Assignment-2

Analysis of Receiver operating characteristic curve

Student id : 18231267

This task involves Analysis of ROC, plotting ROC Curve and finding the Area Under ROC curve for the classification models designed in the first assignment. We need to use the test data for this analysis.

I have used the Knn and SVM models for the given Autoimmune data set in the first assignment.

And continuing the same dataset for this analysis.

About the data set :

Dataset contains 376 observations and 10 features out of which 9 are independent features.

Autoimmune_Disease is the dependent feature which has two values : Positive/Negative.

So basically it's a Binary classification problem. And for the further analysis I split my dataset into 70% training and 30% testing.

ROC Curve and AUROC Curve :

- This is one of the Performance metric for the classification models at different probability threshold settings.
- ROC curve is a probability curve and Area under ROC measures the seperability between 2 different classes.
- Higher this area better the capability of predicting 0's as 0's and 1's as 1's.
- By analogy, higher the AUC , it predicts well between the Autoimmune_disease positive or negative.

The ROC curve is plotted with True Positive Rate(TPR) against False Positive Rate(FPR). Where TPR is on Y- axis and FPR is on X-axis.

All the above values can be computed from the confusion matrix.

Confusion matrix :

This is also one of the metric for effectiveness of our classification model, this tells the how many observations are correctly classified and how many or not.

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

True Positive(TP) : Predicted positive and its True

True Negative(TN) : Predicted negative and its true

False Positive(FP) : Predicted positive and its false

False Negative(FN) : Predicted negative and its false

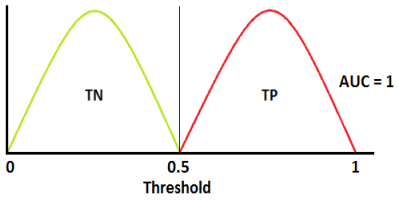
Defining terms used in AUC and ROC curve:

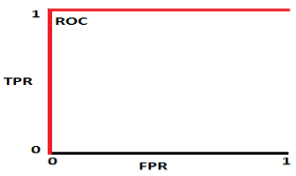
True Positive Rate (TPR/Sensitivity) $= \frac{TP}{TP+FN}$ **Specificity =** $\frac{TN}{TN+FP}$

False Positive Rate(FPR) : FPR = 1- Specificity

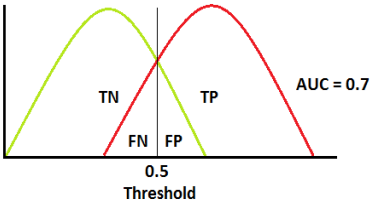
Understanding the ROC Curve :

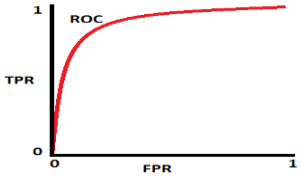
A perfect model which has AUC near to 1 , which means it has good measure of separability. AUC near to 0 indicates the worst measure of separability between the classes. When AUC is 0.5 it means model doesn't have a capability to separate.



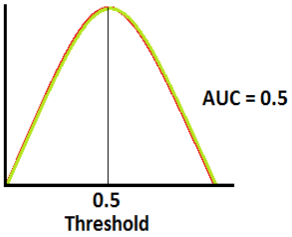


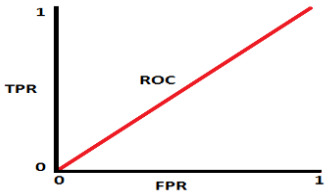
Example 1 : When two curves don't overlap at all means classification model clearly able to distinguish the different classes, this is the ideal case



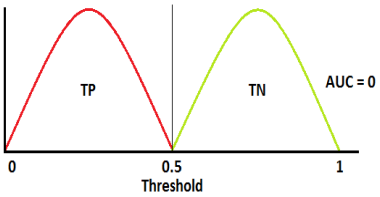


Example 2 : When the two curves overlap , that introduces the FP /FN errors. AUC =0.70 means 70% chance that model will distinguish between the 2 classes.





Example 3: When the two curves overlap, AUC close to 0.5 , means model is not capable of distinguish the different classes



Example 4: in this case AUC is 0 that means model is identifying positive class as negative and negative class as positive.

<u>ROC Curve for SVM and KNN models</u> :

<u>Confusion matrix for Knn</u> : Actual labels :

Predicted :

	Positive	Negative
positive	78	6
nagative	23	18

From the above table : TP=78; TN= 18; FP=6; FN=23

Falsely identified as Negative values are more in our model : that means even though disease is present , our model identified as negative.

Confusion matrix of SVM :

Actual labels :

Predicted :

	Positive	Negative
positive	79	5
nagative	19	22

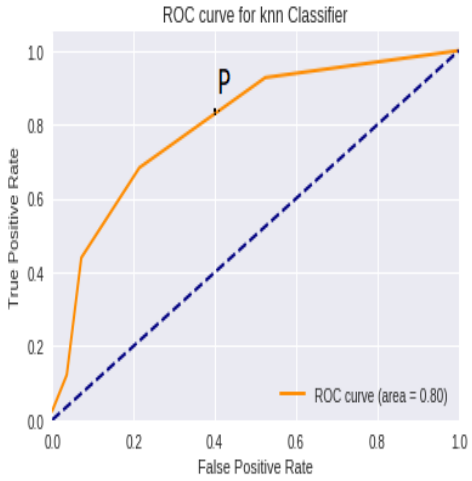
TP=79 ; TN= 22; FP= 5; FN= 19

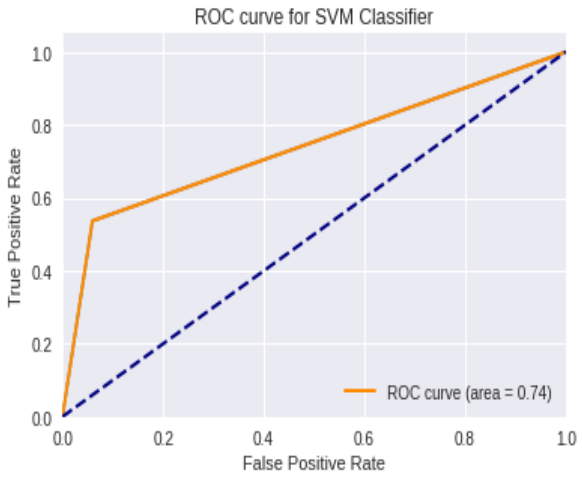
In this model also FN values are more , but compare to the knn model both FP and FN values are less

Observations from ROC curves:

Every point on the ROC curve tells about the sensitivity(TP)and specificity(TN). Its always good to have a high sensitivity and high specificity value.That means any point which has a high TPR and low FPR points on curve satisfies the both High sensitivity and High specificity

Because FPR = 1- Specificity.





1. **Accuracy** :
   Accuracy of the model can be measured with the help of Area Under the Curve
   - For Knn AUROC = 0.80
   - For SVM AUROC = 0.74

2. **Interpretation** :
   For knn ROC curve : let's consider a point P on our ROC curve, which has TPR,FPR pair of (0.40,0.82) that means a sensitivity, specificity value of (0.60,0.82)

That means our model can perform 82% positives(TP) as positives and 60%(TN) negatives as negatives. and remaining 18% is identified as FN and 40% identified as FP.

3. Consider any point (0.1,0.3)on the lower Left side on ROC curve, it fails to predict the True Positives and better with True Negatives. Accordingly we can set the threshold for the application in which TN matters most against the TP
4. Consider any point on the right upper side of the ROC curve i.e around(0.6,0.8) , it fails to identifies the TN and good with the TP.
5. Based on the application and the domain of area we need to chose the optimum cutoff value.

As per our given data its better to have a model which identifies all the True Positives than missing any positives.