AAPS problem set - 1

1) Here we want FP ↓ and TP↑

$$\text{Precision} = \frac{TP}{TP + FP}$$

as precision ↑ with TP↑ and ↓FP this is suitable

2) (a) T : Spam F : not spam

we want to bring down FN as we don't genuine mails to go to spam folder and miss out on them ⇒ RECALL

b) decreases FP

c) definitely increases

(d) +P decreases

(e) uniformly increase

(f) decreases FN

(g) recall probably decreases

(h) non uniformly decrease

3) (b) → as (a), (b) ∈ imbalanced ∴ accuracy not good metric here

4) (a)

5) FPR Vs TPR for different thresholds

6) positive, TP and FP

7) (c) model classifying every datapoint as belonging to +ve class

↓

8) model is doing bad, make the dataset balanced seems like % of -ve samples more than +ve samples

9) TRUE

FALSE

FALSE

10) TN

11) model 1 — overfit, model 2 - good fit

model 3 - underfit

12) (a) ROC

↓

good for balanced dataset

(b) Precision - recall as good for imbalance

CORRECT ANSWERS BELOW → SIR ↓

1. In a future society, a machine is used to predict a crime before it occurs. If you were responsible for tuning this machine, what evaluation metric would you want to maximize to ensure no innocent people (people not about to commit a crime) are imprisoned?

↓FP ⇒ Precision

P = criminal N − not criminal

2. Consider a classification model that separates email into two categories: "spam" or "not spam." Answer the following questions regarding precision and recall (a.k.a. sensitivity or true positive rate) by playing around with the threshold slider on the demo website here:
   (a) Which is a more relevant performance metric in this case: recall or precision? Explain briefly why.
   (b) Increasing the classification threshold generally increases/decreases FP.
   (c) When the classification threshold increases, precision probably increases/probably decreases/definitely increases/definitely decreases.

choose one

choose one

↑ Shady emails ending up in inbox

(a)

P - spam N − not spam

↑ FN − large no. of emails are classified not spam and actually and are actually spam

↑ FP − large of emails are predicted as spam and actually not spam

↳ means missing out important emails

∴ FP is important ↓ ⇒ Precision (a)

(b) clear from pic that increasing ↓ classfⁿ threshold ⇒ ↑ no. sample classifying as Negative

	Predicted: Positive	Predicted: Negative
Actual: Positive	TP: 6	FN: 2
Actual: Negative	FP: 5	TN: 21

FP ↓ FN ← but doesn't happen uniformly

(b) → decreases FP

(a) probably increases

as decrease in FP doesn't happily steadily, it takes time (slow)

(d) Keeping in mind that $TP + FP + TN + FN = n$, which is the number of samples, when the classification threshold is increased, what happens to the quantity TP?

(e) When the classification threshold is increased, the quantities TN and FN both

$$\underbrace{\text{uniformly/non-uniformly}}_{\text{choose one}} \ \underbrace{\text{increase/decrease}}_{\text{choose one}}.$$

(f) Decreasing the classification threshold generally $\underbrace{\text{increases/decreases}}_{\text{choose one}}$ FN.

(g) When the classification threshold is decreased, recall

$$\underbrace{\text{probably increases/probably decreases/definitely increases/definitely decreases}}_{\text{choose one}}.$$

(d) as classⁿ threshold T ↓,

TP & FP ↓, TN & FN ↑

TP decreases

(e) TN and FN increases non uniformly

(f) decreases FN

(g) probably decreases

(deciding right classification threshold → also based on what is important for the scenario → bringing down FN / FP)

(h) When the classification threshold is decreased, the quantities TP and FP both

$$\underbrace{\text{uniformly/non-uniformly}}_{\text{choose one}} \quad \underbrace{\text{increase/decrease}}_{\text{choose one}}.$$

non-uniformly increases

3. In which of the following scenarios would a high accuracy value suggest that the ML model is doing a good job? Explain your answer briefly.

(a) An expensive and critical hydro-electric turbine operates 23 hours a day. An ML model evaluates vibration patterns and predicts when the turbine is operating without anomaly with an accuracy 99.99%.

(b) You are building an ML tool for a retail company which will predict, based on past purchase history and other demographic information, the high end cellphone that the next buyer will potentially buy from an available 10 high end models. Your ML model has an accuracy of 15%.

(a) Assume inference being made every second → anamdy / not

(answer: Bad model)

Turbine manages says keep watch for every second in 23 hrs (critical)?

```
> (23*60*60)*.9999
[1] 82791.72
> (23*60*60)*(1-.9999)
[1] 8.28
>
```

← for 8 seconds in a day, you are saying model is not good job

```
> (23*60*60)*.9999
[1] 82791.72
> (23*60*60)*(1-.9999)
[1] 8.28
> (23*60*60)*(1-.999999)
[1] 0.0828
> (23*60*60)*(1-.99999)
[1] 0.828
> (23*60*60)*(1-.999999)
[1] 0.0828
>
```

accuracy should be this good for model to be okay / relied upon

(b) model without accuracy

consider random model (making prediction without any model → random guessing → 1/10 → 10% accuracy

∴ our model even though low accuracy certainly better than random model ✓ → good model

(c) A deadly, but curable, medical condition afflicts .01% of the population. Your ML model uses symptoms as features and predicts this affliction with an accuracy of 99.99%.

focus is on FN ✓

Here, population also matters

Model is bad ⇒ analogy ⇒

deadly disease → we don't want

even 1 person to die.

4. Consider two models: A and B, that each evaluate the same dataset. Which one of the following statements is true?
   (a) If model A has better precision and better recall than model B, then model A is probably better.
   (b) If model A has better recall than model B, then model A is better.
   (c) If Model A has better precision than model B, then model A is better.

(a) true (st. fwd)

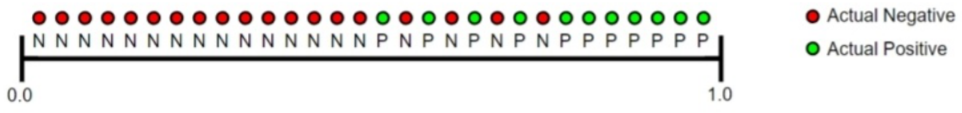
5. An ROC curve is a plot of ___ vs. ___ for different thresholds.

TPR ↓ FPR ↓

6. Lowering the classification threshold classifies more items as positive/negative, thus increasing both TP and FP.

(choose one)

7. AUC (Area under the ROC Curve) provides an aggregate measure of performance across all possible classification thresholds. One way of interpreting AUC is as the probability that the model ranks a random positive example more highly than a random negative example. For example, given the following examples, which are arranged from left to right in ascending order of prediction probabilities:



AUC represents the probability that a random positive (green) example is positioned to the right of a random negative (red) example. AUC ranges in value from 0 to 1. A model whose predictions are 100% wrong has an AUC of 0.0; one whose predictions are 100% correct has an AUC of 1.0.

Suppose we multiplied all of the predictions from a given model by 0.5 (for example, if the model predicts 0.4, we multiply by 0.5 to get a prediction of 0.2), how would it change the model's performance as measured by AUC?

Sudarsan N.S. Acharya sudarsan.acharya@manipal.edu

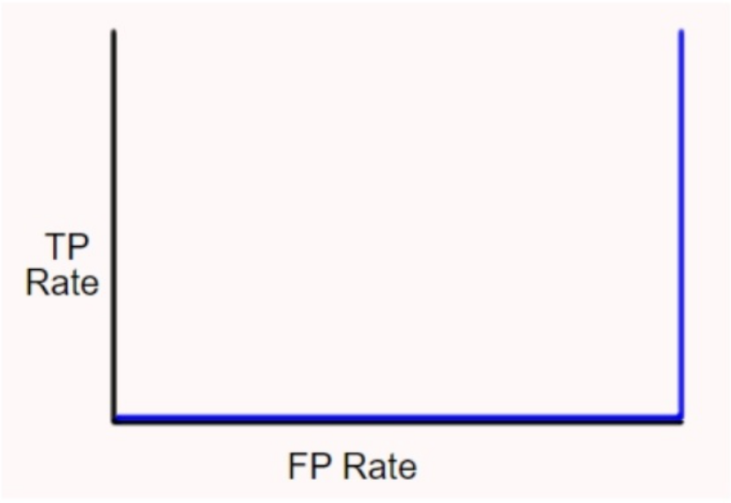
(a) It would make AUC terrible, since the prediction values are now way off.

(b) It would make AUC better, because the prediction values are all farther apart.

(c) No change. AUC only cares about relative prediction probabilities.

(come back here to watch recording he explains more seems important)

8. Your friend shows you his model's ROC curve as follows:

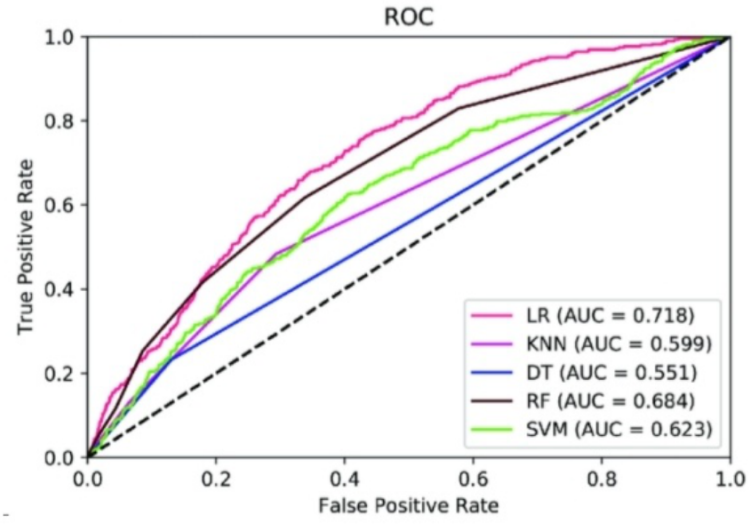


Is your friend's model any good, why? How can you help your friends model go from zero to hero?

this roc curve bad than random guessing → when this is the case → just reverse the predictions → ↑ Roc-AUC :)

↑ y: (1 - probability you got)

9. The figure shows ROC curves for different models.



Classify the following statements as true or false:

- Dashed black line represents random classification. T
- ROC curve for any model can't fall below the dashed black line. F
- The model represented by solid blue line is better than that represent by solid lime.

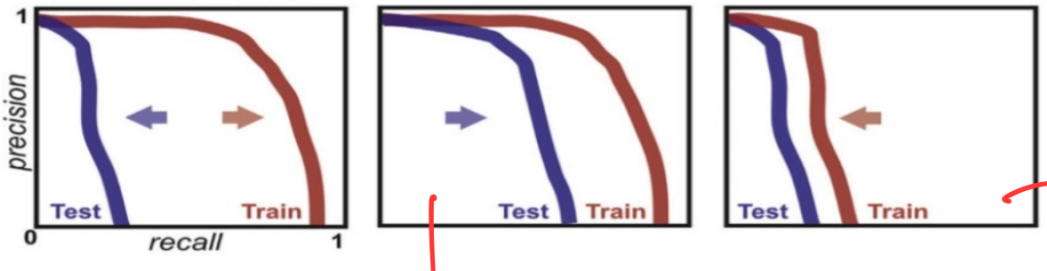
↳ F

10. Which one among TP, TN, FP, FN does not play a role in forming the precision-recall curve? What does the conclusion mean intuitively?

TN, when TN are important

Po-Re curve not useful

11. The figure shows precision-recall curves for different models on the train and test sets.



Identify which model overfits, which one underfits, and which one is a good fit.

overfit

Good fit

underfit

12. Explain which one among area under ROC and area under precision-recall curve would you use for the following scenarios:

(a) Identifying whether a customer will buy a product on discount or not when a customer is equally likely to do so. ROC AUC (Balanced dataset)

(b) Identifying a spam email when generally spam emails constitute 1% of the total emails.

↳ PR - recall (imbalanced dataset)