X.
  - P(X) = $\beta_0 + \beta_1 X$ is our classic linear regression model with issues.
  - Logistic function $p(X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$

Logistic Regression

- The RHS values range between 0 and 1 regardless of X.
- Logistic function is a S shaped curved also called as sigmoid.
- We can algebraically rearrange the logistic function as
  - $\frac{p(X)}{1-p(X)} = e^{\beta_0+\beta_1 X}$
  - LHS is odds with value ranging between 0 and infinity.
  - We can then use P(X = x) > 0.5 to predict Y = 1 for an individual. The choice of 0.5 can be chosen subjectively.
- We can further algebraically rearrange the above function as
- $log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X$ where LHS is log odds or logit.

Logistic Regression

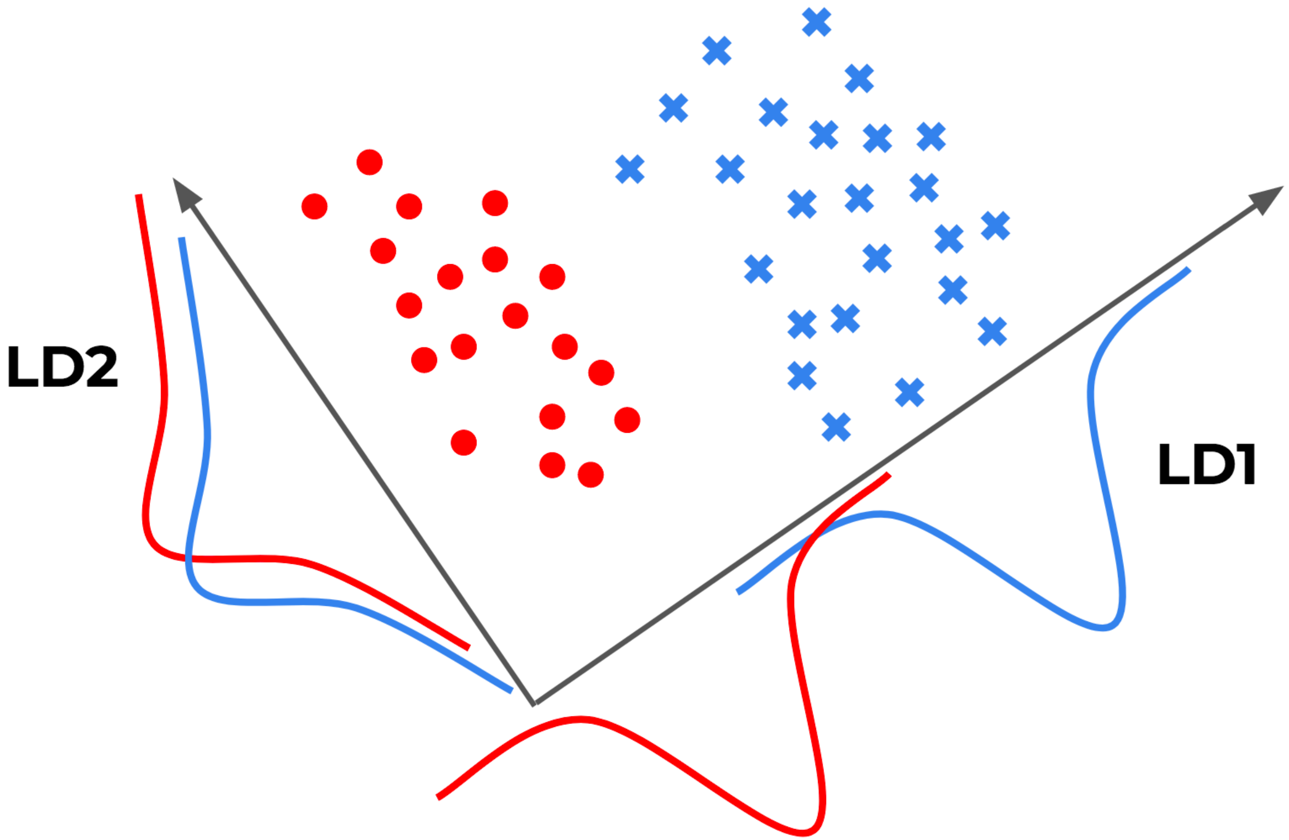
- β_1 represents the rate of change of log – odds of Y = 1 per unit change in X.
- Equivalently, it multiplies the odds by e^{β_1}
- The relationship between p(X) and X is not a straight line.
- β_1 does not correspond to the change of p(X) associated with unit change in X.
- The p(X) changes due to one-unit change in X will depend on the current value of X.
- We fit seek to estimate β_0 and $\beta_1$such that predicted probability $\hat{p}(X)$ for each individual corresponds as closely as possible to observed Y=1

Linear Discriminant Analysis

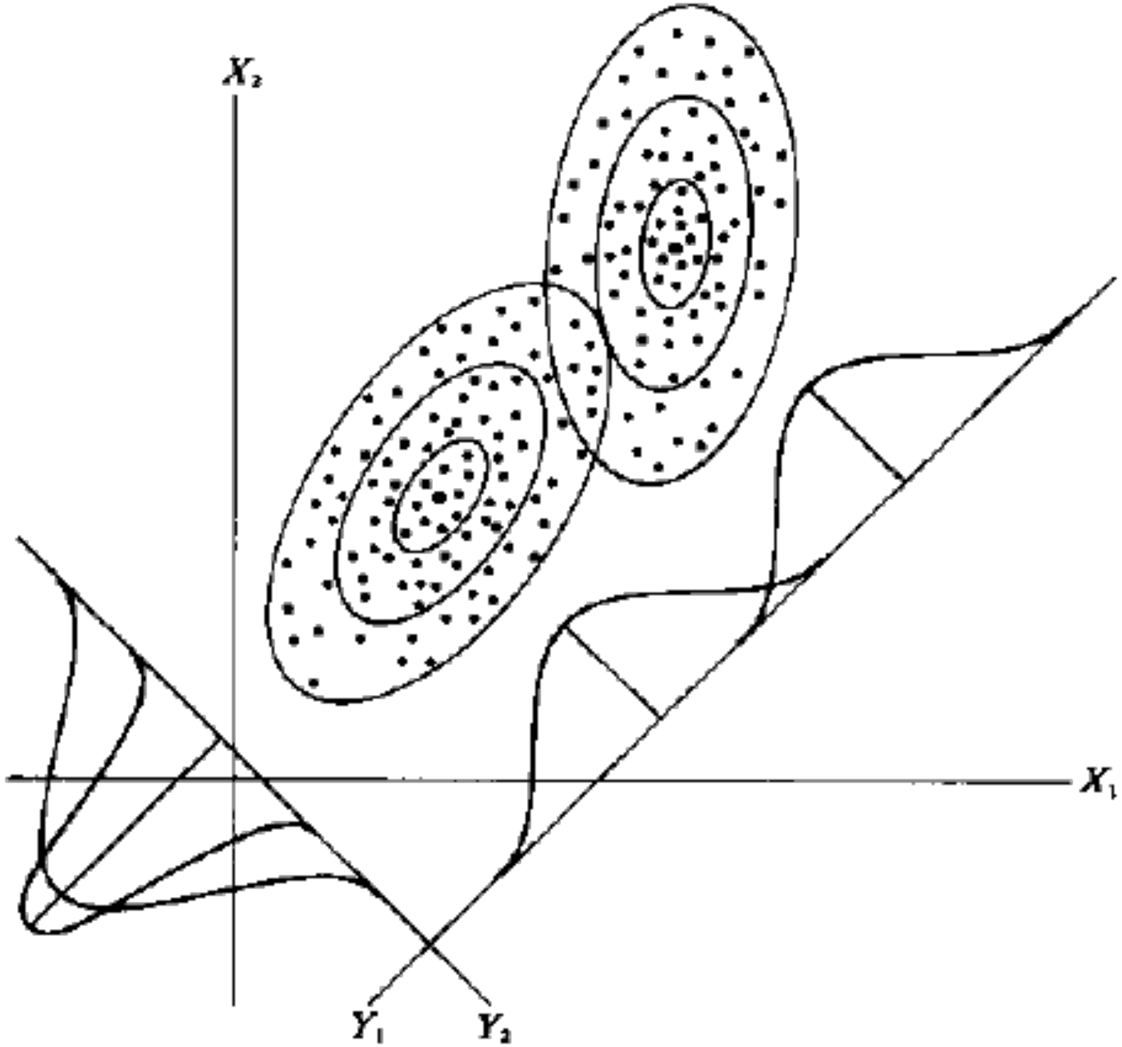
Linear Discriminant Analysis (LDA)

- We model the distribution of the predictors X separately in each of the response classes Y and then use Bayes theorem to flip these around into estimates of P(Y=k|X = x).
- Why LDA
  - When classes are well separated, parameter estimates from LDA is more stable than that from Logit.
  - If n is small and distribution of the predictors X is approximately normal for each of the classes, then again LDA is more stable.
  - LDA is popular when we have more than two response classes.

LDA Intuition



LDA Intuition…



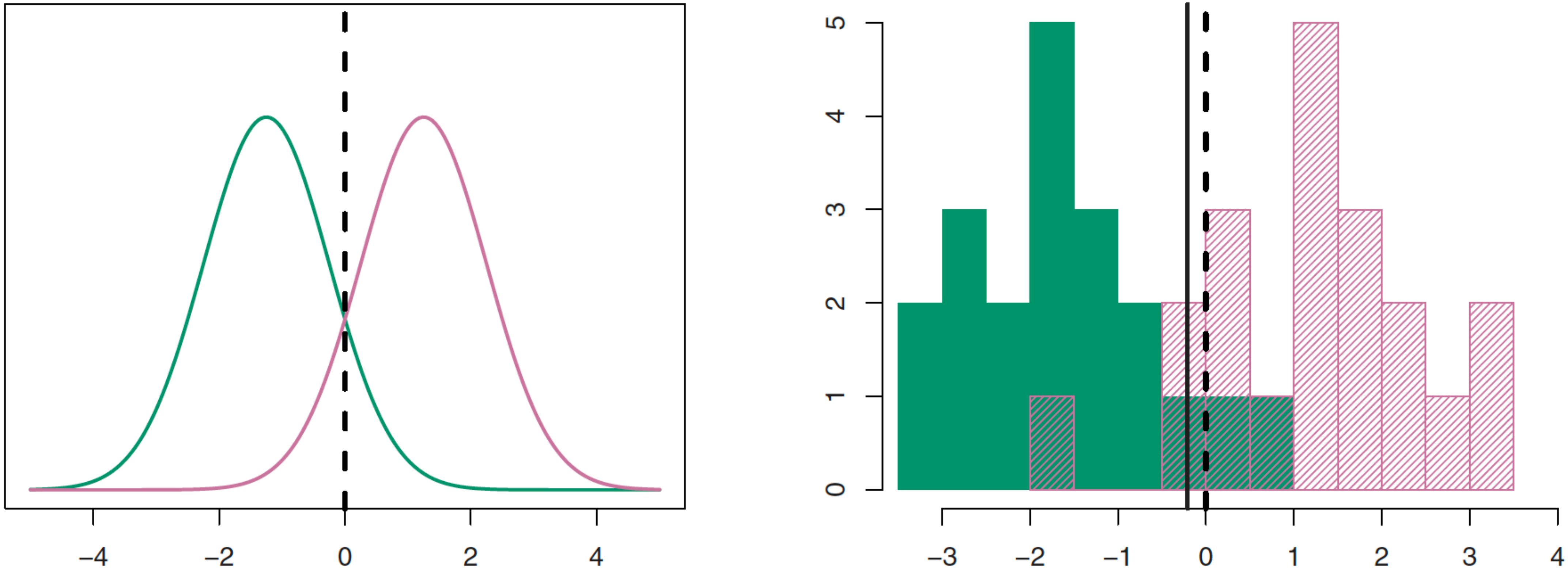
Role of Bayes Theorem

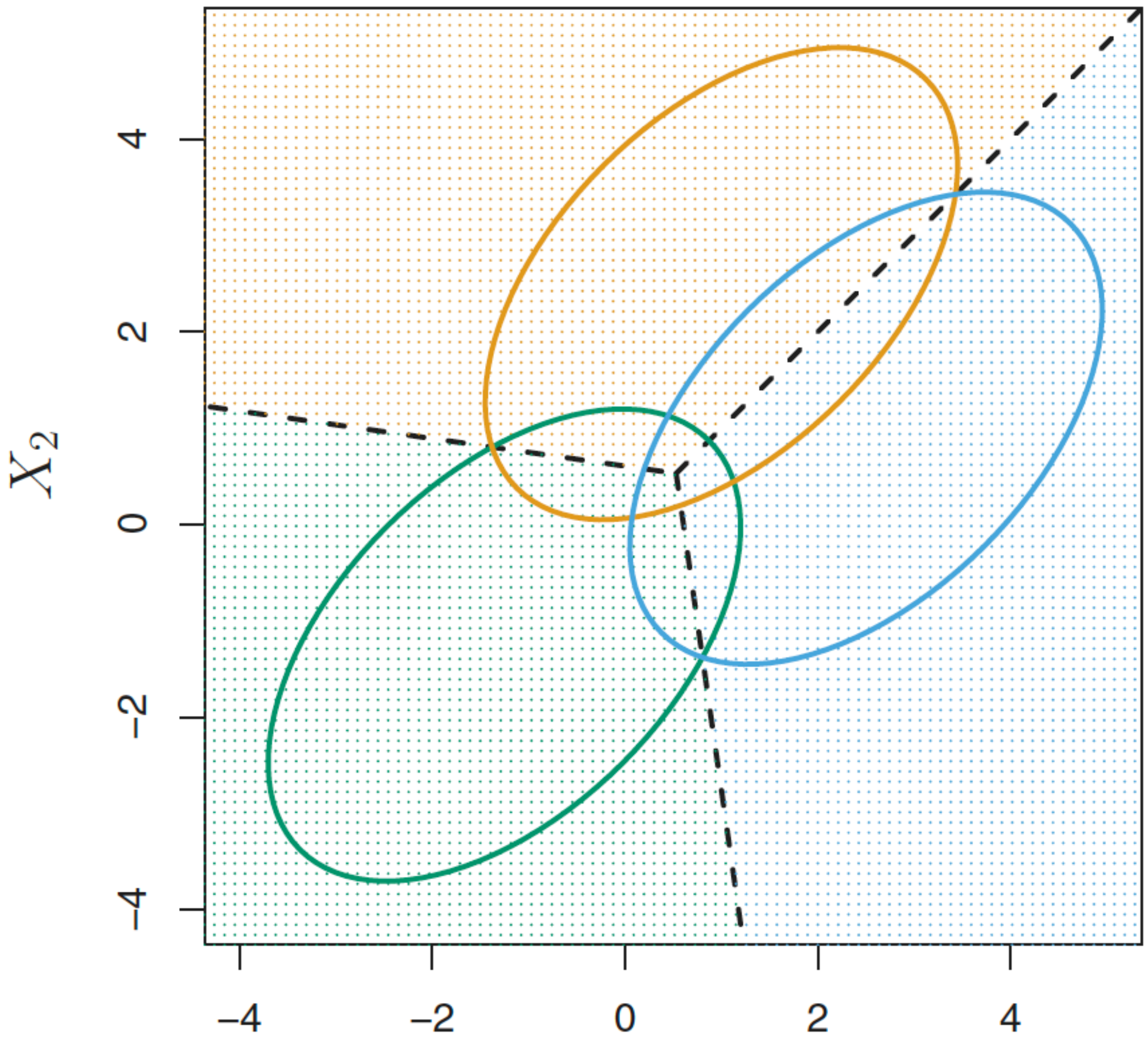
- P(A|B)*P(B) = P(B|A)*P(A)
- Let l_k represent the prior probability that a randomly chosen observation comes from the kth class.
- Let $f_k(X)$ be the density function P(X=x|Y=k) of X given kth class. So
- $p(X) = \; P(Y = k \,| X = x) = {}^{l_k * f_k(X)}/_{\sum l * f(X)}$ are posterior probabilities.
- Computing l_k is easy if Y is randomly distributed. Just find the proportion of kth class in all the classes.
- Bayes classifier has the lowest error rate among all classifiers.
- Estimating $f_k(X)$ is challenging here.

LDA for a single predictor case

- We assume that $f_k(X)$ is normally distributed.
- We assume that there is a shared variance term across all K classes.
- Now we assign X = x to a class for which p(X) is highest.
- For instance, if K =2 and observations are equally distributed on both the classes then the decision boundary is $\frac{\mu_1 + \mu_2}{2}$ for classifying an observation as Class 1 or Class 2.
- LDA approximates the Bayes Classifier by plugging estimates of $f_k(X)$, mean of density function of each class and shared common SD.
- The word linear stems from the fact that classifier function is linear.

LDA for single predictor case

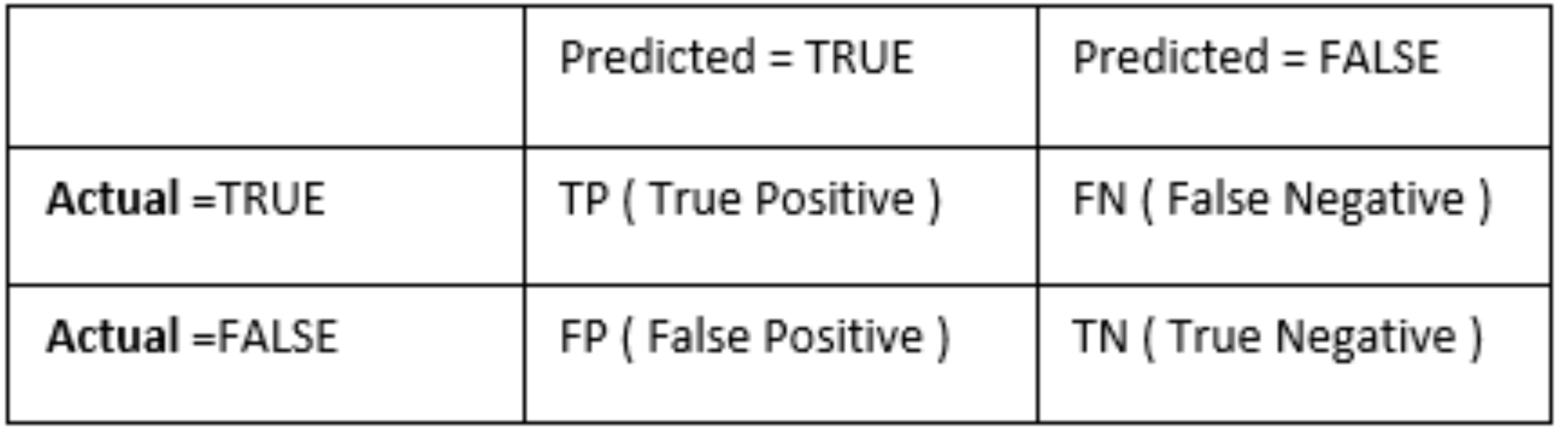




LDA for two predictor case

We use cutting planes instead of line as classifier

Evaluating LDA: Confusion Matrix



	Predicted = TRUE	Predicted = FALSE
Actual =TRUE	TP (True Positive)	FN (False Negative)
Actual =FALSE	FP (False Positive)	TN (True Negative)

Confusion Matrix Example (Boundary: 50%)

		True Default		
		No	Yes	Total
Predicted Default	No	9644	252	9896
	Yes	23	81	104
	Total	9667	333	10,000

- 23 out of 9667 (0.24%) were incorrectly identified as defaulted.
- But 252 out of 333 (75.7%) were missed by LDA.
- Accuracy is High at (97.25%)
- Sensitivity is 24.3% (Identified true defaulters)
- Specificity is 99.76% (Identified non-defaulters.)

Confusion Matrix Example (Boundary: 20%)

		True Default		
		No	Yes	Total
Predicted Default	No	9432	138	9570
	Yes	235	195	430
	Total	9667	333	10,000

- 235 out of 9667 (3.73%) were incorrectly identified as defaulted.
- But 138 out of 333 (41.4%) were missed by LDA.
- Accuracy is still high at (95.7%)
- Sensitivity is 58.6% (Identifying true defaulters)
- Specificity is 97.7% (Identifying non-defaulters.)

What is the right threshold

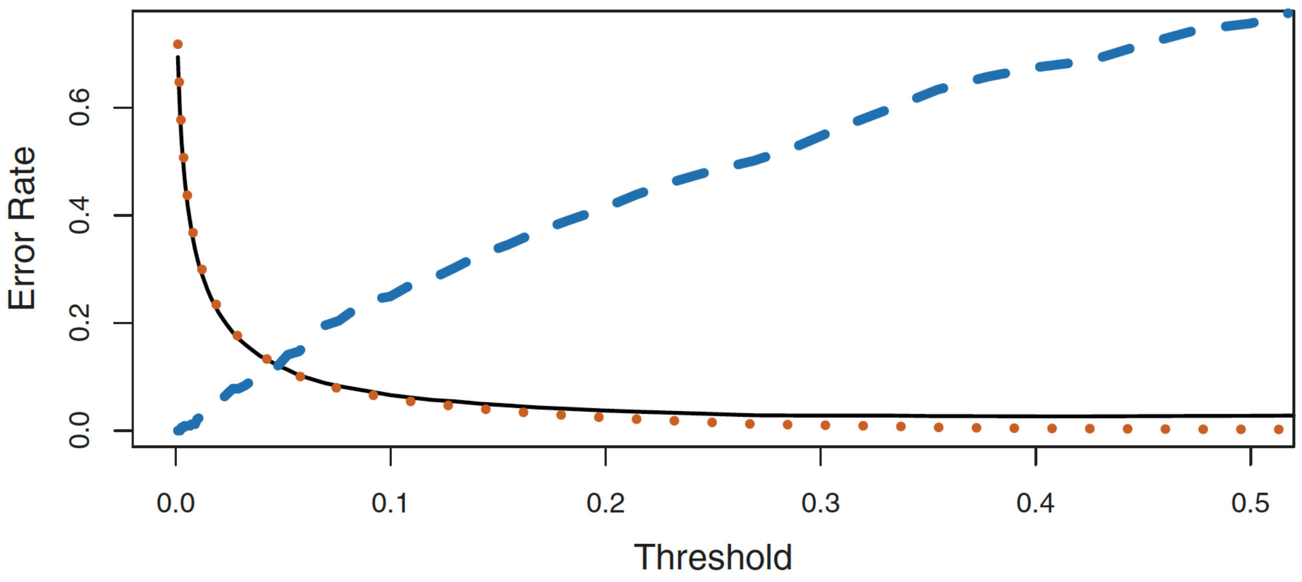


FIGURE 4.7. *For the* `Default` *data set, error rates are shown as a function of the threshold value for the posterior probability that is used to perform the assignment. The black solid line displays the overall error rate. The blue dashed line represents the fraction of defaulting customers that are incorrectly classified, and the orange dotted line indicates the fraction of errors among the non-defaulting customers.*