STEVENS INSTITUTE of TECHNOLOGY
School of Business

FE 582

Lecture 5: Classification Methods

Dragos Bozdog

For academic use only.

Classification Methods

Logistic Regression

Linear Discriminant Analysis (LDA)

Quadratic Discriminant Analysis (QDA)

K-Nearest Neighbors (KNN)

Logistic Regression

Why Not Linear Regression?

Simple Logistic Regression

- Logistic Function
- Interpreting the coefficients
- Making Predictions
- Adding Qualitative Predictors

Multiple Logistic Regression

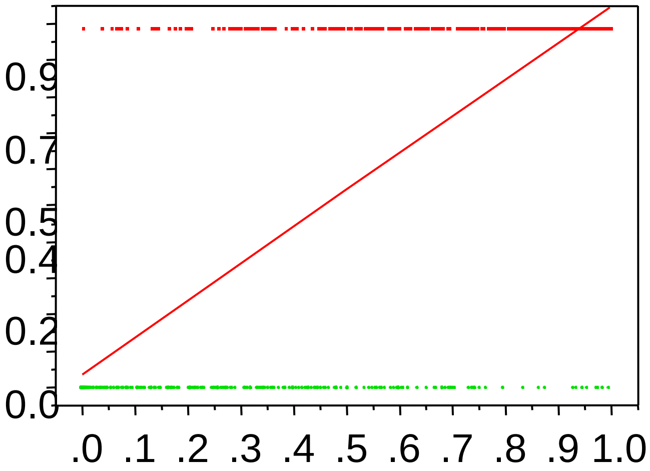
Case Study:

- Credit Card Default Data

Logistic Regression

- The Y variable is categorical: 0 or 1
- The X variable is a numerical value (between 0 and 1)

Can we use Linear Regression when Y is categorical?



Why standard linear regression in inappropriate?

- The regression line $\beta_0 + \beta_1 X$ can take on any value between negative and positive infinity
- Therefore the regression line almost always predicts the wrong value for Y in classification problems

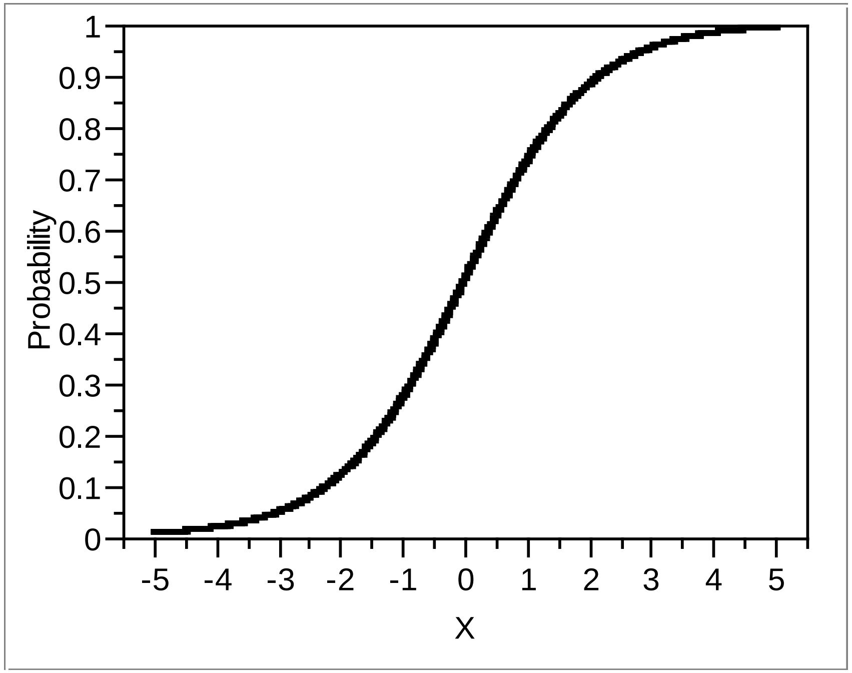
Logistic Regression

Solution: Use Logistic Function

- Instead of trying to predict Y, let’s try to predict P(Y = 1)

- Thus, we can model P(Y = 1) using a function that gives outputs between 0 and 1.

- $p = P(Y = 1) = \frac{e^{\beta_0 + \beta_1\ X}}{1 + e^{\beta_0 + \beta_1\ X}}$

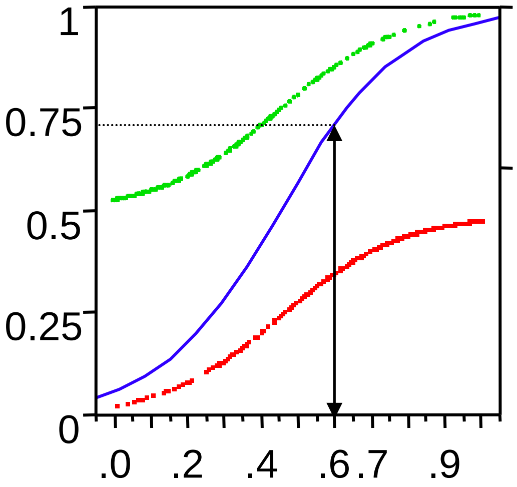


Logistic Regression

Logistic regression is very similar to linear regression

- We come up with b_0 and b_1 to estimate β_0 and β_1.
- We have similar problems and questions as in linear regression

e.g. Is β_1 equal to 0? How sure are we about our guesses for β_0 and β_1?



- If X is about 0.6, then $P(Y) \approx 0.7$

Logistic Regression

Case Study: Credit Card Default Data

- We would like to be able to predict customers that are likely to default

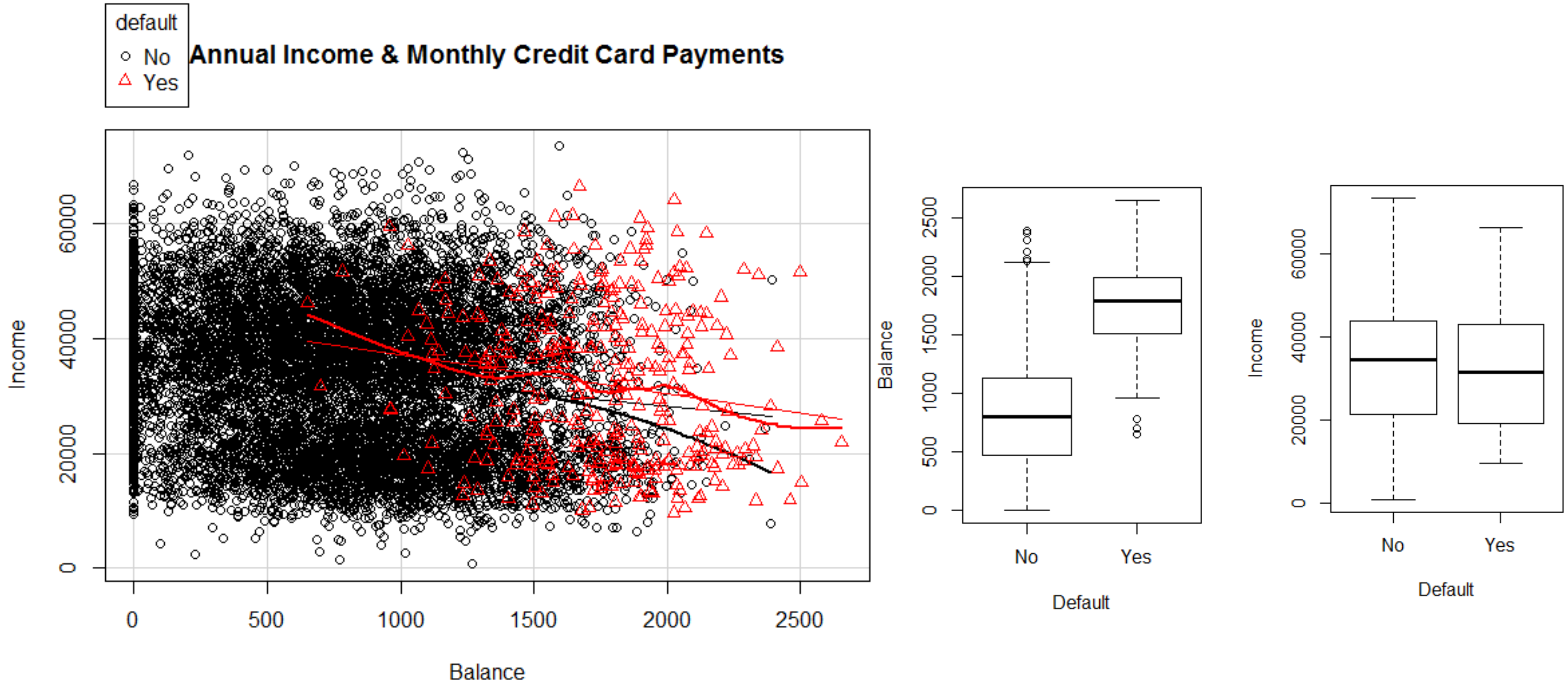
Possible X variables are:

- Annual Income
- Monthly credit card balance

The Y variable (Default) is categorical: Yes or No

How do we check the relationship between Y and X?

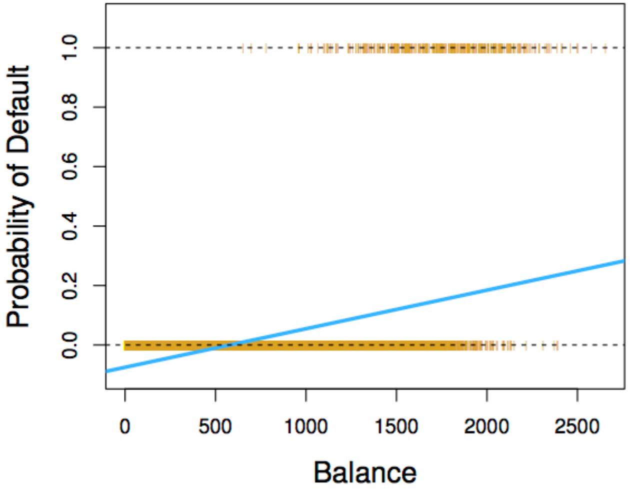
Logistic Regression



Logistic Regression

Why not Linear Regression?

- If we fit a linear regression to the Default data, then for very low balances we predict a negative probability, and for high balances we predict a probability above 1!

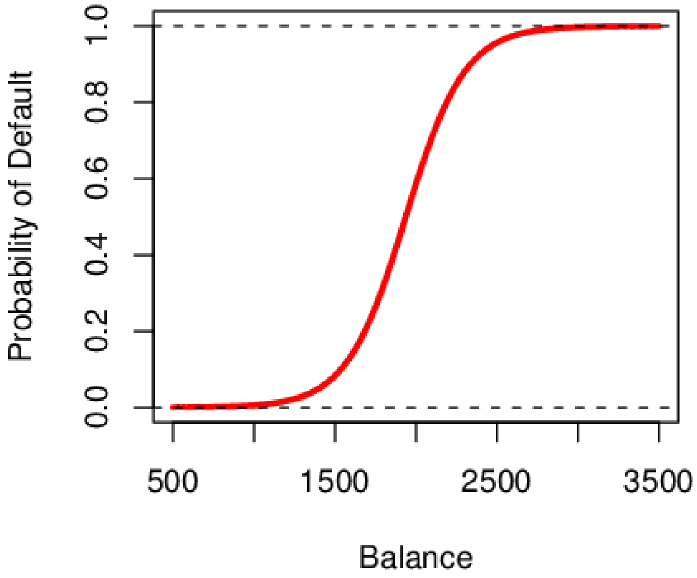


- When Balance < 500, P(default) is negative!

Logistic Regression

Logistic Function on Default Data

- Now the probability of default is close to, but not less than zero for low balances. And close to but not above 1 for high balances



Logistic Regression

Interpreting β_1

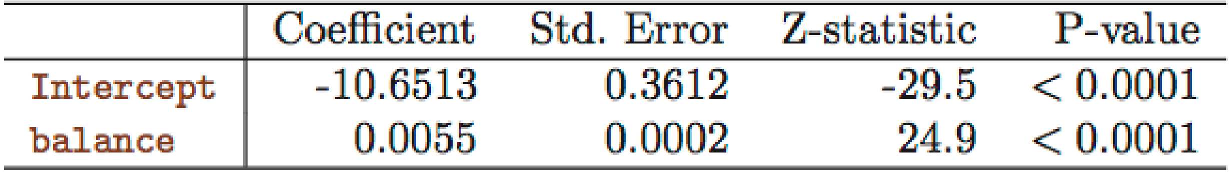
Interpreting what β_1 means is not very easy with logistic regression, simply because we are predicting P(Y) and not Y.

- If β_1 =0, this means that there is no relationship between Y and X.
- If β_1 >0, this means that when X gets larger so does the probability that Y = 1.
- If β_1 <0, this means that when X gets larger, the probability that Y = 1 gets smaller. But how much bigger or smaller depends on where we are on the slope

Logistic Regression

Are the coefficients significant?

- We still want to perform a hypothesis test to see whether we can be sure that are β_0 and β_1 significantly different from zero.
- We use a Z test instead of a T test, but of course that doesn't change the way we interpret the p-value
- Here the p-value for balance is very small, and b_1 is positive, so we are sure that if the balance increase, then the probability of default will increase as well.



	Coefficient	Std. Error	Z-statistic	P-value
Intercept	-10.6513	0.3612	-29.5	< 0.0001
balance	0.0055	0.0002	24.9	< 0.0001

Logistic Regression

Making Prediction

- Suppose an individual has an average balance of $1000. What is their probability of default?

$$\hat{p} = P(Y = 1) = \frac{e^{\widehat{\beta_0} \ \ + \widehat{\beta_1}X}}{1 + e^{\widehat{\beta_0} + \widehat{\beta_1}X}} = \frac{e^{-10.6513 + 0.0055 \times 1000}}{1 + e^{-10.6513 + 0.0055 \times 1000}} = 0.00567$$

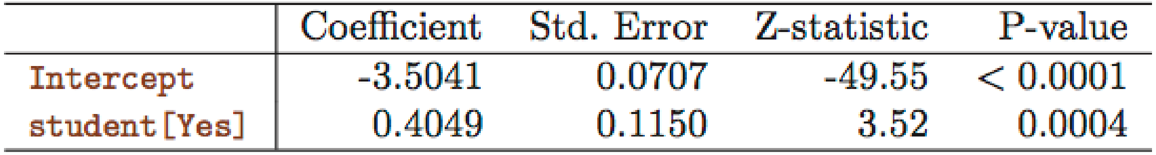
- The predicted probability of default for an individual with a balance of $1000 is less than 1%.

- For a balance of $2000, the probability is much higher, and equals to 0.586 (58.6%).

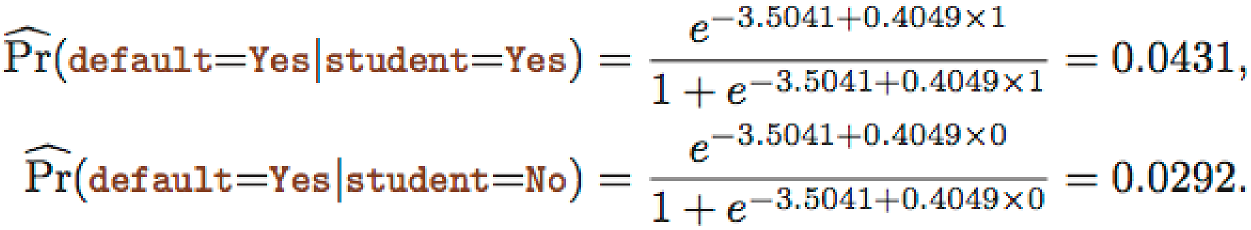
Logistic Regression

Qualitative Predictors in Logistic Regression

- We can predict if an individual default by checking if she is a student or not. Thus we can use a qualitative variable “Student” coded as (Student = 1, Non-student =0).
- b_1 is positive: This indicates students tend to have higher default probabilities than non-students



	Coefficient	Std. Error	Z-statistic	P-value
Intercept	-3.5041	0.0707	-49.55	< 0.0001
student[Yes]	0.4049	0.1150	3.52	0.0004



$$\widehat{\Pr}(\texttt{default=Yes}|\texttt{student=Yes}) = \frac{e^{-3.5041+0.4049\times 1}}{1+e^{-3.5041+0.4049\times 1}} = 0.0431,$$

$$\widehat{\Pr}(\texttt{default=Yes}|\texttt{student=No}) = \frac{e^{-3.5041+0.4049\times 0}}{1+e^{-3.5041+0.4049\times 0}} = 0.0292.$$

Logistic Regression

Multiple Logistic Regression

- We can fit multiple logistic just like regular regression

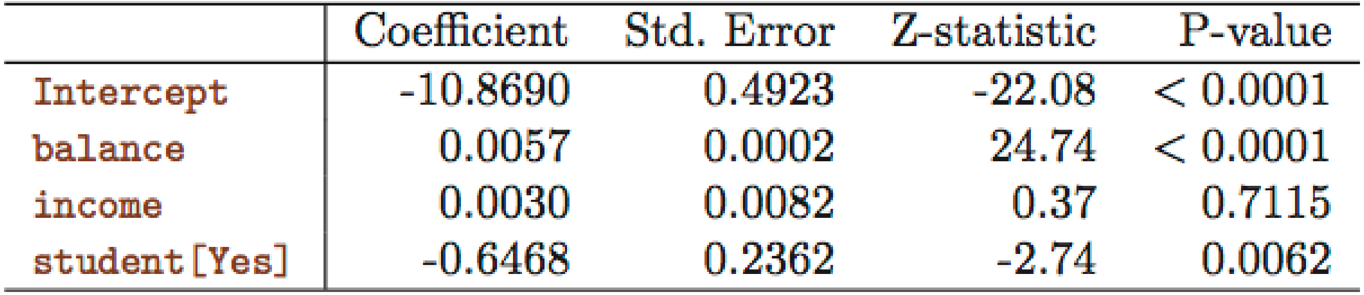
$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}.$$

Logistic Regression

Multiple Logistic Regression- Default Data

Predict Default using:

- Balance (quantitative)
- Income (quantitative)
- Student (qualitative)



	Coefficient	Std. Error	Z-statistic	P-value
Intercept	-10.8690	0.4923	-22.08	< 0.0001
balance	0.0057	0.0002	24.74	< 0.0001
income	0.0030	0.0082	0.37	0.7115
student[Yes]	-0.6468	0.2362	-2.74	0.0062

Logistic Regression

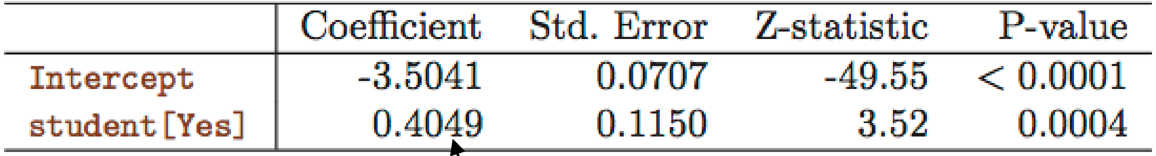
Predicting Default using:

- A student with a credit card balance of $1,500 and an income of $40,000 has an estimated probability of default

$$\hat{p}(X) = \frac{e^{-10.869+0.00574\times 1500+0.003\times 40-0.6468\times 1}}{1+e^{-10.869+0.00574\times 1500+0.003\times 40-0.6468\times 1}} = 0.058.$$

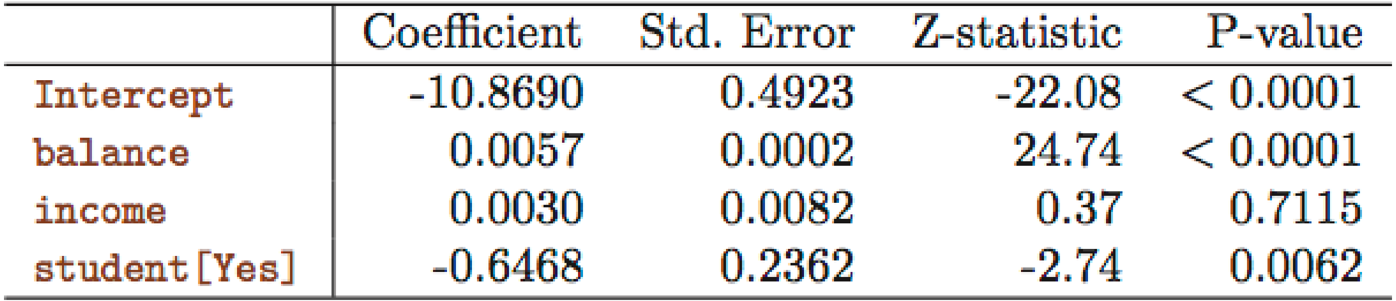
Logistic Regression

An Apparent Contradiction!



	Coefficient	Std. Error	Z-statistic	P-value
Intercept	-3.5041	0.0707	-49.55	< 0.0001
student[Yes]	0.4049	0.1150	3.52	0.0004

Positive

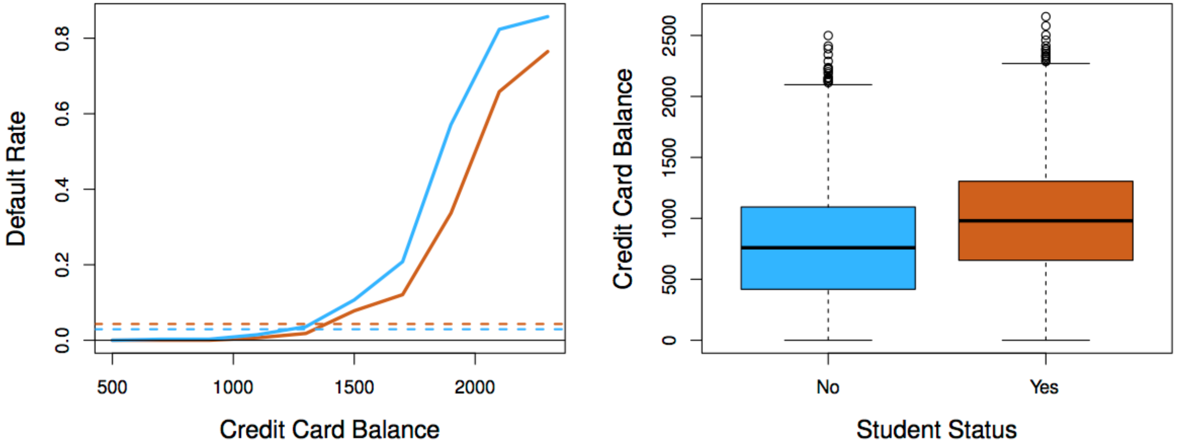


	Coefficient	Std. Error	Z-statistic	P-value
Intercept	-10.8690	0.4923	-22.08	< 0.0001
balance	0.0057	0.0002	24.74	< 0.0001
income	0.0030	0.0082	0.37	0.7115
student[Yes]	-0.6468	0.2362	-2.74	0.0062

Negative

Logistic Regression

Students (Orange) vs. Non-students (Blue)



Logistic Regression

To whom should credit be offered?

- A student is risker than non students if no information about the credit card balance is available

- However, that student is less risky than a non student with the same credit card balance!

Logistic Regression

Lab: The Stock Market Data

- Use Smarket data, which is part of ISLR library

Linear Discriminant Analysis (LDA)

Outline

- Overview of LDA
- Why not Logistic Regression?
- Estimating Bayes' Classifier
- LDA Example with One Predictor (p=1)
- LDA Example with more than One Predictor (P>1)
- LDA on Default Data
- Overview of QDA
- Comparison between LDA and QDA

Linear Discriminant Analysis (LDA)

LDA undertakes the same task as Logistic Regression. It classifies data based on categorical variables

- Making profit or not
- Buy a product or not
- Satisfied customer or not
- Political party voting intention

Linear Discriminant Analysis (LDA)

Why Linear? Why Discriminant?

- LDA involves the determination of linear equation (just like linear regression) that will predict which group the case belongs to.

$$\boxed{D = v_1 X_1 + v_2 X_2 + \ldots + v_i X_i + a}$$

D: discriminant function
v: discriminant coefficient or weight for the variable
X: variable
a: constant

Linear Discriminant Analysis (LDA)

- Choose the v’s in a way to maximize the distance between the means of different categories

- Good predictors tend to have large v’s (weight)

- We want to discriminate between the different categories

- Ex: Portfolio. By changing the proportions (weights) of the constituents, the characteristics of the portfolio will change.

Linear Discriminant Analysis (LDA)

Assumptions of LDA

- The observations are a random sample
- Each predictor variable is normally distributed

Why not Logistic Regression?

- Logistic regression is unstable when the classes are well separated
- In the case where n is small, and the distribution of predictors X is approximately normal, then LDA is more stable than Logistic Regression
- LDA is more popular when we have more than two response classes

Linear Discriminant Analysis (LDA)

Bayes' Classifier

- Bayes' classifier is the golden standard. Unfortunately, it is unattainable.

We can estimate it with two methods:

- KNN classifier
- Logistic Regression

Linear Discriminant Analysis (LDA)

Estimating Bayes' Classifier

- With Logistic Regression we modeled the probability of Y being from the kth class as

$$p(X) = P(Y = k, X = x) = \frac{e^{\beta_0+\beta_1\ X}}{1 + e^{\beta_0+\beta_1\ X}}$$

However, Bayes' Theorem states

$$p(X) = P(Y = k, X = x) = \frac{\Pi_k f_k(x)}{\sum_{i=1}^{k} \Pi_i f_i(x)}$$

Π_k is probability of coming from class k (prior probability)
$f_k(x)$: density function for X given that X is an observation from class k

Linear Discriminant Analysis (LDA)

Estimate Π_k and $f_k(x)$

- We can estimate Π_k and $f_k(x)$ to compute $p(x)$

- The most common model for $f_k(x)$ is the Normal Density

$$f_k(x) = \frac{1}{\sqrt{2\Pi}\sigma_k} e^{\left(-\frac{1}{2\sigma_k^2}(x-\mu_k)^2\right)}$$

- Using the density, we only need to estimate three quantities to compute $p(x)$

$$\mu_k \quad \sigma_k^2 \quad \Pi_k$$

Linear Discriminant Analysis (LDA)

Use Training Data set for Estimation

- The mean μ_k could be estimated by the average of all training observations from the kth class.

- The variance σ_k^2 could be estimated as the weighted average of variances of all k classes.

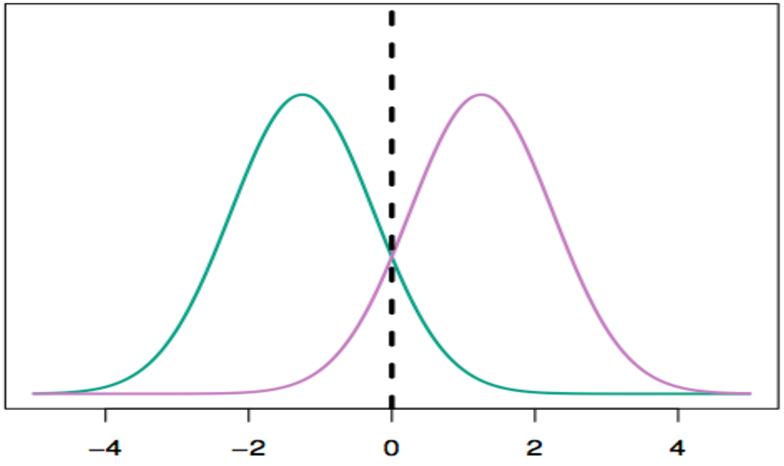
- And, Π_k is estimated as the proportion of the training observations that belong to the kth class.

- $\widehat{\mu_k} = \frac{1}{n_k} \sum_{i:y_i=k} x_i$
- $\hat{\sigma} = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i:\, y_i=k} (x_i - \widehat{\mu_k})^2$
- $\widehat{\pi_k} = \frac{n_k}{n}$

Linear Discriminant Analysis (LDA)

A Simple Example with One Predictor (p =1)

- Suppose we have only one predictor (p = 1)
- Two normal density function $f_1(x)$ and $f_2(x)$, represent two distinct classes
- The two density functions overlap, so there is some uncertainty about the class to which an observation with an unknown class belongs
- The dashed vertical line represents Bayes' decision boundary



Linear Discriminant Analysis (LDA)

Apply LDA

- LDA starts by assuming that each class has a normal distribution with a common variance
- The mean and the variance are estimated
- Finally, Bayes' theorem is used to compute p_k and the observation is assigned to the class with the maximum probability among all k probabilities

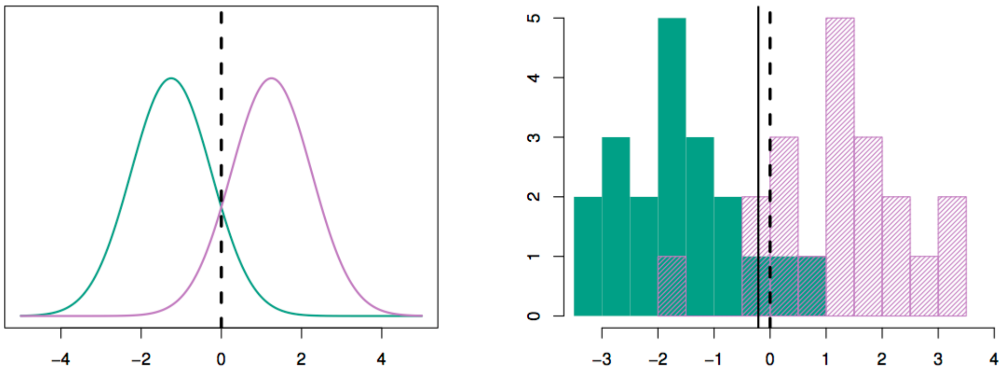
Linear Discriminant Analysis (LDA)

- 20 observations were drawn from each of the two classes
- The dashed vertical line is the Bayes' decision boundary
- The solid vertical line is the LDA decision boundary

  Bayes' error rate: 10.6%

  LDA error rate: 11.1%

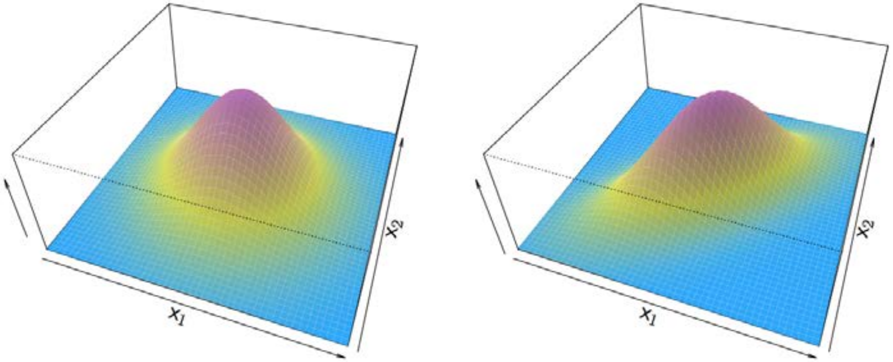
Thus, LDA is performing well!



Linear Discriminant Analysis (LDA)

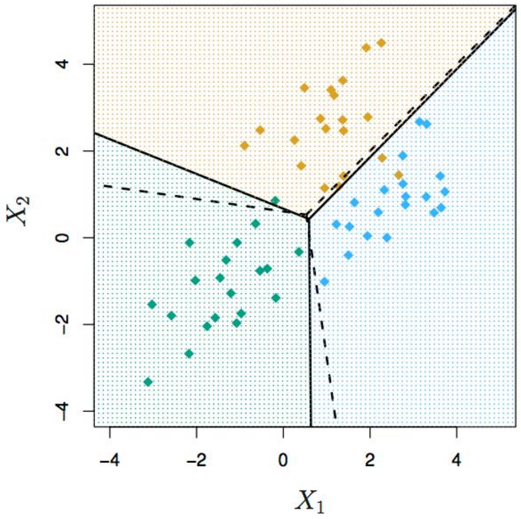
An Example When p > 1

- If X is multidimensional (p > 1), we use exactly the same approach except the density function *f(x)* is modeled using the multivariate normal density



Linear Discriminant Analysis (LDA)

- We have two predictors (p =2)
- Three classes
- 20 observations were generated from each class
- The solid lines are Bayes' boundaries
- The dashed lines are LDA boundaries

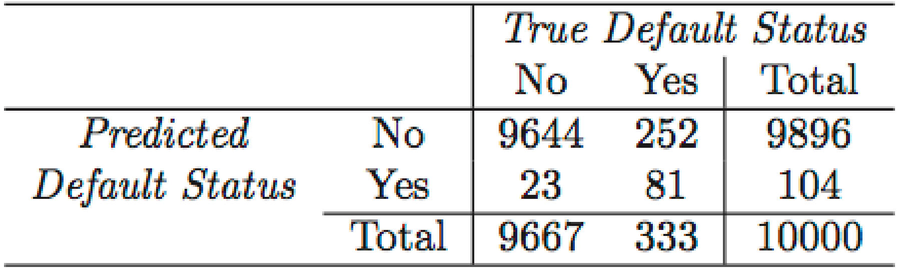


Linear Discriminant Analysis (LDA)

Running LDA on Default Data

- LDA makes 252+ 23 mistakes on 10000 predictions (2.75% misclassification error rate)
- But LDA miss-predicts 252/333 = 75.5% of defaulters!

Perhaps, we shouldn’t use 0.5 as threshold for predicting default?

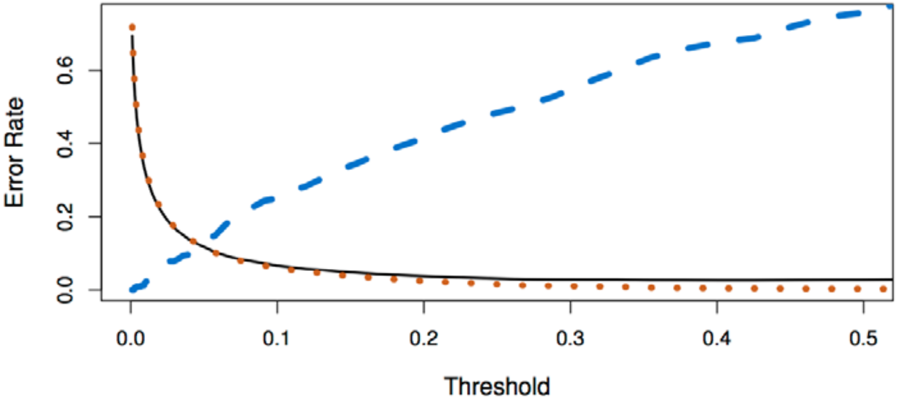


		True	*Default*	*Status*
		No	Yes	Total
Predicted Default Status	No	9644	252	9896
	Yes	23	81	104
	Total	9667	333	10000

Linear Discriminant Analysis (LDA)

Default Threshold Values vs. Error Rates

- Black solid: overall error rate
- Blue dashed: Fraction of defaulters missed
- Orange dotted: non defaulters incorrectly classified



Linear Discriminant Analysis (LDA)

Lab: The Stock Market Data

- Use Smarket data, which is part of ISLR library
- Use lda() function in MASS library

Quadratic Discriminant Analysis (QDA)

- LDA assumed that every class has the same variance/ covariance

- However, LDA may perform poorly if this assumption is far from true

- QDA works identically as LDA except that it estimates separate variances/ covariance for each class

Quadratic Discriminant Analysis (QDA)

Which is better? LDA or QDA?

- Since QDA allows for different variances among classes, the resulting boundaries become quadratic

- Which approach is better: LDA or QDA?

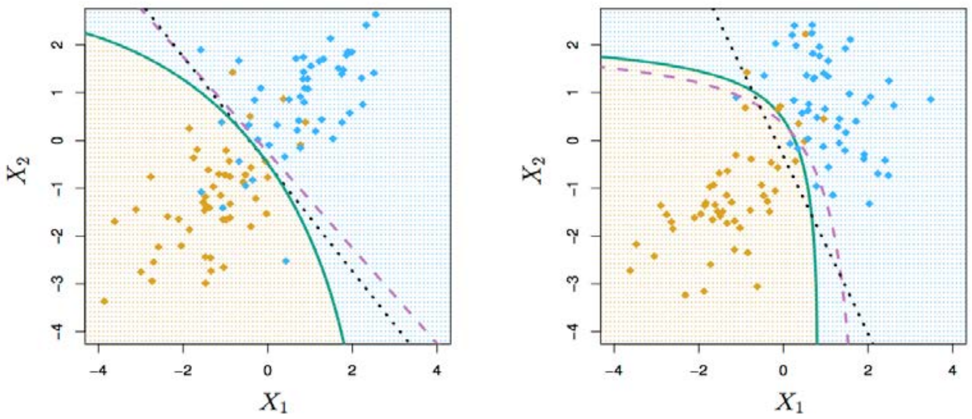
- QDA will work best when the variances are very different between classes and we have enough observations to accurately estimate the variances

- LDA will work best when the variances are similar among classes or we don’t have enough data to accurately estimate the variances

Quadratic Discriminant Analysis (QDA)

Comparing LDA to QDA

- Black dotted: LDA boundary
- Purple dashed: Bayes’ boundary
- Green solid: QDA boundary
- Left: variances of the classes are equal (LDA is better fit)
- Right: variances of the classes are not equal (QDA is better fit)



Quadratic Discriminant Analysis (QDA)

Lab: The Stock Market Data

- Use Smarket data, which is part of ISLR library
- Use qda() function in MASS library

K-Nearest Neighbors (KNN)

Given a positive integer K and a test observation x_0, the KNN classifier first identifies the K points in the training data that are closest to x_0, represented by N_0.

- It then estimates the conditional probability for class j as the fraction of points in N_0 whose response values equal j :

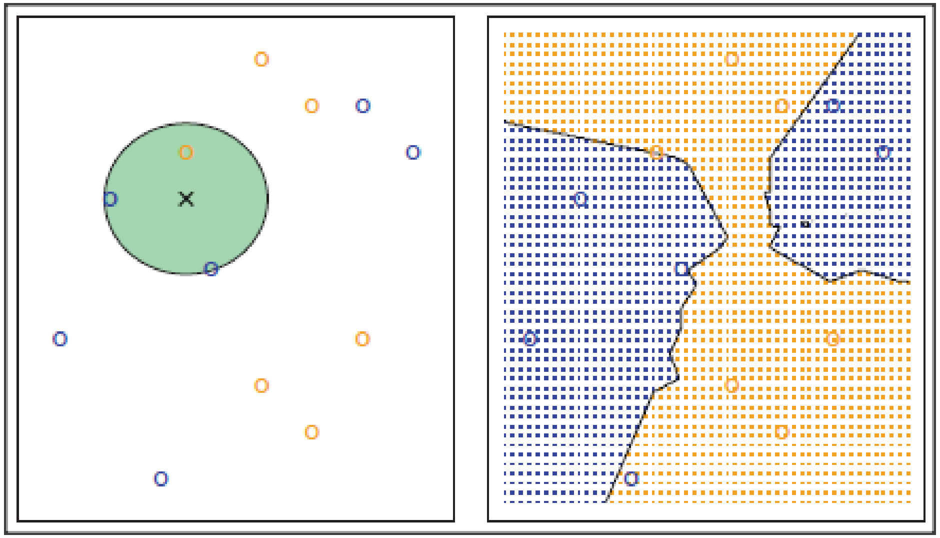
$$P(Y = j \mid X = x_0) = \frac{1}{K} \sum_{i \in N_0} I(y_i = j)$$

- Finally, KNN applies Bayes rule and classifies the test observation x_0 to the class with the largest probability.

K-Nearest Neighbors (KNN)

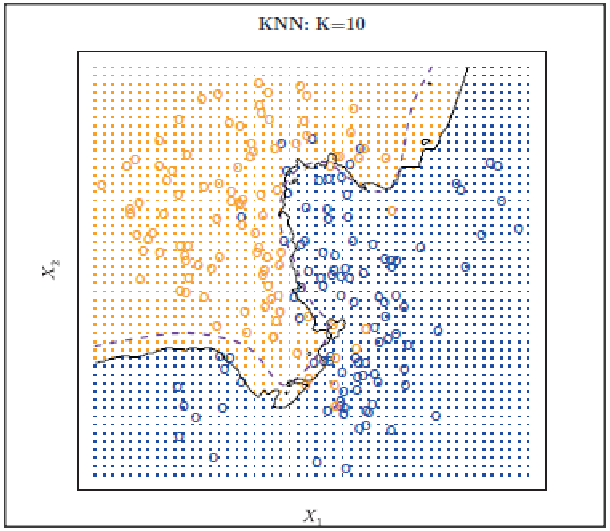
Example of the KNN approach:

- In the left-hand panel, we have plotted a small training data set consisting of six blue and six orange observations.
- Our goal is to make a prediction for the point labeled by the black cross.
- Suppose that we choose *K* = 3. Then KNN will first identify the three observations that are closest to the cross. This neighborhood is shown as a circle. In the right-hand panel of we have applied the KNN approach with *K* = 3 at all of the possible values for X_1 and X_2, and have drawn in the corresponding KNN decision boundary.



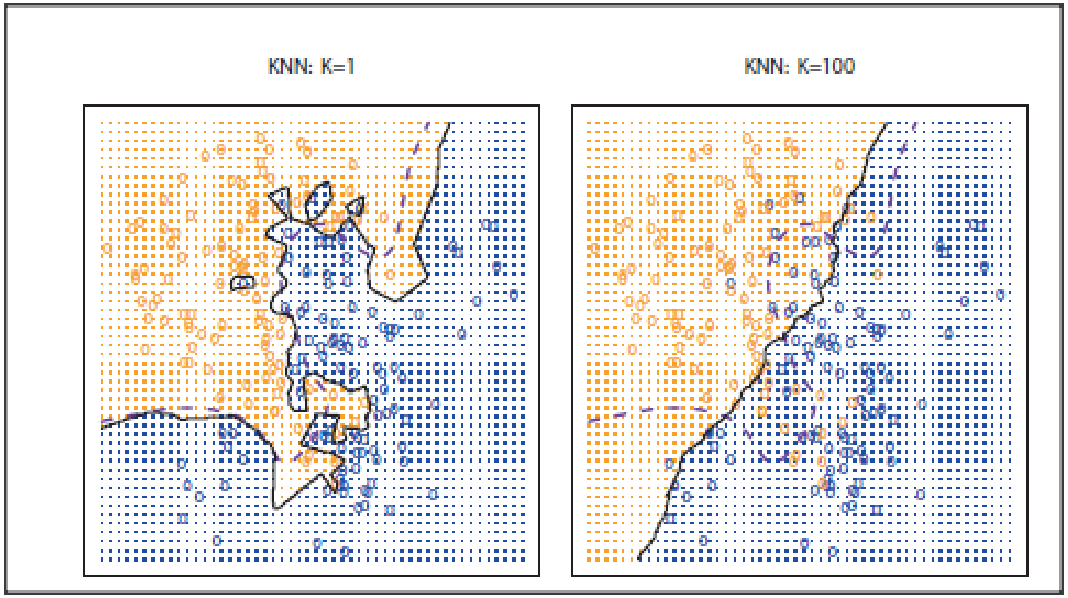
K-Nearest Neighbors (KNN)

- Figure displays the KNN decision boundary, using $K = 10$, when applied to the larger simulated data set.
- Notice that even though the true distribution is not known by the KNN classifier, the KNN decision boundary is very close to that of the Bayes classifier. The test error rate using KNN is 0.1363, which is close to the Bayes error rate of 0.1304.



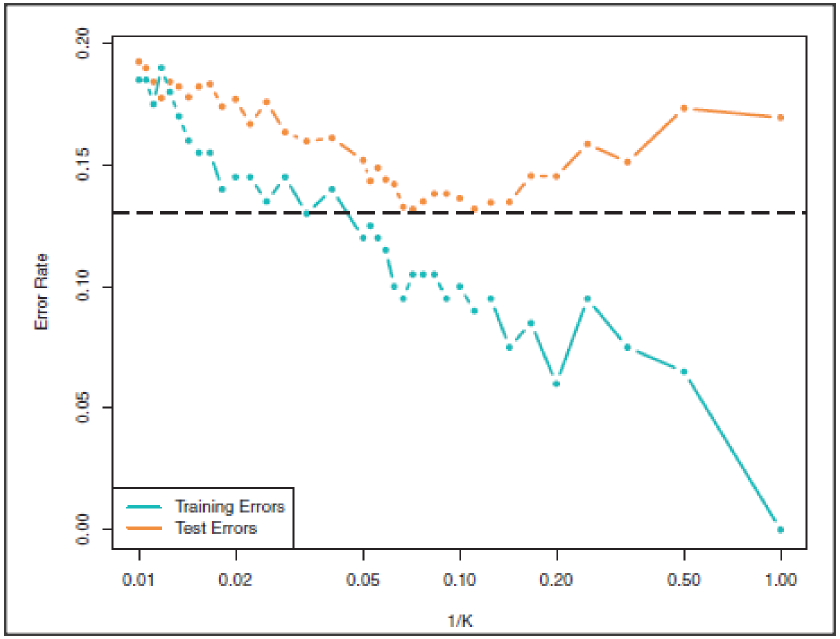
K-Nearest Neighbors (KNN)

- The choice of *K* has a drastic effect on the KNN classifier obtained. Figure displays two KNN fits to the simulated data, using *K* = 1 and *K* = 100. When *K* = 1, this corresponds to a classifier that has low bias but very high variance.
- As *K* grows, the method becomes less flexible and produces a decision boundary that is close to linear. This corresponds to a low-variance but high-bias classifier.



K-Nearest Neighbors (KNN)

- There is not a strong relationship between the training error rate and the test error rate. With *K* = 1, the KNN training error rate is 0, but the test error rate may be quite high.
- In general, as we use more flexible classification methods, the training error rate will decline but the test error rate may not. In Figure, we have plotted the KNN test and training errors as a function of 1/*K*.



K-Nearest Neighbors (KNN)

Lab: The Stock Market Data

- Use Smarket data, which is part of ISLR library

Comparison of Classification Methods

- KNN
- Logistic Regression
- LDA
- QDA

Comparison of Classification Methods

Logistic Regression vs. LDA

- Similarity: Both Logistic Regression and LDA produce linear boundaries

- Difference: LDA assumes that the observations are drawn from the normal distribution with common variance in each class, while logistic regression does not have this assumption. LDA would do better than Logistic Regression if the assumption of normality hold, otherwise logistic regression can outperform LDA

Comparison of Classification Methods

KNN vs. (LDA and Logistic Regression)

- KNN takes a completely different approach
- KNN is completely non-parametric: No assumptions are made about the shape of the decision boundary!
- Advantage of KNN: We can expect KNN to dominate both LDA and Logistic Regression when the decision boundary is highly non-linear
- Disadvantage of KNN: KNN does not tell us which predictors are important (no table of coefficients)

Comparison of Classification Methods

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

- QDA is a compromise between non-parametric KNN method and the linear LDA and logistic regression

- If the true decision boundary is:

    Linear: LDA and Logistic outperforms

    Moderately Non-linear: QDA outperforms

    More complicated: KNN is superior

STEVENS
INSTITUTE *of* TECHNOLOGY
School of Business

Thank You

Dragos Bozdog

For academic use only.