Why not Linear Regression

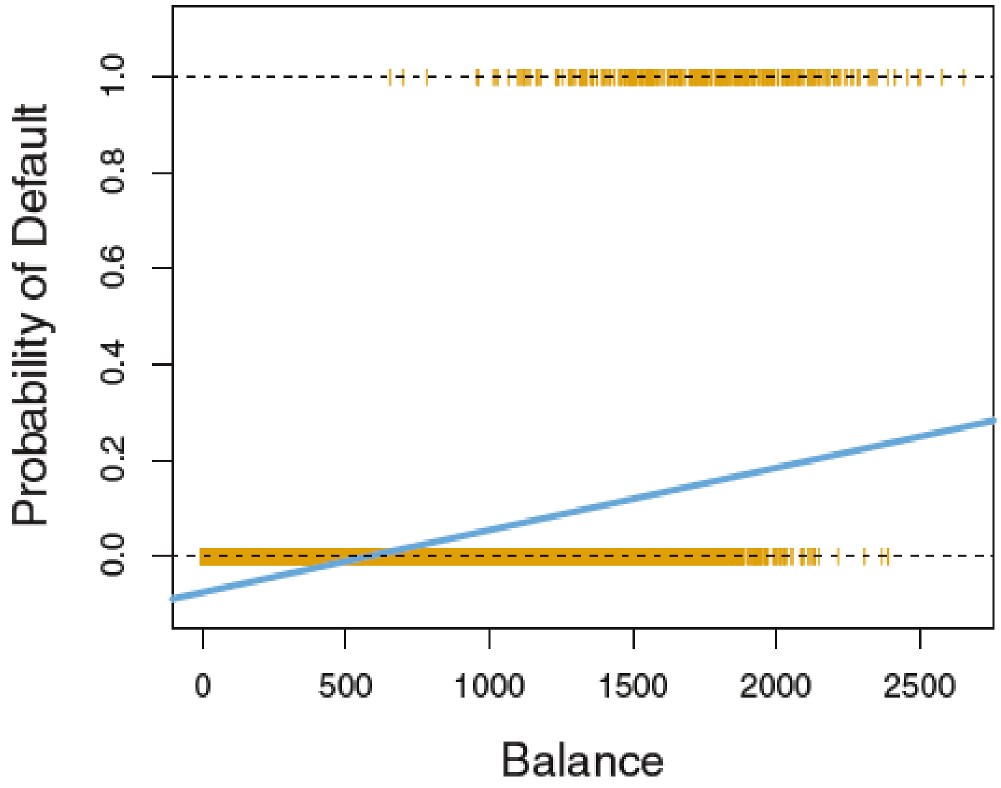
Data

	default	student	balance	income
1	No	No	729.52650	44361.625
2	No	Yes	817.18041	12106.135
3	No	No	1073.54916	31767.139
4	No	No	529.25060	35704.494

Linear regression cannot be used for more than two categories

Why not Linear Regression

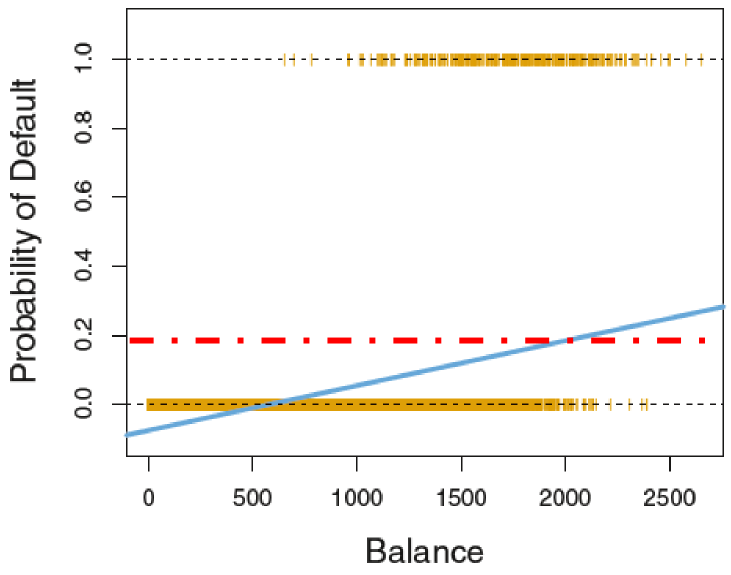
Limitations



Logistic Regression

Data

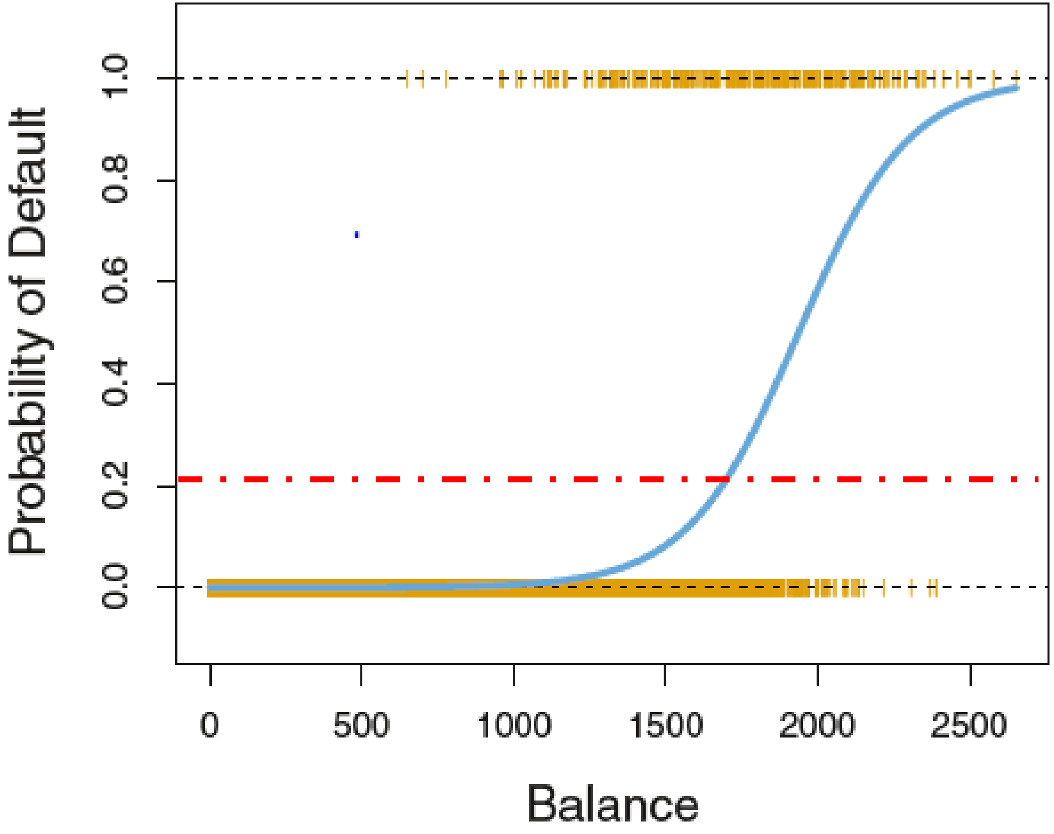
	default	student	balance	income
1	No	No	729.52650	44361.625
2	No	Yes	817.18041	12106.135
3	No	No	1073.54916	31767.139
4	No	No	529.25060	35704.494



Boundary Value

Logistic Regression

Sigmoid Function



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

Logistic Regression

Maximum Likelihood Method

$$\ell(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} (1 - p(x_{i'}))$$

Model	Method
Linear Regression	OLS (Ordinary Least Squares)
Logistic Regression	Maximum Likelihood method

Logistic Regression

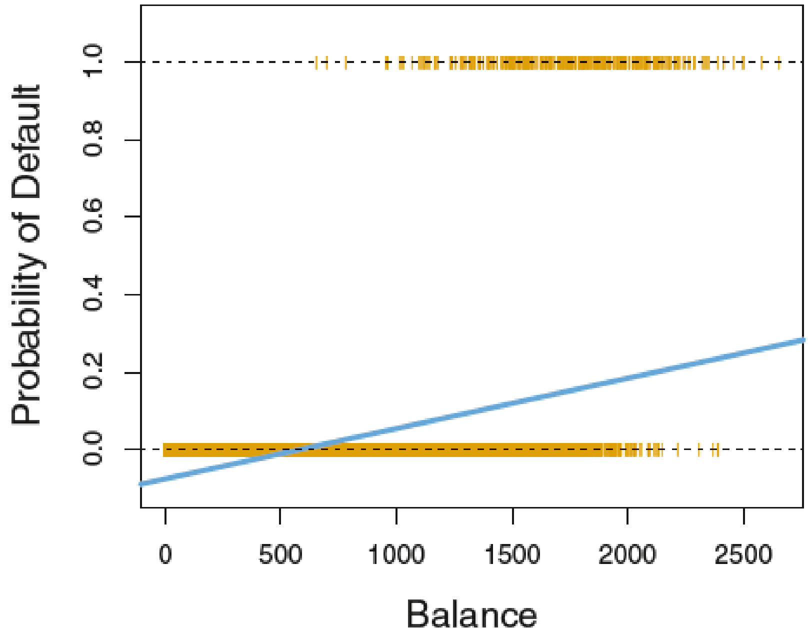
Data

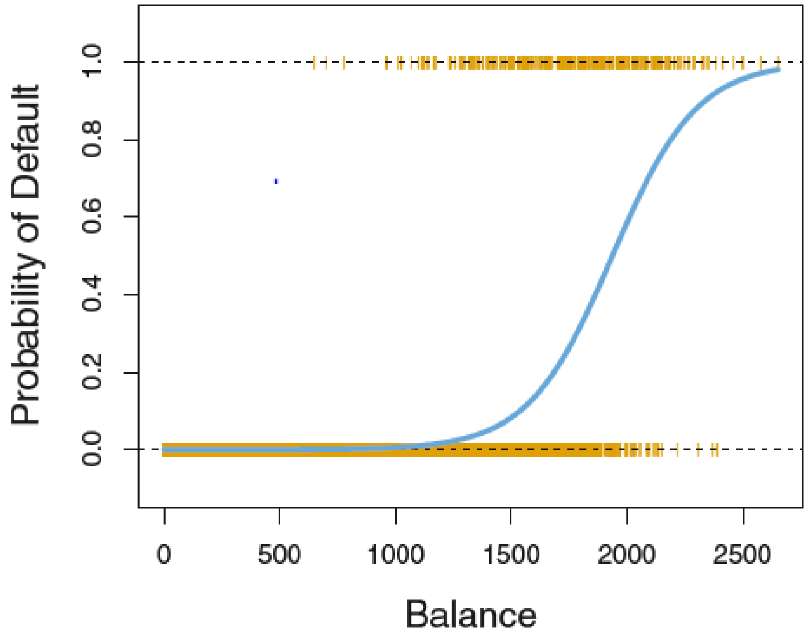
	default	student	balance	income
1	No	No	729.52650	44361.625
2	No	Yes	817.18041	12106.135
3	No	No	1073.54916	31767.139
4	No	No	529.25060	35704.494

Linear regression cannot be used for more than two categories

Logistic Regression

Limitations





Logistic Regression

Result

Result summary

```
Coefficients:
               Estimate Std. Error z value Pr(>|z|)
β0(Intercept)  0.61486    0.24751   2.484 0.012986 *
β1price       -0.03572    0.01045  -3.417 0.000632 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- If β is zero, it means there is no relationship

Ho : There is no relationship between X and Y

Ha : There is some relationship between X and Y

$$H : \beta_1 = 0$$

$$Ha : \beta_1 \neq 0,$$

Logistic Regression

Limitations

- To disapprove Ho, we calculate Z statistic$= \dfrac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)}$

- We also compute the probability of observing any value equal to /z/ or

Larger

- We call this probability the *p-value*
- *A* small p-value means there is an association between the predictor and the response (typically less than 5% or 1 %)

Key Takeaway

P value should be less than 0.05 (Threshold) to establish relationship

Logistic Regression

Multiple Predictors

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

- Use maximum likelihood to calculate Betas
- Fix the Boundary condition as per business requirements

Logistic Regression

Confusion matrix

		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

Linear regression cannot be used for more than two categories

Logistic Regression

Confusion matrix

		True default status		
		No	Yes	Total
Predicted default status	No	9,432	138	9,570
	Yes	$\boxed{235}$	195	430
	Total	9,667	333	10,000

Type 1 Error

Logistic Regression

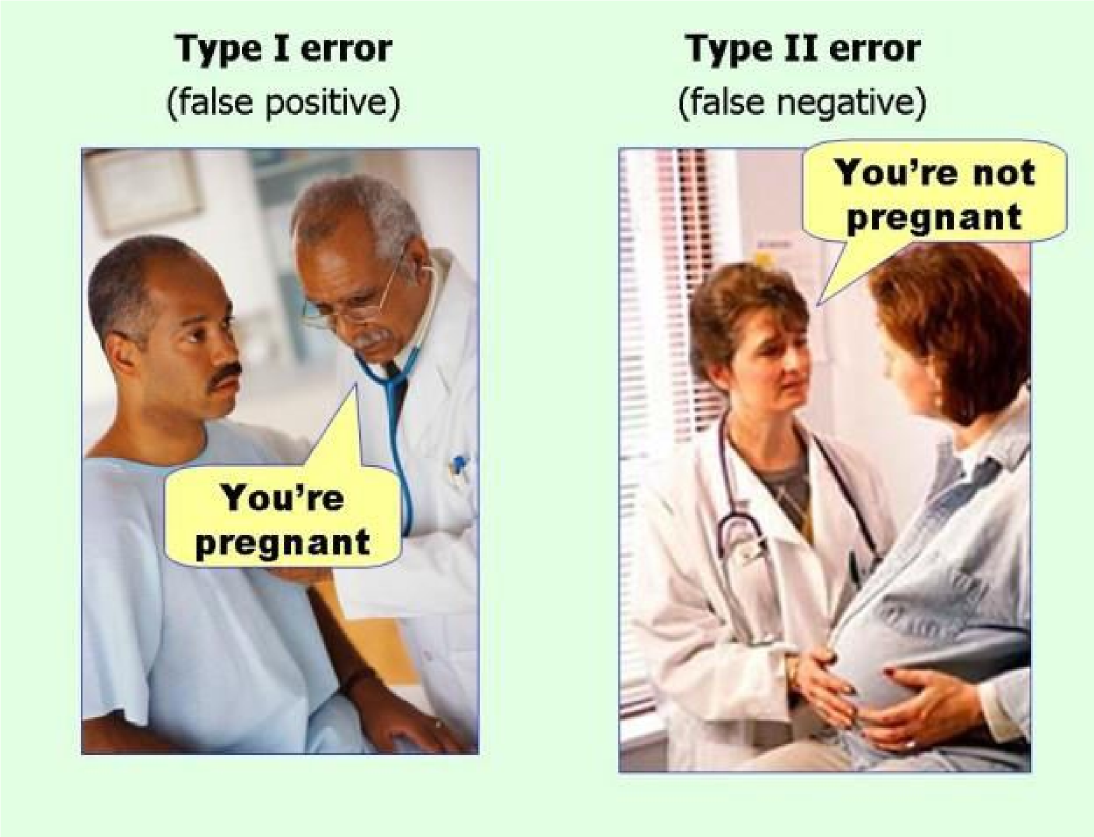
Confusion matrix

Type 2 Error

		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

Logistic Regression

Confusion matrix



Linear Discriminant Analysis

Linear Discriminant Analysis

- Preferred when response variable has more than two classes
- Based on Bayes theorem

Height	Fit	Not Fit	
Low	13	22	35
Medium	15	25	40
High	20	5	25
	48	52	100

Linear Discriminant Analysis

Conditional Probability

Height	Fit	Not Fit	
Low	13	22	35
Medium	15	25	40
High	20	5	25
	48	52	100

- Probability of ‘Fit’ given ‘Medium’ = 15/40
- Bayes Classifier :
  Assigns conditional probability to all classes and assign the class with highest probability

- 15/100= 15/40*40/100 = 15/48*48/100

Linear Discriminant Analysis

Conditional Probability

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

Performance Measures

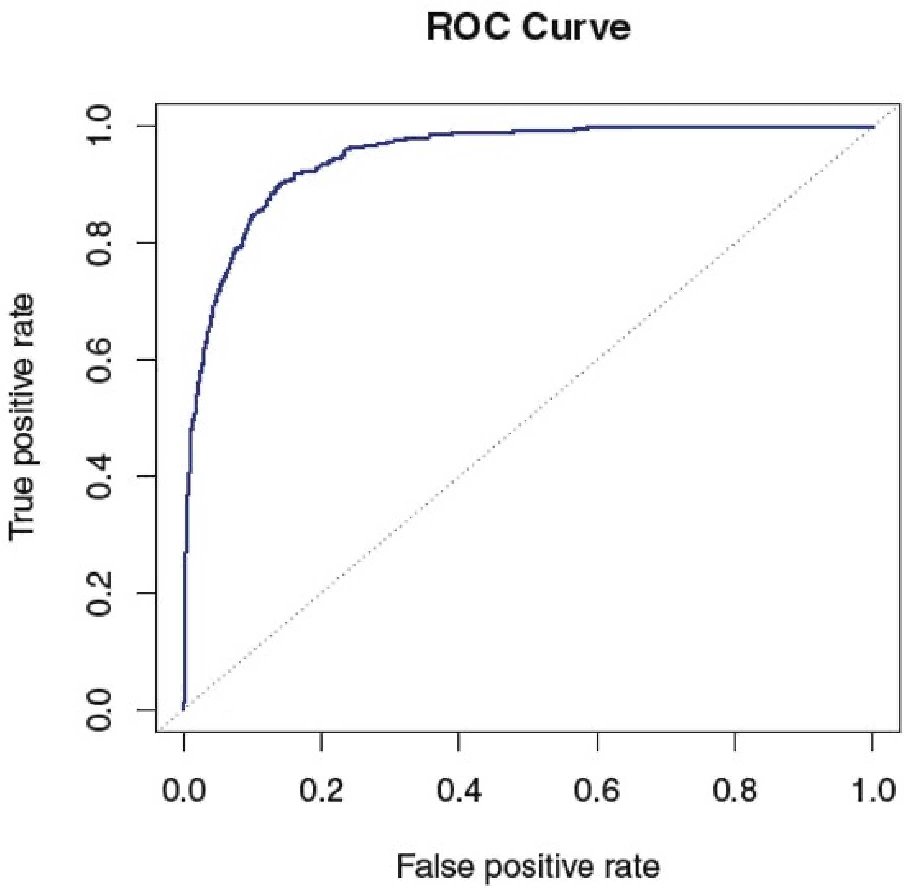
Performance Measures

		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*	

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*	

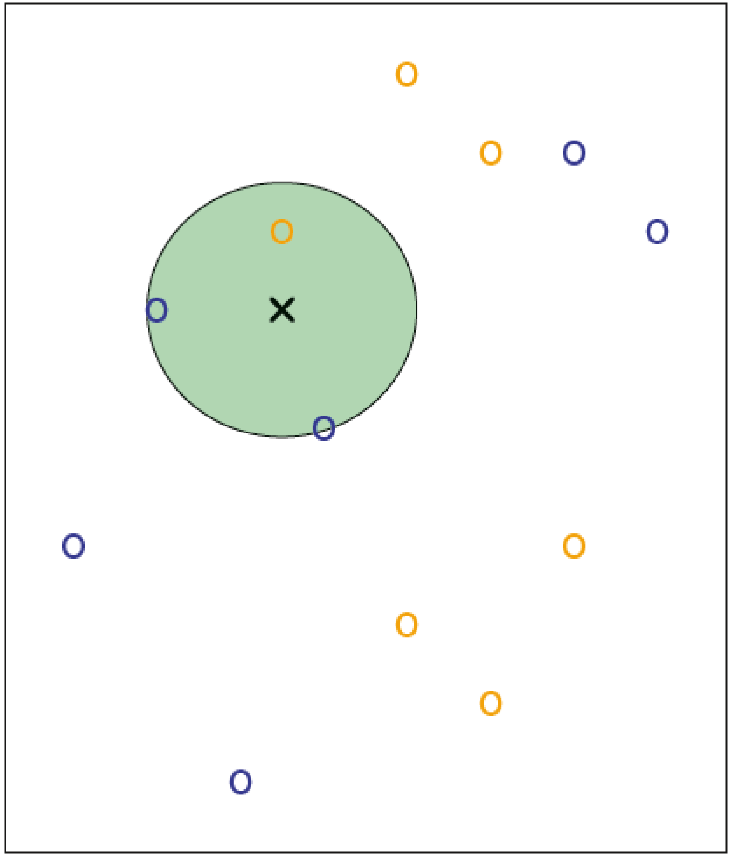
Performance Measures

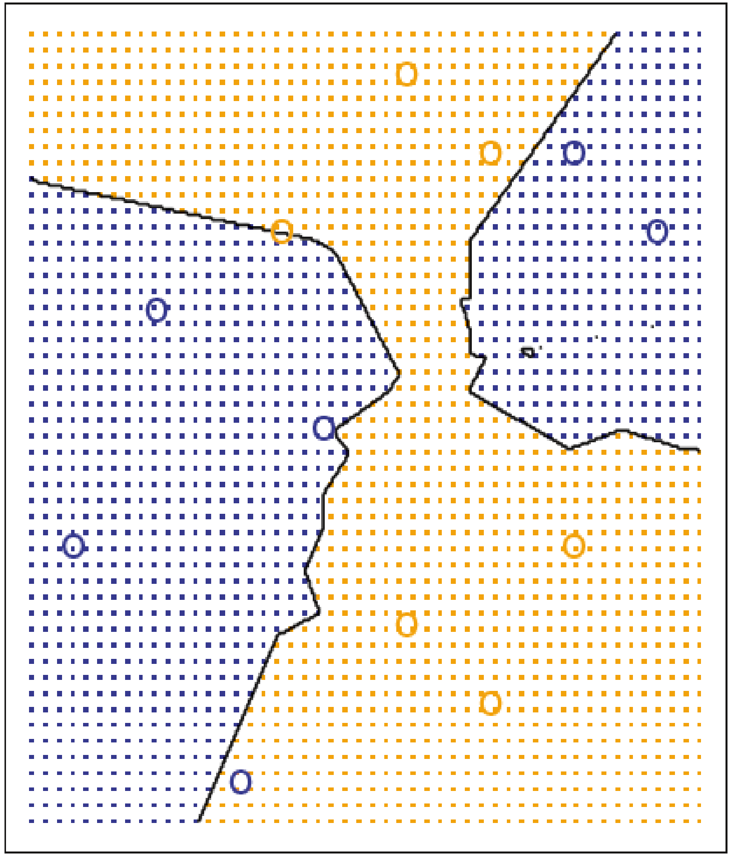
ROC



K-Nearest Neighbors

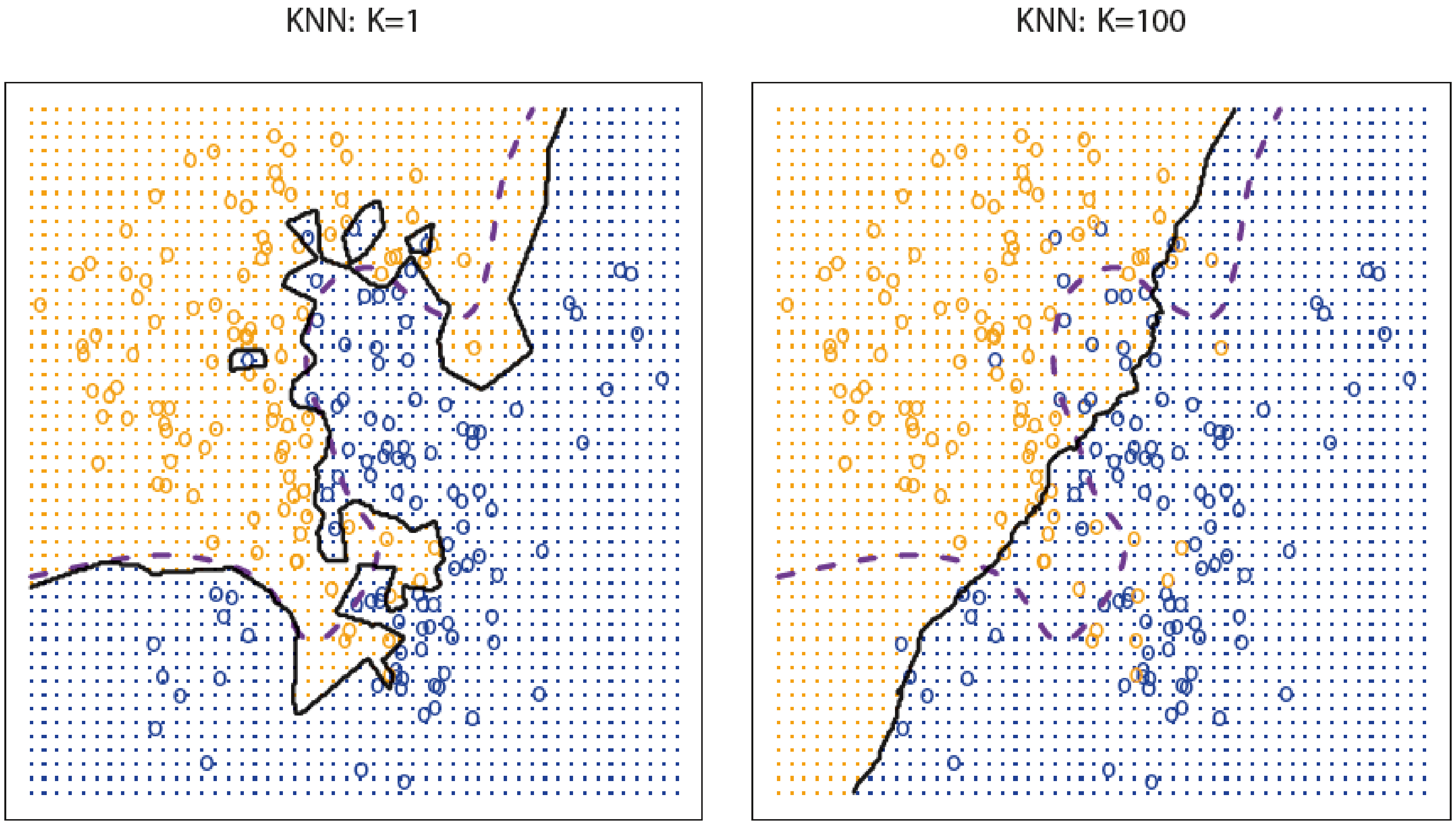
KNN





K-Nearest Neighbors

KNN



K-Nearest Neighbors

KNN

Notes

- In KNN distance between observations impacts the classifier
- Therefore, scale matters
- To handle the problem of scaling we standardize the data

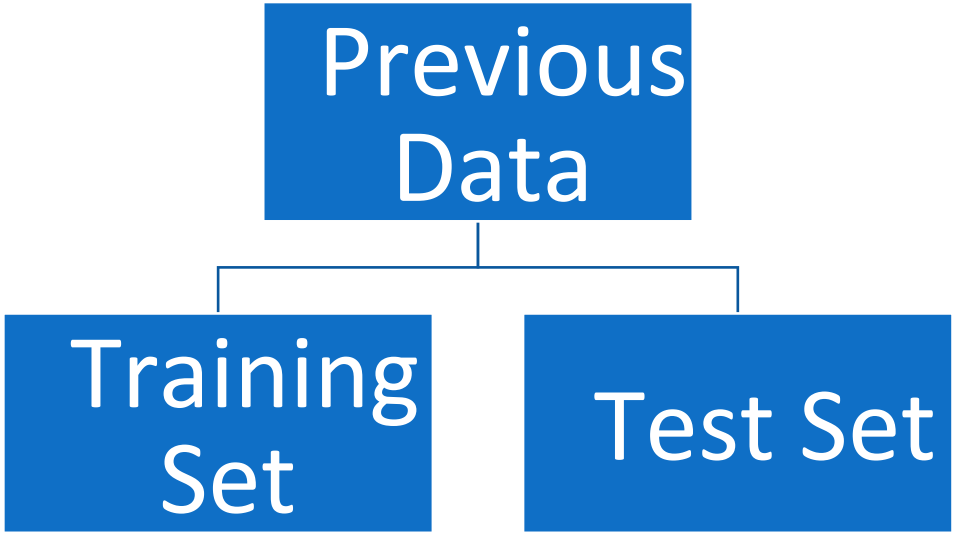
$$\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \overline{x}_j)^2}}$$

Linear Regression

Test-Train Split

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$$

- Training error – Performance of model on the previously seen data
- Test error – Performance of model on the unseen data



Linear Regression

Test-Train Split

Training Set - $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$

Model is trained

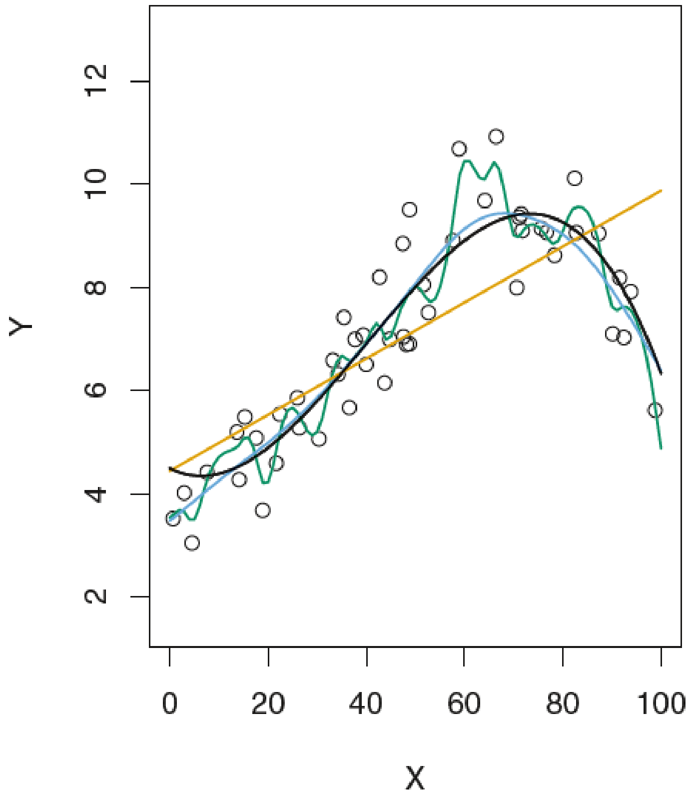
$$y = f(x)$$

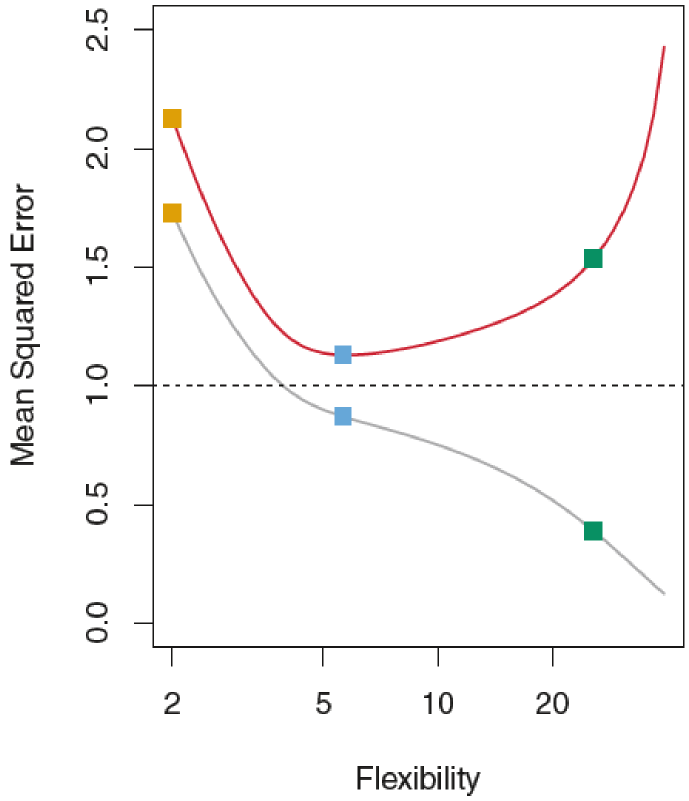
Test Set - Previously unseen d: (x_0, y_0)

Test MSE - $\mathrm{Ave}(\hat{f}(x_0) - y_0)^2$

Other Linear Regression

Test-Train Split





True Function
Output of linear Model
Output of more flexible model

Test error
Training error

Linear Regression

Test-Train Split Techniques

1. Validation set approach
    - Random division of data into two parts
    - Usual split is 80:20 (Training : Test)
    - When to use – In case of large number of observations

2. Leave one out cross validation
    - Leaving one observation every time from training set

3. K-Fold validation
    - Divide the data into k set
    - We will keep one testing and K-1 for training

Results

Logistic Regression

```
Coefficients:
                            Estimate Std. Error z value Pr(>|z|)    
(Intercept)                -3.786667   3.023162  -1.253 0.210369    
price                      -0.289955   0.039074  -7.421 1.17e-13 ***
resid_area                  0.040238   0.031089   1.294 0.195575    
air_qual                   -6.689560   3.038370  -2.202 0.027687 *  
room_num                    1.418795   0.333412   4.255 2.09e-05 ***
age                        -0.002811   0.007611  -0.369 0.711843    
teachers                    0.297946   0.072028   4.137 3.53e-05 ***
poor_prop                  -0.211818   0.040039  -5.290 1.22e-07 ***
airportYES                  0.033861   0.245330   0.138 0.890223    
n_hos_beds                  0.176256   0.083340   2.115 0.034439 *  
n_hot_rooms                -0.079553   0.056361  -1.412 0.158097    
waterbodyLake              -0.062983   0.370489  -0.170 0.865011    
`waterbodyLake and River`  -0.199015   0.361962  -0.550 0.582442    
waterbodyRiver              0.080375   0.293049   0.274 0.783877    
rainfall                   -0.005667   0.009691  -0.585 0.558725    
parks                      20.411874  27.453336   0.744 0.457172    
avg_dist                   -0.427118   0.115154  -3.709 0.000208 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Results

Results

Method	Confusion Matrix	Accuracy

Logistic Regression	`pred 0 1`
	`NO 42 16`
	`YES 26 36`
		65%

LDA	`lda.class 0 1`
	`0 44 16`
	`1 24 36`
		66.6%

KNN (k=3)	`testy`
	`knn.pred 0 1`
	`0 38 24`
	`1 30 28`
		55%

Summary

Steps

- Data Collection
- Data Pre-processing
  - Outlier Treatment
  - Missing value imputation
  - Variable transformation
- Model training
  - Test-Train Split
  - Use template to train
  - Do iterations
  - Compare performance of different methods using test set
- Select the best model
  - For prediction purposes use model with best accuracy
  - For interpretation purposes look at the coefficient values of parametric models