INTERVIEW DAILY QUESTION BANK 01

Bias Variance, Confusion Matrix

Q1. What is the difference between supervised and unsupervised machine learning?

- **Supervised Machine learning:** Supervised machine learning requires training labelled data.
- **Unsupervised Machine learning:** Unsupervised machine learning doesn't require labelled data.

Q2. What is bias, variance trade off?

Bias:

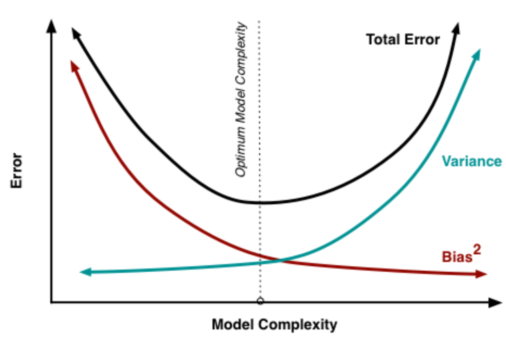
- "Bias is error introduced in your model due to over simplification of machine learning algorithm."
- It can lead to **under-fitting**.
- When you train your model at that time model makes simplified assumptions to make the target function easier to understand.
- **Low bias machine learning algorithms** — Decision Trees, k-NN and SVM (because they have very less assumptions)
- **High bias machine learning algorithms** — Linear Regression, Logistic Regression (because they have very high assumptions)

Variance:

- "Variance is error introduced in your model due to complex machine learning algorithm, your model learns noise also from the training data set and performs bad on test data set."
- It can **lead high sensitivity and over fitting**.
- Normally, as you increase the complexity of your model, you will see a reduction in error due to lower bias in the model.
- However, this only happens till a particular point. As you continue to make your model more complex, you end up **over-fitting your model and hence your model will start suffering from high variance**.

Bias, Variance trade off:

- The goal of any supervised machine learning algorithm is to have low bias and low variance to achieve good prediction performance.
- The k-nearest neighbours algorithm has low bias and high variance, but the trade-off can be changed by increasing the value of k which increases the number of neighbours that contribute to the prediction and in turn increases the bias of the model.
- The support vector machine algorithm has low bias and high variance, but the trade-off can be changed by increasing the C parameter that influences the number of violations of the margin allowed in the training data which increases the bias but decreases the variance.
- There is no escaping the relationship between bias and variance in machine learning.
- **Increasing the bias will decrease the variance. Increasing the variance will decrease the bias. So that is why it is Trade off.**

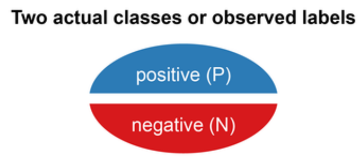


Q3. What is a confusion matrix?

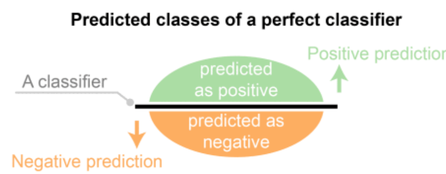
The confusion matrix is a **2X2 table** that contains **4 outputs provided by the binary classifier**. Various measures, such as error-rate, accuracy, specificity, sensitivity, precision and recall are derived from it.

		Predicted class	
		P	N
Actual Class	P	True Positives (TP)	False Negatives (FN)
	N	False Positives (FP)	True Negatives (TN)

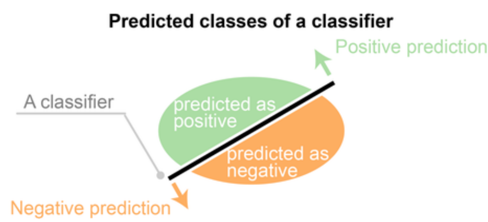
A data set used for performance evaluation is called **test data set**. It should contain the **correct labels and predicted labels**.



The predicted labels will exactly the same if the performance of a binary classifier is perfect.



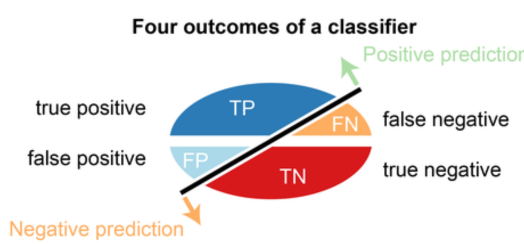
The predicted labels usually match with part of the observed labels in real world scenarios.



Q4. Explain True Positive, True Negative, False Positive, False Negative, Type 1 and Type 2 Error

A binary classifier predicts all data instances of a test dataset as either positive or negative.
This produces four outcomes-

- **True positive (TP)** — Correct positive prediction
- **False positive (FP)** — Incorrect positive prediction. Also known as **TYPE 1 ERROR**
- **True negative (TN)** — Correct negative prediction
- **False negative (FN)** — Incorrect negative prediction. Also known as **TYPE 2 ERROR**



Q5. Give an Example of Each

Take an example of my project :

Project 1 : A customer will invest in Fixed Deposit (1) or Not (0)

- **True positive (TP)** — Model predicted Customer will invest in FD (1) and Customer did invest in FD (1)
- **False positive (FP)** — Model predicted Customer will invest in FD (1) and Customer did not invest in FD (0)
- **True negative (TN)** — Model predicted Customer will not invest in FD (1) and Customer did not invest in FD (0)
- **False negative (FN)** — Model predicted Customer will not invest in FD (0) and Customer did invest in FD (1)

In this case, False Negative is Critical because we do not want to lose FD customer as there are very limited segment in the whole base so **TYPE 2 Error is Critical**

Project 2 : Classification of SMS into Salary (1) or Non Salary (0) for making base of Salaried Customer so that they can be approached for Salary Personal Loan

- **True positive (TP)** — Model predicted Customer is Salaried (1) and Customer in reality is Salaried (1)
- **False positive (FP)** — Model predicted Customer is Salaried (1) and Customer in reality is NOT Salaried (0)
- **True negative (TN)** — Model predicted Customer is NOT Salaried (1) and Customer in reality is NOT Salaried (0)
- **False negative (FN)** — Model predicted Customer is NOT Salaried (0) and Customer in reality is Salaried (1)

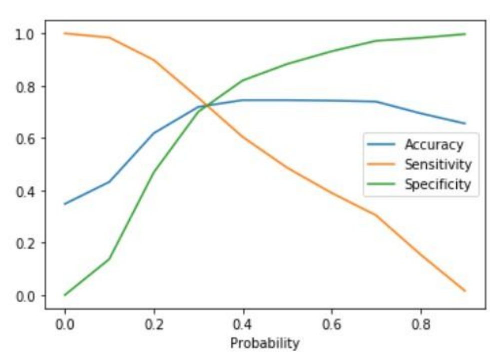
In this case, False Positive is Critical because if we predicted Customer as Salaried, he/she may be considered for Salary Personal Loans which is not correct so **TYPE 1 Error is Critical**

Q6. Basic measures derived from the confusion matrix

Let's take an Example of People invested in FD or Not

- Total Case - 200, True Positive - 20, True Negative - 120, False Positive - 20, False Negative - 40

1. **Accuracy**
    a. (TP+TN)/(TP+TN+FP+FN)
    b. = (20+120) / (20+120+20+40) = (140) / (200) = 70%
2. **Sensitivity or Recall (True positive rate)**
    a. TP/P
    b. TP / TP + FN
    c. Recall quantifies the number of Positive Class predictions made out of all Positive Examples
    d. = (20) / (20+120) = (20) / (140) = 14%
3. **Specificity (True negative rate)**
    a. TN/N
    b. TN / TN + FP
    c. Evaluates model ability to predict True Negative in each category
    d. = (120) / (120+20) = 120 / 140 = 85%
4. **Precision (Positive predicted value)**
    a. TP/(TP+FP)
    b. Precision quantifies number of Positive Class Prediction that actually belong to Positive Class
    c. = (20) / (20+20) = 20/40 = 50%
5. **F-Score (Harmonic mean of precision and recall)**
    a. It measure a single score that balances both the concerns of Precision and Recall
    b. = [2 * Precision * Recall] / [Precision + Recall]
    c. This is Harmonic Mean
6. **What is the RIght Score ?**



Plot and Choose the Trade off between all three parameters

Explain how AUC and ROC curve works?

ROC curve

- An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds.
- This curve plots two parameters: True Positive Rate and False Positive Rate
- True Positive Rate (TPR) is a synonym for **Recall or Sensitivity** and is therefore defined as follows: TP / TP + FN
- False Positive Rate (FPR) is defined as follows: FP / FP + TN
- An ROC curve plots TPR vs. FPR at different classification thresholds.
- Lowering the classification threshold classifies more items as positive, thus increasing both False Positives and True Positives.

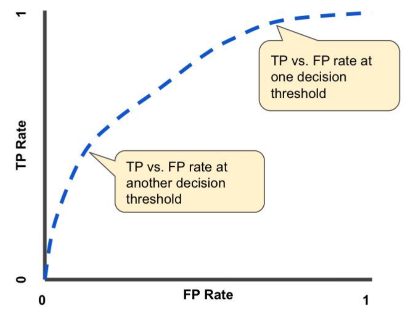


Figure 4. TP vs. FP rate at different classification thresholds.

-

AUC: Area Under the ROC Curve

- AUC stands for "Area under the ROC Curve."
- That is, AUC measures the entire two-dimensional area underneath the entire ROC curve (think integral calculus) from (0,0) to (1,1).
- Sample AUC (Area under the ROC Curve).

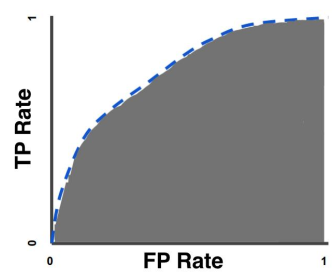


Figure 5. AUC (Area under the ROC Curve).

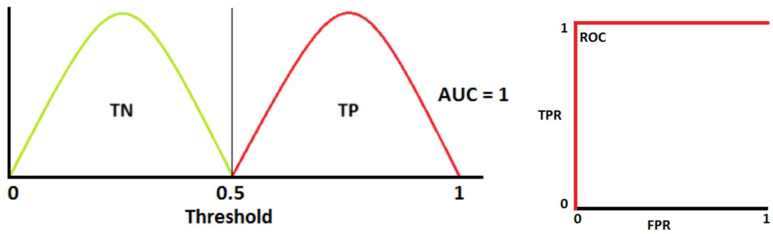
- AUC provides an aggregate measure of performance across all possible classification thresholds.

How to speculate about the performance of the model?

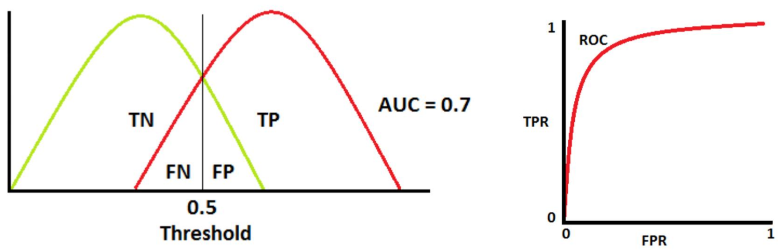
- An **excellent model** has AUC near to the 1 which means it has a good measure of separability.
- A **poor model** has an AUC near 0 which means it has the worst measure of separability. In fact, it means it is reciprocating the result. It is predicting 0s as 1s and 1s as 0s.
- And when AUC is 0.5, it means the model has no class separation capacity whatsoever.

As we know, ROC is a curve of probability. So let's plot the distributions of those probabilities:

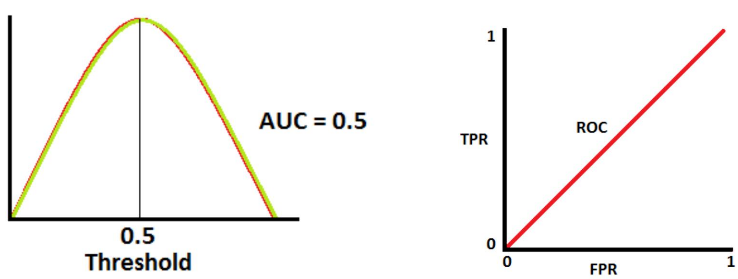
Note: Red distribution curve is of the positive class (patients with disease) and the green distribution curve is of the negative class(patients with no disease).



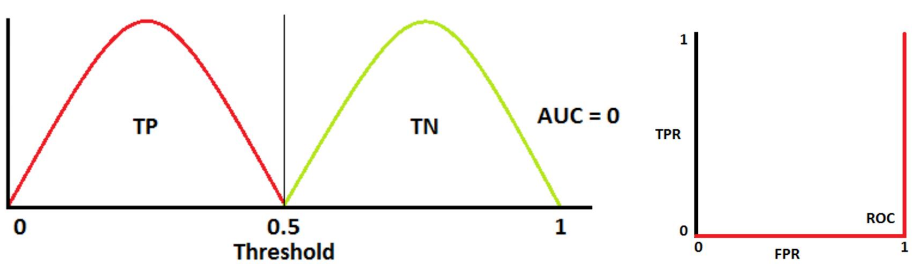
This is an ideal situation. When two curves don't overlap at all, the model has an ideal measure of separability. It is perfectly able to distinguish between positive class and negative class.



When two distributions overlap, we introduce type 1 and type 2 errors. Depending upon the threshold, we can minimize or maximize them. When AUC is 0.7, it means there is a 70% chance that the model will be able to distinguish between positive class and negative class.



This is the worst situation. When AUC is approximately 0.5, the model has no discrimination capacity to distinguish between positive class and negative class.



When AUC is approximately 0, the model is actually reciprocating the classes. It means the model is predicting a negative class as a positive class and vice versa.