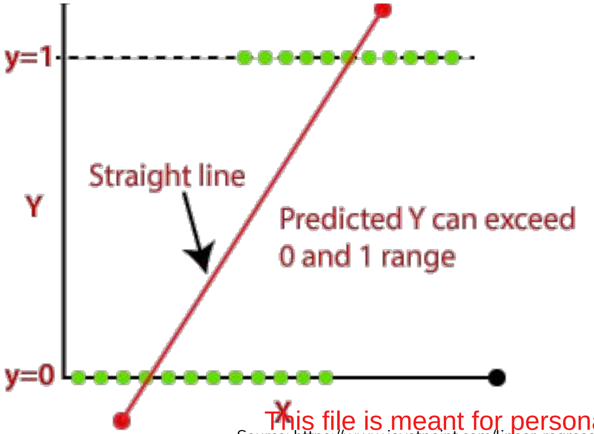
Logistic Regression

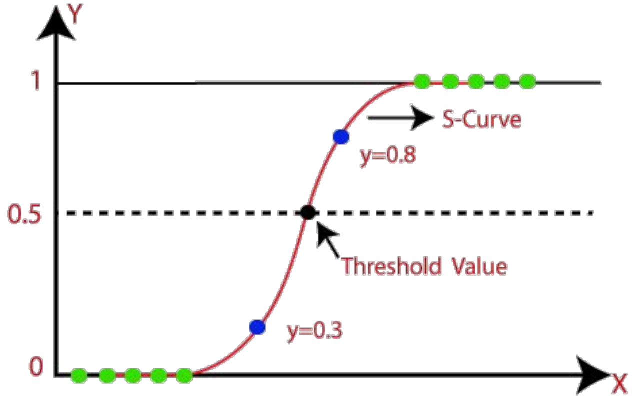
Discussion Questions

1. Why do we need Logistic Regression?
    a. What is the difference between Regression and Classification?
    b. Can we use Linear Regression for a classification problem?
    c. What is the difference between Linear and Logistic Regressions?
    d. What is the output of a Logistic Regression model?
2. How do we measure the performance of a Logistic Regression model?
3. Why is accuracy not always a good performance measure?
4. How do we choose the optimal threshold value?
5. What are some other ways to look at the model performance?

Why do we need Logistic Regression?

- Linear Regression uses a set of independent variables to predict a continuous dependent variable whereas Logistic Regression is used where the dependent variable is categorical/discrete.
- Linear Regression can not be used to predict probabilities because we can not restrict its output between 0 and 1.
- Logistic Regression is a supervised learning algorithm that estimates the log of odds for an event which can be used to predict the probability of the occurring of that event.
- The predicted probabilities can be converted to classes based on the threshold (the default threshold is 0.5).





Source: https://www.javatpoint.com/linear-regression-vs-logistic-regression-in-machine-learning

Confusion Matrix

It is a tabular representation of the predicted vs actual classes. We can use it to assess the performance of a Logistic Regression model.

- **True Positives (TP):** The model predicted the class as positive, and it is actually positive.
- **True Negatives (TN):** The model predicted the class as negative, and it is actually negative.
- **False Positives (FP):** The model predicted the class as positive, but it is actually negative.(Also known as a "Type I error").
- **False Negatives (FN):** The model predicted the class as negative, but it is actually positive.(Also known as a "Type II error").

Actual Values

Predicted Values	Positive (1)	Negative (0)
Positive (1)	TP	FP
Negative (0)	FN	TN

Metric's that are often computed from a confusion matrix for a binary classifier:

- **Accuracy** = TP + TN / (TP + FP + FN + TN)
- **Precision** = TP / (TP + FP)
- **Recall or sensitivity** = TP / (TP+FN)
- **Specificity** = TN / (TN + FP)

Why is accuracy not always a good performance measure?

Accuracy is simply the overall % of correct predictions and can be high even for very useless models.

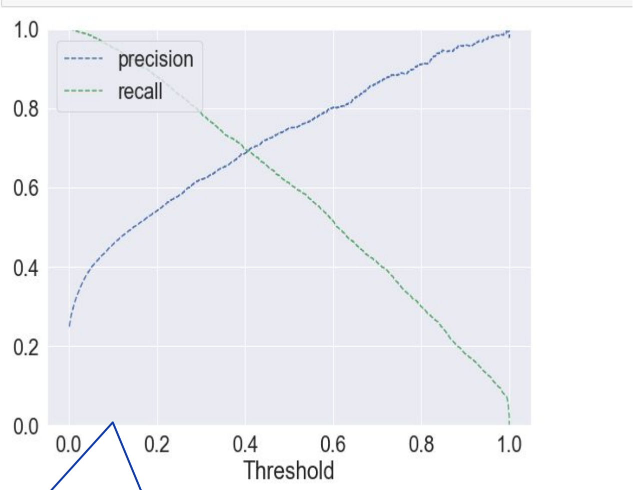
Total Patients – 100 # of Patients having cancer- 2 → Cancer rate – 2% → Model predicts that no one has cancer → Accuracy – 98% → Misses out critical patients having cancer

- Here Accuracy will be 98%, even if we predict all patients do not have cancer.
- In this case, Recall should be used as a measure of the model performance, High recall implies less False Negatives.
- Less False Negatives imply fewer chances of predicting a patient having cancer as a patient not having cancer.
- This is where we need other metrics to evaluate model performance.

- The other measures are Recall and Precision
  - Recall - What % of actuals 1s did we capture in my prediction?
  - Precision - What % of our predicted 1s are actually 1?
- There is a tradeoff - as you try to increase Recall, Precision will reduce and vice versa
- This tradeoff can be used to figure out the right threshold to use for the model

How do we choose the optimal threshold value?

- Precision-Recall is a useful measure of success of prediction when the classes are imbalanced.
- The Precision-Recall curve shows the tradeoff between Precision and Recall for different thresholds.
- It can be used to select optimal threshold as required to improve the model improvement
- Here we can see, Precision and Recall are the same when the threshold is 0.4
- If we want higher Precision, we can increase the threshold.
- If we want higher Recall, we can decrease the threshold.



Choosing a threshold can completely change the model performance assessment
It is important to think about what is the 'sweet spot'.

Is there a performance measure that can cover both Precision and Recall?

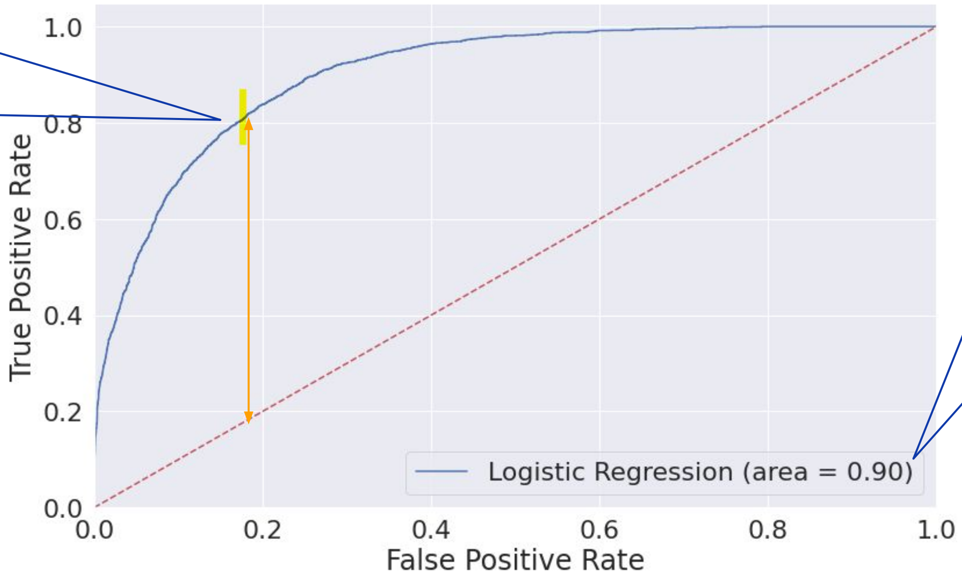
- F1 Score is a measure that takes into account both Precision and Recall.
- F1 Score is the harmonic mean of Precision and Recall. Therefore, this score takes both False Positives and False Negatives into account.

$$F1\ Score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

- The highest possible value of an F1 Score is 1, indicating perfect precision and recall, and the lowest possible value is 0.

What are some other ways to look at the model performance?

We can choose a threshold based on the point where the vertical distance between the plot and the baseline is maximum.



It generally varies from 0.5 to 1.

If it is less than 0.5, then there is something terribly wrong with the model as it is doing worse than random/baseline.

AUC = Area under the ROC Curve

Appendix

How does Logistic Regression work?

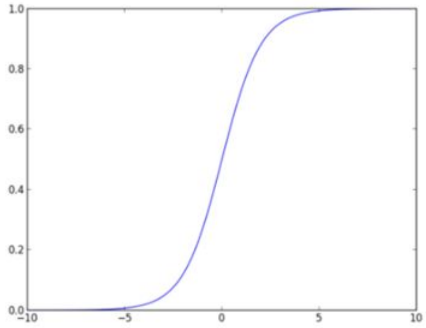
To understand the concept of a Logistic Regression, it is important to understand the concept of **Odds Ratio**, **Logit function**, and **Sigmoid function (or Logistic function)**

Odds Ratio (OR):

- Odds Ratio (OR) is the odds in favor of a particular event.
- Let P be the probability of subjects affected, then
  Odds = P/(1-P)

Logit Function:

- Logit function is the logarithm of the Odds Ratio (log-odds). It takes input values in the range 0 to 1 and then transforms them to value over the entire real number range.
- If P is the probability, then
  Logit(P) = Log(P/(1-P))



Sigmoid function:

- The inverse of the logit function is the **sigmoid** function.
- The Sigmoid Function can take any real value and map it to a value between 0 and 1.
- It is also called Logistic Function and gives an S shaped curve.
  Sigmoid(x) = 1 / (1 + e^(-x))

What is the relationship between Logit, Sigmoid and Logistic Regression

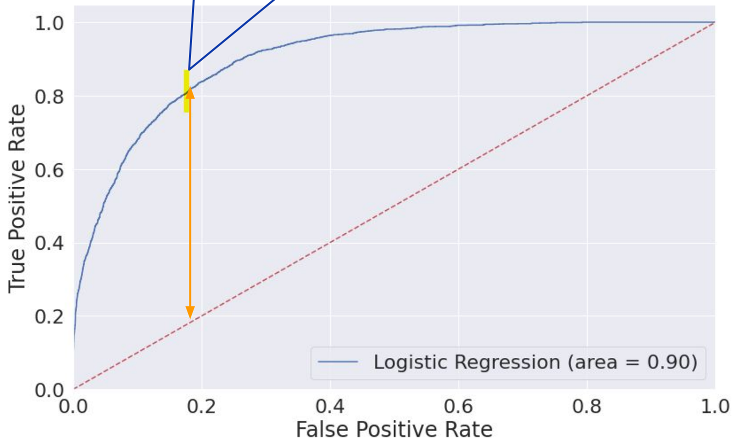
- Linear Regression Equation
  - **Y = a1 + a2*x + error**
- If the dependent variable Y is the logit function
  - **Logit(P) = Y = a1 + a2*x + error**

    where P = the probability of sample belonging to a class
  - **log(P/1-P) = a1 + a2*x + error**
- Apply the sigmoid function over LHS and RHS to get probabilities,
  - **sigmoid(log(P/1-P)) = sigmoid(a1 + a2*x + error)**
- So, we get,
  - **P = 1 / (1 + e^-(a1 + a2*x + error))**
  - **This 'P' is the output of the Logistic Regression model,** i.e. we are getting the probability of sample belonging to a class.
- Usually if P>0.5, we mark it as positive, and if P<0.5, we mark it as negative
- This cut-off point, known as **Threshold**, can be changed between 0 to 1, depending on the context of the problem.

ROC Curve

1. It is a plot between TPR(True Positive Rate)/Sensitivity and FPR(False Positive Rate)/(1 - Specificity) for varying thresholds in the same model
2. The area under the ROC curve (AUC) is a measure of how good a model is - The higher the AUC, the better the model is, at distinguishing between classes
3. The red diagonal represents a model whose predictions are as good as random
4. The further the ROC curve is from the diagonal line, the better the model is, at distinguishing between positive and negative classes
5. We can use this curve for getting a better threshold value as per our requirement.

We can choose threshold based on the point where the vertical distance between the plot and the baseline is maximum.



greatlearning
Power Ahead

Happy Learning !