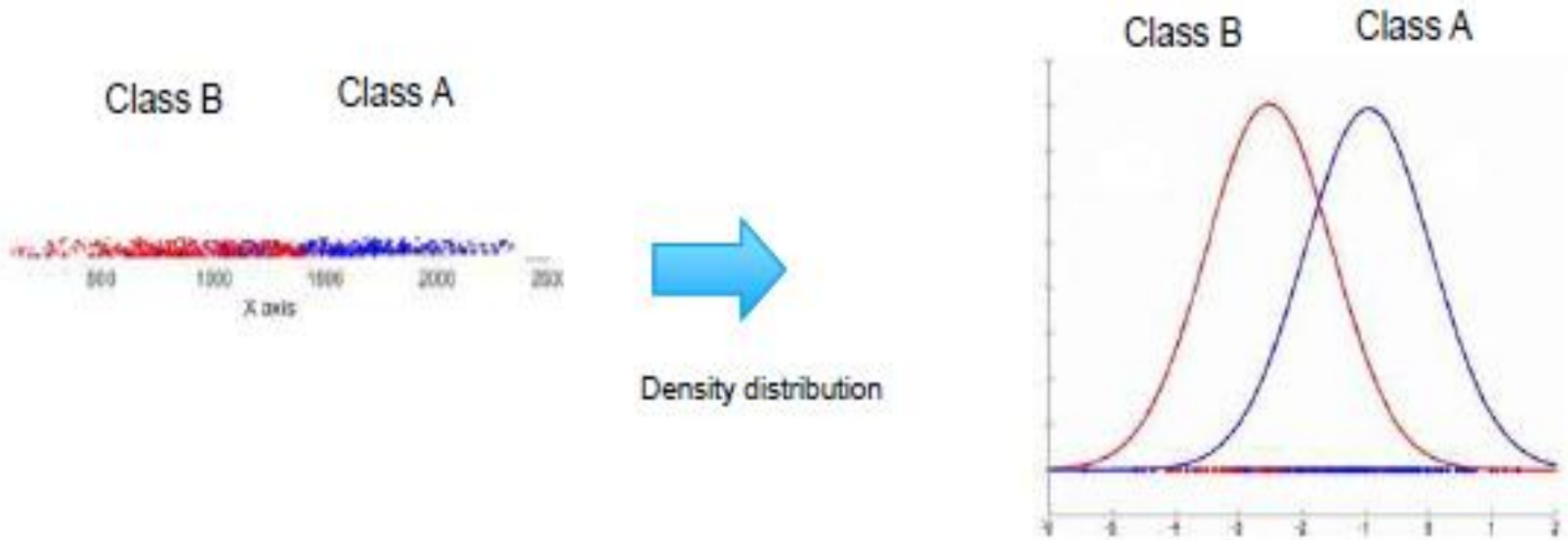


Logistic Regression

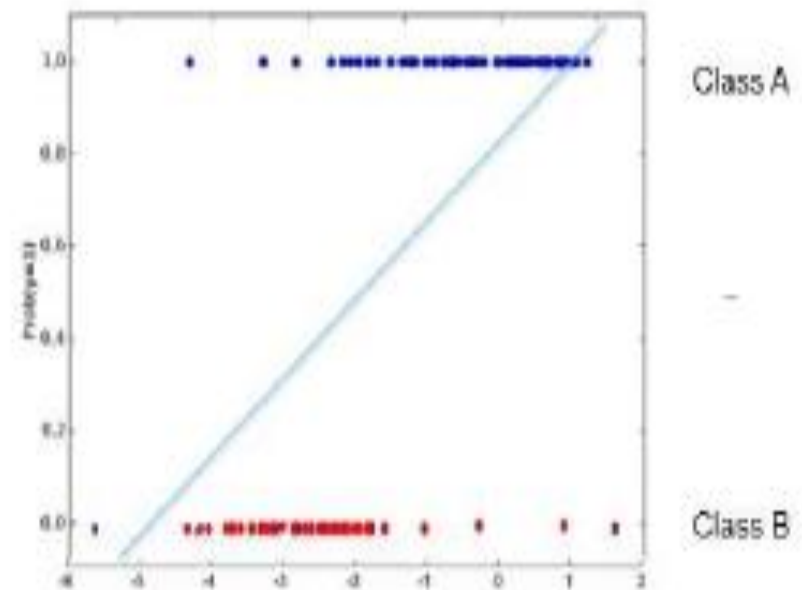
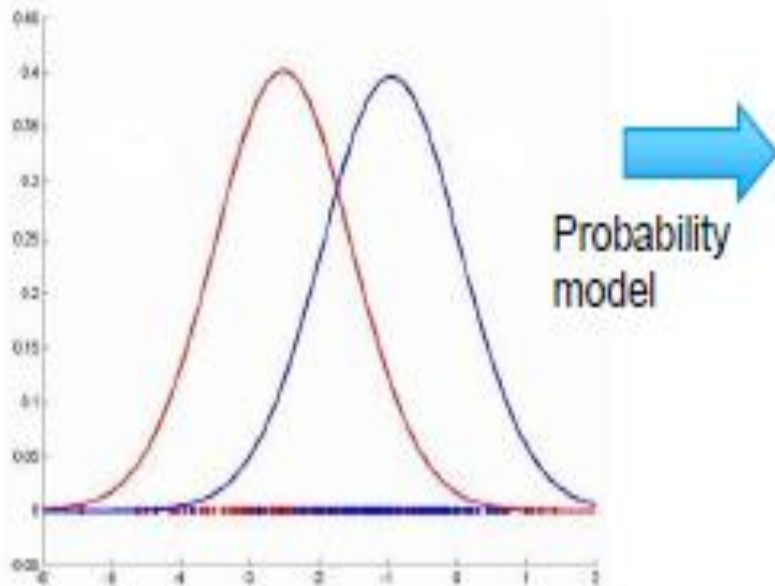
Logistic Regression

- A classification method built on the same concept as linear regression. The response variable is categorical. In its simplest form, the response variable is binary i.e. belongs to one class or the other
- Given the value of predictor (variable x), the model estimates the probability that the new data point belongs to a given class say “A”. Probability values can range between 0 and 1.



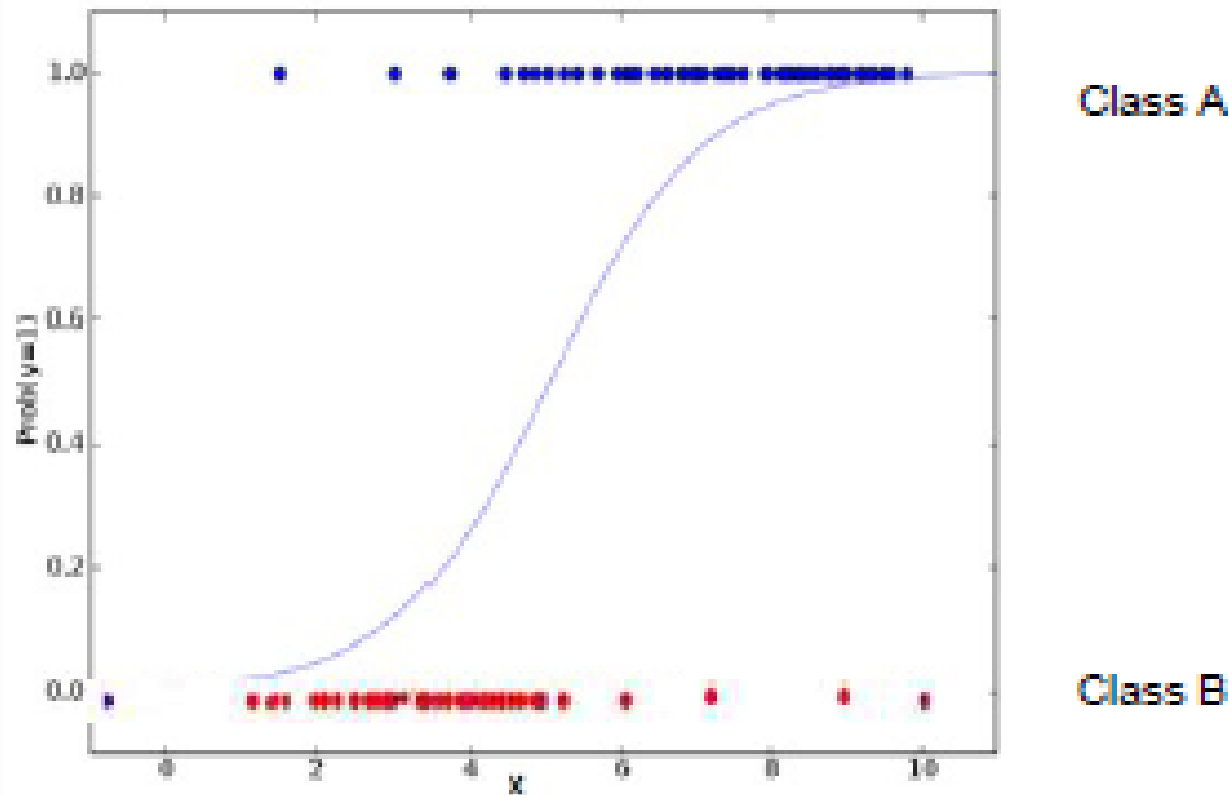
Logistic Regression

- A new data point (shown with “?”) needs to be classified i.e. does it belong to class A or B.
- Given the distribution, closer the point is to the origin, it is unlikely to belong to class A. Farther away it is from the origin, likely it belongs to class A
- One can try to fit a simple linear model ($y = mx + c$) where y greater than a threshold means point most probably belongs to class A. The challenge is, for extreme values of x , probability is <0 or >1 which is absurd



Logistic Regression

- The linear model is passed to a logistic function $p = 1 / (1 + e^{-t})$ the result of which is values between 0 and 1. Thus p represents probability a data point belongs to class "A" given x



Logistic Regression

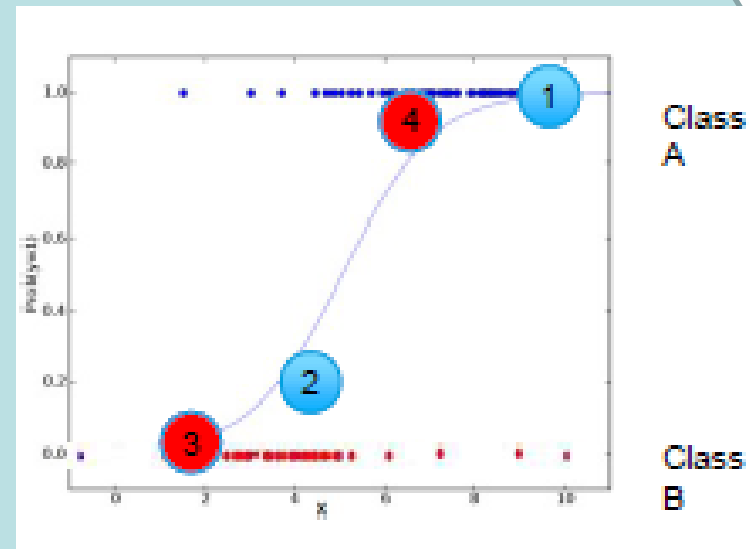
- Uses logloss function to find the best fit line from the infinite possibilities where

$$\logLoss = \frac{-1}{N} \sum_{i=1}^N (y_i(\log p_i) + (1 - y_i) \log(1 - p_i))$$

- The objective is to minimize logLoss as much as possible
- There can be four different cases for the value of y_i and p_i
 - Case1: $y_i = 1, p_i = High, 1 - y_i = 0, 1 - p_i = Low$ Correct classification
 - Case2: $y_i = 1, p_i = Low, 1 - y_i = 0, 1 - p_i = High$ Incorrect classification
 - Case3: $y_i = 0, p_i = Low, 1 - y_i = 1, 1 - p_i = High$ Correct classification
 - Case4: $y_i = 0, p_i = High, 1 - y_i = 1, 1 - p_i = Low$ Incorrect classification
- Correct classification contributes very minimal to the sum while a incorrect classification contributes large magnitudes

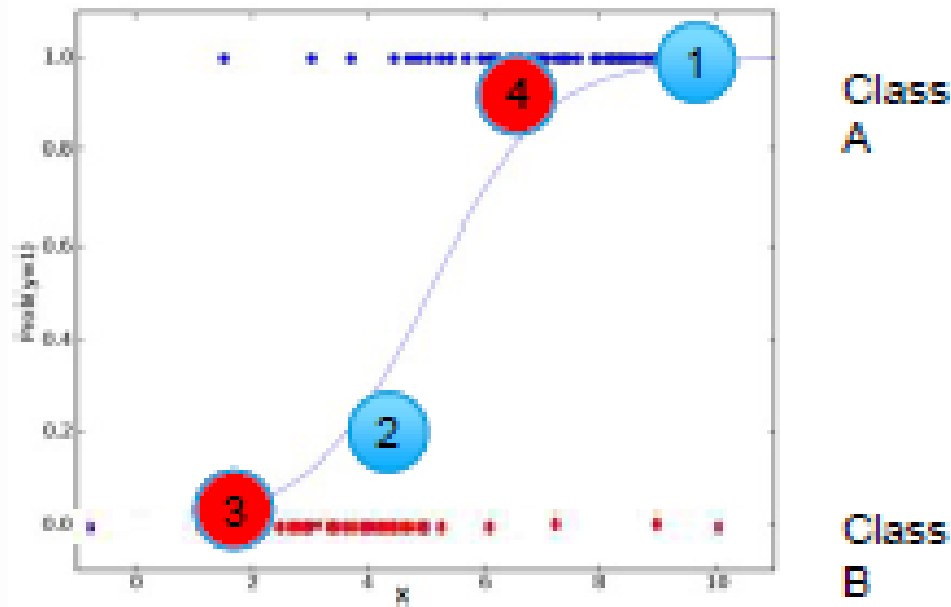
Logistic Regression

- In case1 $y_i = 1$ and $p_i = High$ implies that we have got things right!. It would not inflate the sum because, $y_i * \log(p_i)$ would be small while the other term in the sum would be zero since $1 - y_i = 1 - 1 = 0$
- So more occurrences of Case 1 would not inflate the sum
- In case2, $y_i = 1$ and p_i is low which is incorrect classification. $p_i = Low$, $y_i * \log(p_i)$ would inflate the sum significantly, the second term would be zero since $1 - y_i$ would be zero. So Case 2 would inflate the sum a lot.
- Similarly the occurrences of Case 3 would not inflate the sum significantly because first term would be 0 and second term will be small i.e. $(1 - y_i) * \log(1 - p_i)$
- Case 4 first term will be 0 while in second term due to high p_i , $(1 - p_i)$ will also be small hence contribution will be significant



Logistic Regression

- More of Case 1s and Case 3s does not increase the magnitude of the sum inside the logLoss formula
- More Case2s and Case4s will impact on the overall value significantly
- The objective is to find the logistic curve that makes the overall logLoss as small as possible



Logistic Regression





■ Advantages

- Makes no assumptions about distributions of classes in feature space
- Easily extended to multiple classes (multinomial regression)
- Natural probabilistic view of class predictions
- Quick to train
- Very fast at classifying unknown records
- Good accuracy for many simple data sets
- Resistant to overfitting
- Can interpret model coefficients as indicators of feature importance

■ Disadvantages

- Constructs linear boundaries

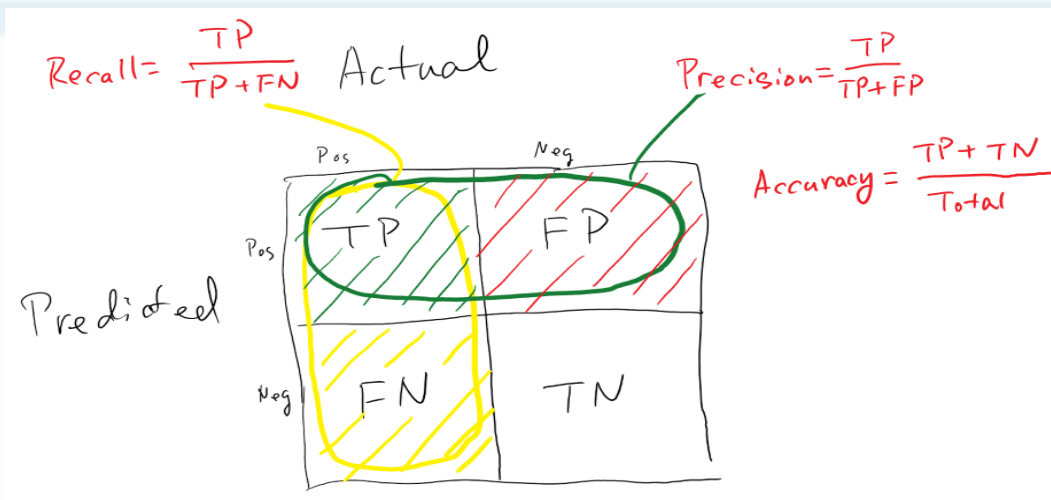
Classification Metric : Confusion Matrix

		Actual Values	
		1	0
Predicted Values	1	TRUE POSITIVE 	FALSE POSITIVE  TYPE 1 ERROR
	0	FALSE NEGATIVE  TYPE 2 ERROR	TRUE NEGATIVE 

Observe that we describe predicted values as Positive and Negative and actual values as True and False

- **True Positive**
 - Interpretation: You predicted positive and it's true
 - You predicted that a woman is pregnant and she actually is
- **True Negative**
 - Interpretation: You predicted negative and it's true
 - You predicted that a man is not pregnant and he actually is not
- **False Positive (Type 1 Error)**
 - Interpretation: You predicted positive and it's false
 - You predicted that a man is pregnant but he actually is not
- **False Negative (Type 2 Error)**
 - Interpretation: You predicted negative and it's false
 - You predicted that a woman is not pregnant but she actually is

Classification Metric : Confusion Matrix



- Recall = $TP / (TP + FN)$
 - Out of all positive classes, how much we predicted correctly. It should be high as possible
- Precision = $TP / (TP + FP)$
 - Out of all predicted positive classes, how many are correctly predicted. It should be as high as possible
- F-measure = $(2 * Recall * Precision) / (Recall + Precision)$
 - It is difficult to compare two models with low precision and high recall or vice versa. So to make them comparable, we use F-Score. F-score helps to measure Recall and Precision at the same time. It uses Harmonic Mean in place of Arithmetic Mean by punishing the extreme values more
- **Exercise: What are Recall, Precision and F-measure for PIMA Indians diabetes prediction?**

Classification Metric : Confusion Matrix

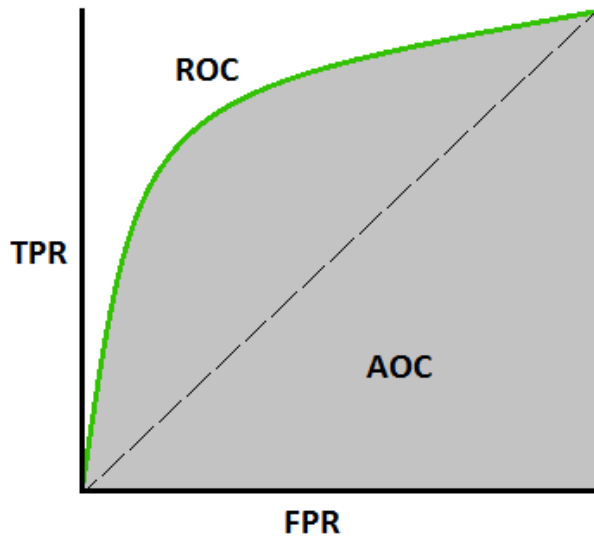
- A tool to assess the performance of a classification model such as logistic regression model

Total = 231	Actual 0	Predicted 1	Row Total
Predicted 0	130	38	168
Predicted 1	17	46	63
Col Total	147	84	231

- Of the 84 actual diabetes case, the model correctly classified only 46 as diabetic
- Of the 147 non diabetic cases, the model correctly classified 130 as non-diabetic
- 17 cases who are normal but identified as diabetic are called Type 1 error
- 38 cases of diabetic patients identified as normal is Type II error

Classification Metric : ROC-AUC

- AUC - ROC curve is a performance measurement for classification problem at various thresholds settings.
- ROC is a probability curve and AUC represents degree or measure of separability. It tells how much model is capable of distinguishing between classes.
- Higher the AUC, better the model is at predicting 0s as 0s and 1s as 1s.
- By analogy, Higher the AUC, better the model is at distinguishing between patients with disease and no disease.
- The ROC curve is plotted with TPR against the FPR where TPR is on y-axis and FPR is on the x-axis.



- $\text{TPR (True Positive Rate) / Recall / Sensitivity} = \text{TP} / (\text{TP} + \text{FN})$
- $\text{Specificity} = \text{TN} / (\text{TN} + \text{FP})$
- $\text{FPR (False Postive Rate)} = 1 - \text{Specificity} = \text{FP} / (\text{TN} + \text{FP})$

Lab5

- Consider pima_indians Data Set
- Build the best model to predict diabetic patient
- Analyse the results
- Separate data into train and test data sets. Build the model with train data and validate the model with test data. Compare the results of train and test data sets. Any observations?

Lab6

- Consider German Credit Data Set
- Build the best model to predict default
- Analyse the results
- Separate data into train and test data sets. Build the model with train data and validate the model with test data. Compare the results of train and test data sets. Any observations?

Lab7

- Consider Heart Data Set
- Build the best model to predict coronary heart disease(CHA)
- Analyse the results
- Separate data into train and test data sets. Build the model with train data and validate the model with test data. Compare the results of train and test data sets. Any observations?

Lab8

- Consider HR Attrition Data Set
- Build the best model to predict employees who are going to leave the organization
- Analyse the results
- Separate data into train and test data sets. Build the model with train data and validate the model with test data. Compare the results of train and test data sets. Any observations?

Questions?



Thanks!