Ans 3:
(a)
Relationship between ROC and PR curve is:
We evaluate the performance of machine learning algorithm of ROC and PR curve on same dataset.There are certain theorems which explain the relationship between them.

(Refer from :https://www.biostat.wisc.edu/~page/rocpr.pdf)(these 3 the)

-Theorem 3.1:

For a given dataset of positive and negative examples, there exists a one-to-one correspondence between a curve in ROC space and a curve in PR space, such that the curves contain exactly the same confusion matrices, if Recall not equal 0.

*Here one to one correspondence values should be uniquely mapped with one another.

*Suppose Recall is equal to 0 the the we cant able to find out False positive , also we can't able to find out True negative.

*If recall is equal to 0 the True positive is 0 ,then if we see the formula of Recall then numerator is 0 so we can't able to find out False Negative.So if recall not equal to 0 then we can never recover False Positive and hence We can never find out True negative also.

Formulas: Definitions of metrics

Recall = T P /T P +F N

Precision = T P/ T P +F P

True Positive Rate = T P/ T P +F N

False Positive Rate = F P /F P +T N

*hence this one to one mapping between confusion matrices and points in PR space this acknowledge we can translate curve in ROC space to PR space and vice versa.

-Theorem 3.2:

*For a fixed number of positive and negative examples, one curve dominates a second curve in ROC space if and only if the first dominates the second in Precision-Recall space.this explanation is further in 2nd part of this question.

* if one curve dominate in other curve in ROC same thing will be happened in PR curve and vice versa explained in this theorem.

<u>-Corollary:</u>

Let assume you have a set of given points in PR space ,then there is a PR curve plot that dominates the other valid PR curves that could be constructed with these points.(Thi part will be further explained in part 3 of this question)

*It is observed that ROC would present over optimistic view of algorithm but PR does't ,when there is large skewed in class distribution of data points.

* As we are talking about curves differentiation of PR and ROC in between different algorithms then we observe that PR curve can expose difference in algorithms that are not apparent in ROC space.

.

(b) Prove that a curve dominate in ROC if and only if it dominate in PR curve.

As explained in the paper given, this can be proves by contradiction:

*Claim 1: If a curve dominate in ROC space then it dominate in PR space.

ILet suppose there are 2 curves I and II and assume curve I in ROC space dominates while curve I will never dominate when we translate this curve in PR space.

Let assume point A on curve II and point B on curve I with same Recall have lower Precision.

Precision(A)>Precision(B) yet Recall(A)=Recall(B),Recall is identical to TPR then TPR(A) = TPR(B). Since we know curve I dominates curve II in ROC space then FPR(A) ≥ FPR(B).

(these formulas taken as it is)

As we know that total positives and total negatives are fixed and since TPR(A) = TPR(B): TPR(A) = TPA/ Total Positives ,TPR(B) = TPB /Total Positive,we now have TPA = TPB and thus denote both as TP. Remember that FPR(A) ≥ FPR(B) and

FPR(A) = FPA /Total Negatives, FPR(B) = FPB /Total Negatives

This implies that FPA ≥ FPB because, PRECISION(A) = TP/ FPA + TP ,

PRECISION(B) = TP/FPB + TP ,

we now have that PRECISION(A) ≤ PRECISION(B). But this contradicts our original assumption that PRECISION(A) > PRECISION(B).Hence proved.

*Claim 2:

If a curve dominate in PR space then it dominate in ROC space.

By contradiction let assume we have 2 curves like in claim 1 ,curve I and curve II.And assumn curve I dominate curve II in in PR space but curve I no longer dominate once it translated into ROC space curve.As we saw curve I no longer dominate in ROC space then there should exist a point A on Curve IIs such that point B in curve I both have same TPR and also;

FPR(A)<FPR(B).Also Recall and TPR are same then Recall(A)=Recall(B).Also we saw curve I dominate in PR space so Precision(A)<=Precision(B). As FP(A)>=FP(B) then FPR(A)>=FPR(B) thats contradict our assumption that we assume initially.

RECALL(A) = TPA/ Total Positives, RECALL(B) = TPB/ Total Positives

As we know PRECISION(A) ≤ PRECISION(B) and PRECISION(A) = TP/ TP + FPA, PRECISION(B) = TP/ TP + FPB

we find that FPA ≥ FPB.

Now we have FPR(A) = FPA ,Total Negatives FPR(B) = FPB/ Total Negatives

This concluded that if a curve dominated in PR space over curve II(consider other curves) then same thing will happen if transaction done to ROC space of curves.

(c) It is incorrect to interpolate between points in PR space. When and why does this happen? How will you tackle this problem?

(Refer from :https://www.biostat.wisc.edu/~page/rocpr.pdf)(these 2 points taken as it is)

* <u>To achieve the same we need to do linear interpolation.</u>

  Convex hull of points in ROC space represents the curve that dominates all the valid curve that can be constructed with the same points. Such a curve is called achievable PR curve.

  *<u>Performing linear interpolation leads to overly optimistic estimate of the performance.</u>
  But doing linear interpolation to achieve the same is incorrect . Because as recall varied precision does not varies linearly with recall as it has FP instead of FN in denominator.
  *Following are procedure :
  = Problem of getting the achievable PR curve by converting th points of PR space to ROC space by the help of convex hull.first make the convex hull for ROC curve the

translate it to convex hull of PR space for getting the achievable PR curve.This will resulted in same as required curve but near to ture PR curve.

=Modified interpolation to get an intermediate points A and B which are apart used the following values;FP(A),FP(B),TP(A),TP(B).Then new points to get values are created by adding TP(A)+x , where x values between 1 to TP(B)-TP(A)And then we get FB by increasing the false positive linearly for each new points by local skews where local skewed is formulized as the ration of -FP(B)-FP(A) and TP(B)-TP(A) ,which are corresponding to each point of the data. Then intermediate PR points will be created as per given in paper.