Applied Probability and Statistics

Assignment - 2

Karthik. L
24105801 6
Big Data Analytics

1. You have tracked the performance of the local meteorologist and complied the following data

P (forecast rain & actual rain) = 0.4

P (forecast rain & no rain) = 0.2.

P (forecast no rain & actual rain) = 0.15.

P (forecast no rain & no rain) = 0.25

a) How often does she forecast rain?

forecast rain = P (forecast rain & Actual rain) + P (forecast rain & no rain)

= 0.4 + 0.2

= 0.6

b) How often does she make a mistake

mistake occurs when forecast doesn't match the actual weather

mistake = P (fr & NO rain) + P(f NO rain & Ar)

= 0.2 + 0.15

= 0.35

c) Given that she just forecast rain, what is the chance that it will actually rain?

$P(Ar \mid fr) = ?$ $\quad \frac{P(Ar \cap fr)}{P(fr)} = \frac{P(\text{actual rain \& forecast rain})}{P(\text{forecast rain})}$

$= \frac{0.4}{0.6} = 0.67$

d) Given that it rains today, what is the probability that she had forecast rain in last night's broadcast?

$P(fr/Ar) = ?$ $\quad \frac{P(fr \cap Ar)}{P(Ar)} = \frac{0.4}{0.55} = 0.72$

P(actual rain) = P (fr + ar) + P(no rain & Ar)

= 0.4 + 0.15

P (Ar) = 0.55

2. Fifty-two percent of the students at a certain college are females. Five percent of the students in this college are majoring in Computer Science. Two percent of the students are women majoring in computer science. If a student is selected at random, find the conditional probability that.

a) The student is female given that the student is majoring in computer science

P(W) = 52% = 0.52

P(CS) = 0.05

P(CS ∩ W) = 2% = 0.02

$P(CS/W) = \frac{P(CS \cap W)}{P(W)} = \frac{0.02}{0.52}$

= 0.038

$P(W/CS) = ?$ $\quad \frac{P(CS/W)\ P(W)}{P(CS)} = \frac{0.038 \times 0.52}{0.05} = 0.3952$

= 0.40

Baye's theorem

b) This student is majoring in computer science given that the student is female

$$P(cs|w) = 0.038\% = \frac{P(cs \cap w)}{P(w)} = \frac{0.02}{0.52} = 0.038$$

3) Suppose that an insurance company classifies people who buy medical insurance from them into one of three classes: good, average & bad risks. As a data scientist for the company, you have access to the following customer data for the calendar year 2023-24.

Class	% of customers	% who had a major health problem
Bad risk	20%	25%
Average risk	40%	20%
Good risk	40%	10%

what is the probability that a new customer will not meet with a major health problem during 2024-25?

$P(Bad) = 0.20$ $\quad P(mh|Bad) = 0.25$

$P(Avg) = 0.40$ $\quad P(mh|Avg) = 0.20$

$P(Good) = 0.40$ $\quad P(mh|good) = 0.10$

$= P(Bad) \times P(mh|bad) + P(Avg) \times P(mh|Avg) + P(good) \times P(mh|good)$

$= 0.20 \times 0.25 + 0.40 \times 0.20 + 0.40 \times 0.10$

$= 0.05 + 0.08 + 0.04$

$= 0.17$

$= 1 - 0.17$

$P(\text{new customer not meet}) = 0.83$

4. Suppose there are four failure modes are respectively 0.002, 0.002, 0.01 and 0.001. Given that it is a structural failure, there is a 25% chance the plane will crash. the crash probabilities given the other three failure model are 30%, 90% & 10% respectively for engine control system & human Error failure model. If a plane has crashed, what is the Probability that it was due to a control system failure.

4 failure models for Single Engine plane

Structural $P(S) = 0.002$ $\quad P(cr/S) = 25\% = 0.25$

engine $P(e) = 0.002$ $\quad P(cr/e) = 30\% = 0.30$

control system $P(cs) = 0.01$ $\quad P(cr/cs) = 90\% = 0.90$

human Error $P(he) = 0.001$ $\quad P(cr/he) = 10\% = 0.10$

$P(cs/cr) = ?$ $\quad \dfrac{P(cr/cs)\ P(cs)}{P(cr)} = \dfrac{0.90 \times 0.01}{0.0102} = \underline{\underline{0.8823}}$

Baye's theorem

$P(cr) = P(S) \times P(cr/s) + P(e) \times P(cr/e) + P(cs) \times P(cr/cs) + P(he) \times P(cr/he)$

$= 0.002 \times 0.25 + 0.002 \times 0.30 + 0.01 \times 0.90 + 0.001 \times 0.10$

$= 0.0005 + 0.0006 + 0.009 + 0.0001$

$\boxed{P(cr) = 0.0102}$

5. A robot, which only has a camera as a sensor, can either be in one of two locations: L_1 or L_2. the robot doesn't know Exactly where it is but based on all past observations, the robot thinks that there is an 80% chance that it is in L_1 and a 20% chance that it is in L_2. location L_2 is the only one that has a window. the robot's vision algorithm detects a window but its image recognition algorithm is not perfect; the probability of observing a window given there is no window at its location is 0.2 & the probability of observing a window given there is a window is 0.9. After incorporating the observation of a window, what is the robot's updated probability that it is in 1) L_1 2) L_2?

$P(L_1) = 0.80$ $\quad P(w/L_1) = 0.2$ $\quad P(L_1|w) = ?$

$P(L_2) = 0.20$ $\quad P(w/L_2) = 0.9$ $\quad P(L_2|w) = ?$

$$P(w) = P(L_1) \times P(w/L_1) + P(L_2) \times P(w/L_2)$$

$$= 0.8 \times 0.2 + 0.2 \times 0.9$$

$$= 0.16 + 0.18$$

$$P(w) = 0.34$$

Baye's theorem

$$P(L_1/w) = \frac{P(w/L_1)\ P(L_1)}{P(w)} = \frac{0.2 \times 0.8}{0.34} = 0.470$$

$$P(L_2/w) = \frac{P(w/L_2)\ P(L_2)}{P(w)} = \frac{0.9 \times 0.2}{0.34} = 0.529$$

Given that 52% chance will be there is window in location L_2

7. In a future society, a machine is used to predict a crime before it occurs. If you were responsible for tunning this machine, what Evaluation metric would you want to maximi to Ensure no innocent people (people not about to commit a crim are imprisoned.

→ Here to Ensure no innocent people are imprisoned we should make the False positive, ~~to zero so that any~~ thus the False positive low definitely precision high for the machine

8. Consider a classification model that separates email into two categories: "spam" or "not spam". Answer the following questions regarding precision & recall (a.k.a. Sensitivity or TPR) by playing around with the threshold sliders on the demo website here:

a) which is more relevant performance metric in this case: recall or precision? Explain breifly why?

False positive:- It's not a spam email but is detect & sent as a spam email

False Negative:- It's a spam Email but it's detect & sent as a not a spam Email

b) Increasing the classification threshold generally increases/decrease FP.

→ FP is decreases

c) when the classification threshold increase, precision

→ If you increase the threshold, False positive decrease then automatically Precision is definitely increases

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?

→ when the classification threshold is increased, True positive decreases

e) when the classification threshold is increased, the quantities TN & FN both

→ classification threshold is increased, the True Negative & False negative both will increase uniformly.

f) Decreasing the classification threshold generally increases/ decreases FN .....

→ If the classification threshold decreased, False Negative will be decrease.

g) when the classification threshold is decreased, recall

→ If the classification threshold is decreased, False Negative also decreases then recall definitely increases

h) when the classification threshold is decreased, the quantities TP & FP both

→ the classification threshold decrease, the True positive & False positive both will increased.

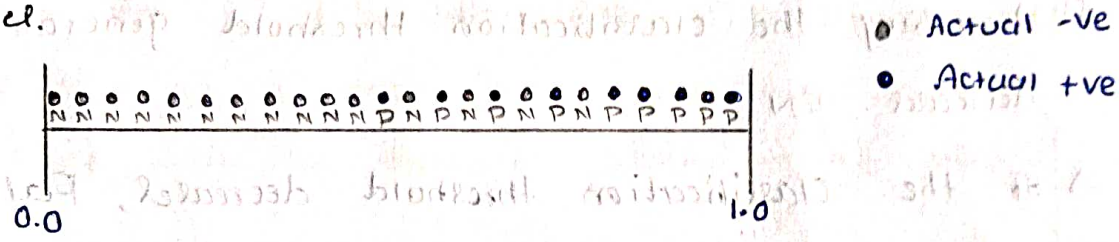
10. consider two models: A and B, that Each Evaluate the same dataset, which one of the following statements is true?

a) If model A has better precision & better recall than model B, then model A is probably better. (✓)

11. An ROC curve is a plot of True positive Rate vs False positive Rate for different thresholds

12. Lowering the classification threshold classifies more items as positive / negative, thus increasing both True Positive False positive

13. AUC (Area under the ROC curve), one way of interpreting AUC is as the probability that the model ranks a random +ve example more highly than a random -ve ex. which are arranged from left to right in ascending order of prediction probabilities.

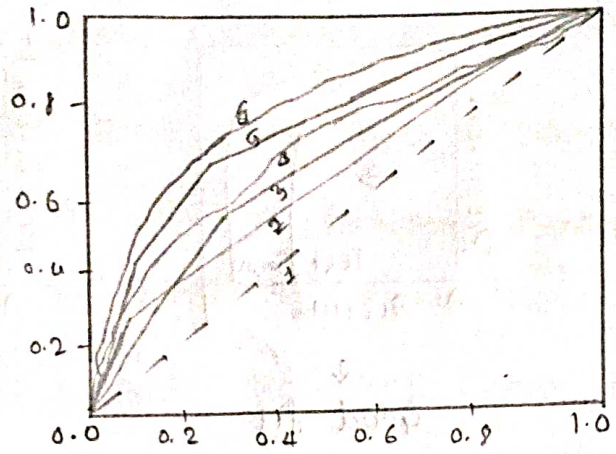


random positive (Blue) to right of a random -ve (Pencil) ex. AUC ranges from 0 to 1. A model predictions are 100% wrong. has an AUC of 0.0; one whose predictions are 100% correct AUC of 1.0.

Suppose we multiplied all of the predictions from a given model by 0.5 (for ex, 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?

→ No change. AUC only cares about relative prediction probabilities.

15. The figure shows Roc curves for different models



1. RF AUC = 0.698
2. DT AUC = 0.551
3. KNN AUC = 0.599
4. SVM AUC = 0.623
5. LR AUC = 0.718

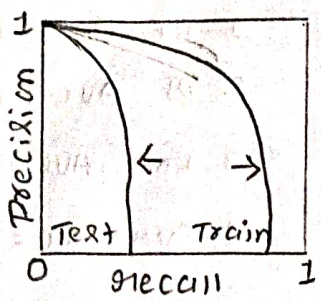
- Dashed black line represents random classification → True
- Roc curve for any model can't fall below the dashed black line → False (Area must be high)
- The model represented by solid blue line is better than that represent by solid line. → False

16. which one among TP, TN, FP, FN does not play a role in forming the Precision - recall curve? what does the conclusion mean intuitively?

→ True Negative, for precision = $\frac{TP}{TP+FP}$, recall = $\frac{TP}{TP+FN}$

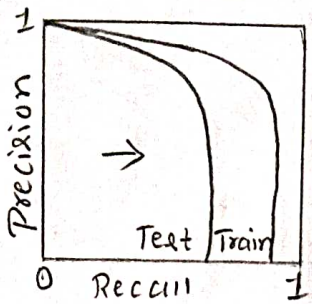
So if you plot a graph for Precision & recall need only TP, FP, & FN there is no need of TN so that's why TN does not play a role in forming the Precision-recall curve.

17. Identify which model overfit, underfit & which one is good fit.



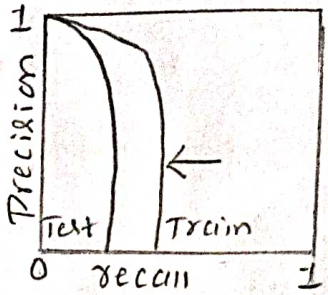
↓

Over fit



↓

Good fit



↓

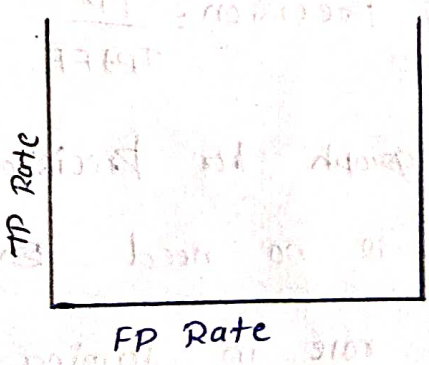
Under fit

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

a) Identifying wheather a customer will buy a product on discount or not when a customer is equally likely to do so → ROC

b) Identifying a spam email when generally spam emails constitute 1% of the total emails → Precision, Recall

14. 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?

Yes it's is good we can take decisions in the opposite way of the model predictions, this way friend model became zero to hero.

6. Suppose a person needs a blood transfusion. we show the compatibility chart x means compatibility

a) what is the probability that a random person will be able to donate to another random person given no information about blood types of either the giver or the receiver?

→ P [Donor (All) and recipient (who all can take)

mutual Exclusive events and Donor and recipents independent so that add all the Donors and recipents

$P(O^- \cap (\text{who all can take})) + P(O^+ \cap (O^+, A^+, B^+, AB^+)) + P(A^- \cap (A^-, A^+, AB^-, AB^+)) + P(A^+ \cap (A^+, AB^+)) + P(B^- \cap (B^-, B^+, AB^-, AB^+)) + P(B^+ \cap (B^+, AB^+)) + P(AB^- \cap (AB^-, AB^+)) + P(AB^+ \cap AB^+)$

$P(O^- \cap (\text{who all can take})) = 0.066 [0.066 + 0.374 + 0.063 + 0.357 + 0.015 + 0.085 + 0.006 + 0.034] = \underline{0.066}$

$P(O^+ \cap (\text{who can take})) = 0.374 [0.374 + 0.357 + 0.085 + 0.034]$

$= \underline{0.3179}$

$P(A^- \cap (\text{who can take})) = 0.063 [0.063 + 0.357 + 0.006 + 0.034]$

$= \underline{0.0289}$

$P(A^+ \cap (\text{who can take})) = 0.357 [0.357 + 0.034] = \underline{0.1395}$

$P(B^- \cap (\text{who can take})) = 0.015 [0.015 + 0.085 + 0.006 + 0.034] = \underline{0.0021}$

$P(B^+ \cap (\text{who can take})) = 0.085 [0.085 + 0.034] = 0.0101$

$P\ [AB^- \cap (\text{who can take})] = 0.006\ [0.006 + 0.034] = 0.0002$

$P\ [AB^+ \cap (\text{who can take})] = 0.034\ [0.034] = 0.0011$

$P\left[\text{Donor (All) and } (\text{who all can take})\right] =$

$0.066 + 0.3179 + 0.0289 + 0.1395 + 0.0021 + 0.0101 + 0.0002$

$+ 0.0011 = 0.56$

$P\left(\text{Donor (All) \& who all can take}\right) = 56\%$

b) Given the above data, what can you say about blood transfusion policy in a hospital regarding blood drives target & blood tranfusion priority?

O negative is the universal donor the blood drives should focus more on o negative blood group the transfusion policies is prioritize in keeping stocks of O⁻ blood for Emergencies. blood transfusion priority should be according to the rear blood types likes AB^- and look for other matching or reciving groups here if the blood is in stock of O negative if the recipient need of O negative should be priortized first.

c) In a battle field hospital a soldier is brought in for immediate blood transfusion. only blood type A+ is available in the supply for immediate use. we do not know the wounded soldier's blood type. there are two other soldiers present who are willing to donate their blood. we have time to do one blood typing before time becomes critical. what should we do?

→ The given case where we have limited time to transfusion of blood. because of the limited time and only A +ve blood type is available ~~is~~ in blood bank. so it is best to do the blood sampling for the wounded soldier here we can get out of the blood group but if the blood type is A +ve (or) AB +ve of the wounded soldier we can immediate start transfusion, so if we got like this we have a two probabilites. if the blood group comes out to be of different blood type we are helpless.

9) 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.

→ 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~~ anomaly with an accuracy 99.99%