

Introduction to Machine Learning

Day 5: Metrics Beyond Accuracy

Joanna Bieri DATA301

Get out your handwritten notes!

Accuracy

ASK

Our wine model is 88.7 percent accurate
on 300 validation wines.

Is that good?

Hint. How many of the 300 are actually good wines?

Accuracy

ANSWER

- Only **41** of the 300 are good.
- Always saying “not good” scores **$259 / 300 = 0.863$** .
- So our model is just **2.3 points** better than doing nothing.
- When the thing you want is rare, accuracy is mostly the base rate.
Q1 in your notes.

Confusion Matrix

PREDICT

wine	1	2	3	4	5	6	7	8
truth	1	0	1	0	1	0	0	0
predicted	1	0	0	0	0	0	0	1

```
confusion_matrix(truth, predicted)
```

What comes out? All four numbers.

Hint. Rows are the truth, columns are the prediction, and 0 comes first.
Go wine by wine and drop each one into its box.

Confusion Matrix

ANSWER

[[4 1]	TN FP
[2 1]]	FN TP

Wine 1 is a TP. Wines 3 and 5 are good wines the model missed, so FN is 2. Wine 8 is a false alarm, FP 1. The other four are TN. Q2 in your notes.

Precision and Recall

PREDICT

```
[[4 1]  
 [2 1]]
```

```
precision_score(truth, predicted)  
recall_score(truth, predicted)
```

What are the two numbers?

Hint. Precision: of the wines the model called good, how many were?

Recall: of the wines that really are good, how many did it find?

Precision and Recall

ANSWER

```
precision_score(truth, predicted)    0.5  
recall_score(truth, predicted)      0.333
```

Precision is the right column: 1 of 2. Recall is the bottom row: 1 of 3.
Same matrix, two questions.

Precision and Recall

ASK

A shop buys every wine the model calls good.

A judge tastes every wine the model shortlists.

Who wants precision? Who wants recall?

Hint. For each one, which mistake costs more: a bad wine called good, or a good wine missed?

Precision and Recall

ANSWER

- **The shop wants precision.** A wrong yes is a case of bad wine they paid for.
- **The judge wants recall.** A wrong yes costs one sip. A missed wine might have been the winner.
- F1 treats the two mistakes as equal, and neither of these people does. Q3 in your notes.

The Threshold

PREDICT

```
y_prob = np.array([0.9, 0.4, 0.6, 0.2, 0.5])
```

```
(y_prob >= 0.5).astype(int)
```

```
(y_prob >= 0.7).astype(int)
```

What are the two arrays?

Hint. `>=` gives True or False. `astype(int)` turns True into 1.

The Threshold

ANSWER

```
(y_prob >= 0.5).astype(int)    [1 0 1 0 1]  
(y_prob >= 0.7).astype(int)    [1 0 0 0 0]
```

The first one is exactly what `predict()` does. Watch the 0.5: `>=` keeps it. Raising the threshold can only take yeses away.

predict_proba

PREDICT

`model.predict_proba(X)` gave this for three wines:

wine	column 0	column 1
A	0.9	0.1
B	0.3	0.7
C	0.55	0.45

What does each column mean? What would `predict()` say for each wine?

Hint. Each row is one wine. The two numbers in a row add to 1.

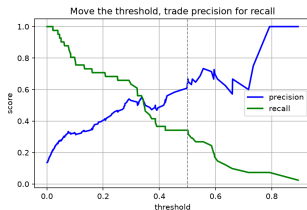
predict_proba

ANSWER

- Column 0 is P(not good), column 1 is P(good).
- `[:, 1]` keeps column 1: `[0.1 0.7 0.45]`. That is `y_prob`.
- At 0.5, `predict()` says A no, B yes, C no.
- C was a close call, and `predict()` hides that.

The Threshold

ASK



On validation at 0.5: precision 0.67, recall 0.34. We moved to 0.3 and recall went up to 0.61.

What happened to precision? Why?

Hint. At 0.3 the model says yes more often. Are the new yeses wines it was sure about?

The Threshold

ANSWER

- Precision **fell** to 0.52.
- The new yeses are the wines the model was least sure about, so more of them are wrong.
- Down: catch more, more false alarms. Up: fewer false alarms, miss more. Q4 in your notes.

ROC AUC

PREDICT

wine	truth	score
A	0	0.1
B	0	0.4
C	1	0.35
D	1	0.8

```
roc_auc_score(truth, score)
```

What comes out?

Hint. AUC is the chance a random good wine scores higher than a random bad wine.

Step 1: pair each good wine (C, D) with each bad wine (A, B).

Step 2: in each pair, did the good wine get the higher score? Count the yeses.

ROC AUC

ANSWER

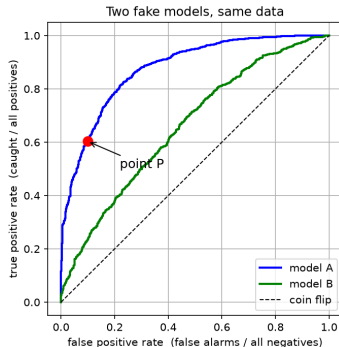
good wine	bad wine	good one higher?
C 0.35	A 0.1	yes
C 0.35	B 0.4	no
D 0.8	A 0.1	yes
D 0.8	B 0.4	yes

```
roc_auc_score(truth, score)    0.75
```

3 of 4 pairs. Notice there is no threshold anywhere in that table, which is why AUC does not change when you move the threshold.

Reading an ROC Curve

ASK



Which model is better? What does point P mean in words?

Reading an ROC Curve

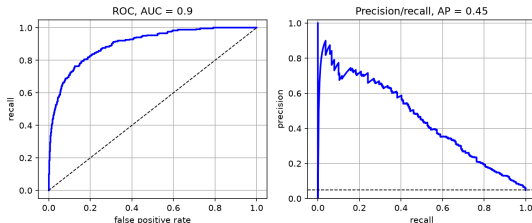
ANSWER

- **Model A.** At every false alarm rate it catches more. AUC 0.88 against 0.66.
- **Point P:** at one threshold, model A catches 60 percent of the positives and raises a false alarm on 10 percent of the negatives.
- At that same 10 percent, model B catches only 23 percent.
- The curve is every threshold at once. That is why ROC is for comparing models, not for picking a threshold.

ROC or PR?

ASK

Same fake model, 1 positive in 20



Same fake model, same predictions. Only 1 case in 20 is a positive.

Why does one plot look great and the other so-so?

Hint. False positive rate divides by all the negatives, 19 for every positive. Precision divides by what the model said yes to.

ROC or PR?

ANSWER

- **Both are true.** They count the same false alarms against different things.
- Against 19 negatives per positive, lots of false alarms is still a small rate, so ROC looks great.
- Against the model's yeses, the same false alarms are a big share, so PR looks so-so.
- Rare positives and you care about them: report **average precision**. Our wine did the same thing, AUC 0.87 and AP 0.53. Q6 in your notes.

Calibration

ASK

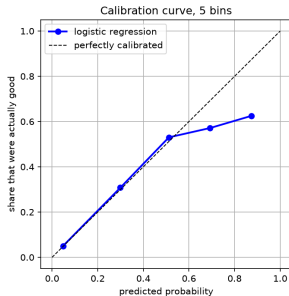
The model gives 100 wines a score of 0.3.

If that number means what it says,
how many of them are good?

Hint. 0.3 is supposed to be a 30 percent chance.

Calibration

ANSWER



About 30. A model whose probabilities mean what they say is **calibrated**. Ours is: wines it scored near 0.3 were good 31 percent of the time. Q7.

The Test Set, Once

ASK

We chose the threshold on validation.

Then we scored the test set, once.

Can we go back and change the threshold now?

Hint. If the test score helps you choose, what job is the test set doing?

The Test Set, Once

ANSWER

- **No. Report it and stop.**
- If the test score helps you choose, the test set is just a second validation set, and nothing honest is left.
- A small difference either way is normal. Ours: validation precision 0.52 and recall 0.61, test 0.54 and 0.63. Q9 in your notes.

Choosing a Threshold

ASK

A screening test for a rare cancer.

Threshold above or below 0.5?

Precision or recall first?

Hint. Which is worse for the patient: a missed cancer, or an extra scan?

Choosing a Threshold

ANSWER

- **Below 0.5, recall first.** A missed cancer is far worse than an extra scan.
- Choose the threshold on validation data, then check it on the test set once.
- This is Problem 2 and Problem 5 in the homework.

Questions?

Questions?