

COMPSCI 762 2022 S1 Week 4 Solution

Luke Chang

April 7, 2022

Question 1

- Check Week 2 lecture slides from page 40 to 48.

Question 2

- The weather is either Shower or Clear. This is a binary classification task. Let Shower be Positive and Clear be Negative:

		Predicted		Total
		P	N	
Actual	P	4	1	5
	N	3	2	5
Total	7	3		10

- Acc. = $\frac{6}{10} = 0.6$
- Precision (P) = $\frac{TP}{TP+FP} = \frac{4}{4+3} \approx 0.571$
- Recall (R) = $\frac{TP}{TP+FN} = \frac{4}{4+1} \approx 0.8$
- $F_1 = 2 \frac{P \times R}{P+R} = 2 \times \frac{0.571 \times 0.8}{0.571 + 0.8} \approx 0.667$
- **Note:** A model with high Recall may also have high FPR (Type I Error).
- ROC curve:

Class	Prediction	Thresholds					
		0	0.2	0.4	0.6	0.8	1
P	0.95	1	1	1	1	1	0
N	0.85	1	1	1	1	1	0
P	0.78	1	1	1	1	0	0
P	0.66	1	1	1	1	0	0
N	0.6	1	1	1	1	0	0
P	0.55	1	1	1	0	0	0
N	0.53	1	1	1	0	0	0
N	0.52	1	1	1	0	0	0
N	0.51	1	1	1	0	0	0
P	0.4	1	1	1	0	0	0

- Counting TP and FP:

Table 1: Counting TP and FP

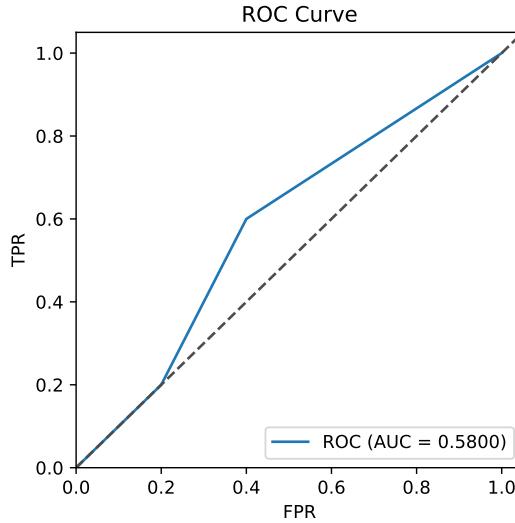
Threshold	0	0.2	0.4	0.6	0.8	1
TPR	1	1	1	0.60	0.2	0
FPR	1	1	1	0.4	0.2	0

- Sort the results:

Table 2: Sort the results

Threshold	1	0.8	0.6	0.4	0.2	0
TPR	0	0.2	0.6	1	1	1
FPR	0	0.2	0.4	1	1	1

- Final ROC plot:



- AUC is the area under the ROC curve.

Question 3

- After we find the optimal parameters for the model using CV, we have to train the model one more time with all data points.
- We don't want to do a train-test split. We want to use all the data.
- The actual test set is hidden from us. When we submit the model, Kaggle will report a ranking score. However, we can not access the test set.