Classifier Evaluation

Procheta Sen

UNIVERSITY OF LIVERPOOL

We have a classifier/model/system…

- How **good** is it?
- Levels of **goodness**
  - **Absolute goodness**: When we run our trained model on the “wild” it does what we expect it to do
    - no way of knowing *before* we deploy the model and when we do know, too late!
  - **Relative goodness**: We have a small representative sample of **test data**
    - We compare the output produced by our classifier on this test dataset (gold standard) and measure how well it resembles the labels in the dataset

Gold Standard

- A dataset that we use for evaluation purpose. Also known as **test data**.

- Each test instance in the test data has its correct label annotated.

- Numerous measures exist (as we will shortly see) to **compare** the **predicted** labels by the trained classifier and actual (**target**) labels in the test dataset

- Never train on test data!!!

Confusion Matrix (error matrix)

		Actual	
		Actual YES(+)	**Actual** NO(-)
Predicted	**Predicted** YES(+)	True Positives (TP)	False Positives (FP)
	Predicted NO(-)	False Negatives (FN)	True Negatives (TN)

The matrix makes it easy to see if the system is **confusing** two classes

Definitions

- **True Positive**

We predicted as positive and it is indeed positive

- **True Negative**

We predicted as negative and it is indeed negative

- **False Positive**

We predicted as positive but it turns out to be negative

- **False Negative**

We predicted as negative but it turns out to be positive

	Actual YES(+)	**Actual** NO(-)
Predicted YES(+)	True Positives (TP)	False Positives (FP)
Predicted NO(-)	False Negatives (FN)	True Negatives (TN)

Example

- We have a set of images some of which contain a car
- We want to detect photos with a car
- Positive class (**+**): all photos with a car
- Negative class (**-**): all other photos

Example: detecting cancer

- Let assume we trained a classifier to detect cancer based on some features.
- Predicting YES means we predict that the patient has cancer.
- We predicted the patient as having cancer but further tests revealed that the patient does not have cancer
  - False Positive
- We predicted the patient as not having cancer (so no further tests were done) but the patient died with cancer!
  - False Negative
- The moral of the story
  - FP and FN have very different importance in real-world data mining tasks.

Evaluation measures

- Accuracy
- Precision
- Recall
- F-score
- Many more (see e.g. https://en.wikipedia.org/wiki/Confusion_matrix)

	Actual YES(+)	**Actual** NO(-)
Predicted YES(+)	True Positives (TP)	False Positives (FP)
Predicted NO(-)	False Negatives (FN)	True Negatives (TN)

Evaluation measures: Accuracy

Answers to

what proportion of all objects were **correctly classified**?

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

	Actual YES(+)	**Actual** NO(-)
Predicted YES(+)	True Positives (TP)	False Positives (FP)
Predicted NO(-)	False Negatives (FN)	True Negatives (TN)

Evaluation measures: Precision

Answers to

what proportion of **predicted Positives** is truly Positive?

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

	Actual YES(+)	**Actual** NO(-)
Predicted YES(+)	True Positives (TP)	False Positives (FP)
Predicted NO(-)	False Negatives (FN)	True Negatives (TN)

Evaluation measures: Recall

Answers to

what proportion of **actual Positives** is correctly classified?

$$\text{Recall} = \frac{TP}{TP + FN}$$

	Actual YES(+)	**Actual** NO(-)
Predicted YES(+)	True Positives (TP)	False Positives (FP)
Predicted NO(-)	False Negatives (FN)	True Negatives (TN)

What is more important: Precision or Recall?

Application 1 (cancer detection)

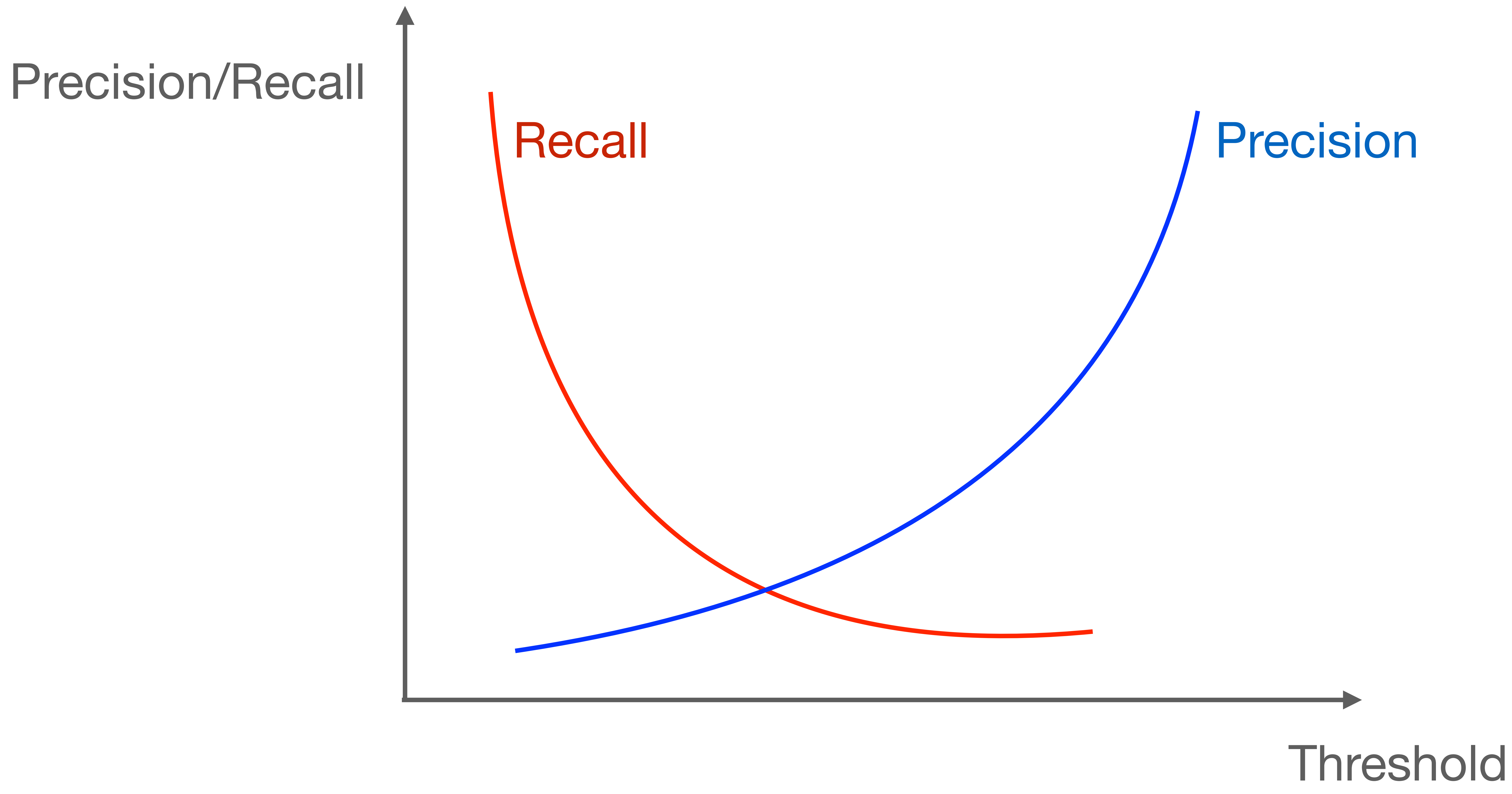
- Positive: has cancer
- Negative: healthy
- We want to detect as many patients with cancer as possible
- We want to have **high recall**

Application 2 (product recommendation)

- Positive: relevant products
- Negative: non-relevant products
- We want to make sure that almost all recommended products are relevant to the user
- We want to have **high precision**

There is a trade-off between precision and recall:
improving precision often results in lowering recall and vice versa

Precision-Recall Trade-off



By simply varying the threshold of our cancer detector we can get a high precision OR low recall system. There is a *trade-off*.

Evaluation measures: F-score

$$\text{F-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

	Actual YES(+)	Actual NO(-)
Predicted YES(+)	True Positives (TP)	False Positives (FP)
Predicted NO(-)	False Negatives (FN)	True Negatives (TN)

F-score is the **harmonic mean** between precision and recall

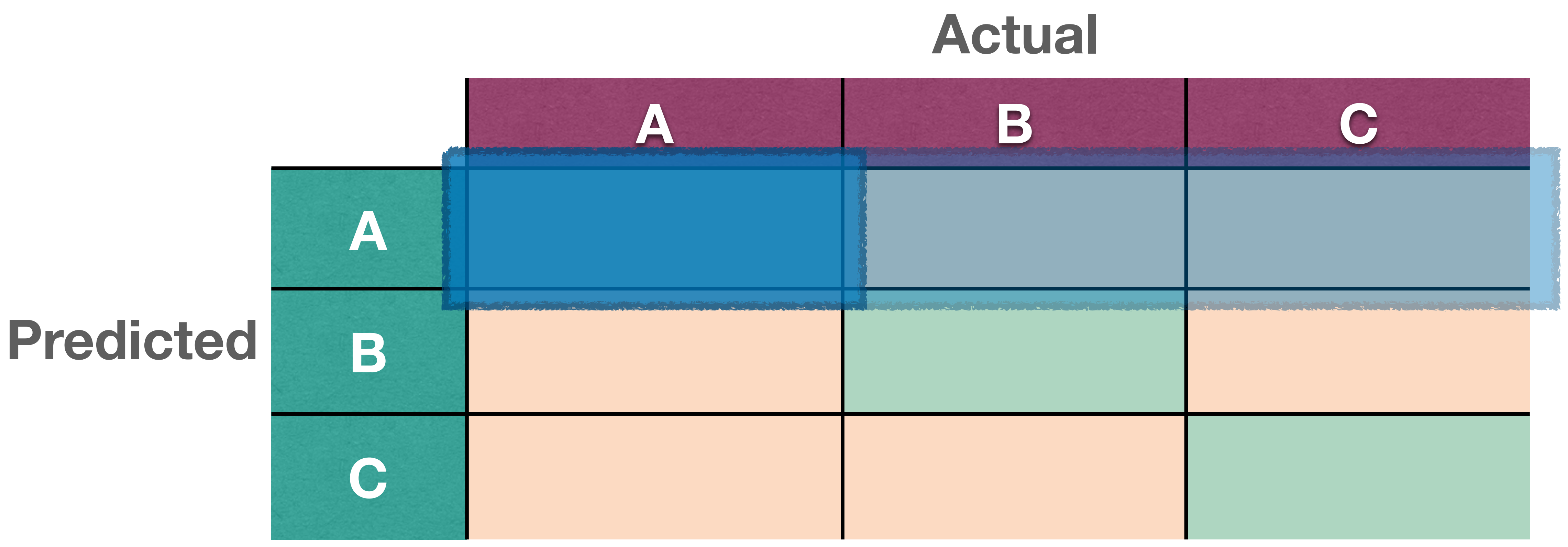
$$\text{F-score} = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}}$$

- F-score is in between precision and recall
- F-score gives a larger weight to lower numbers

Evaluation measures for multiple classes

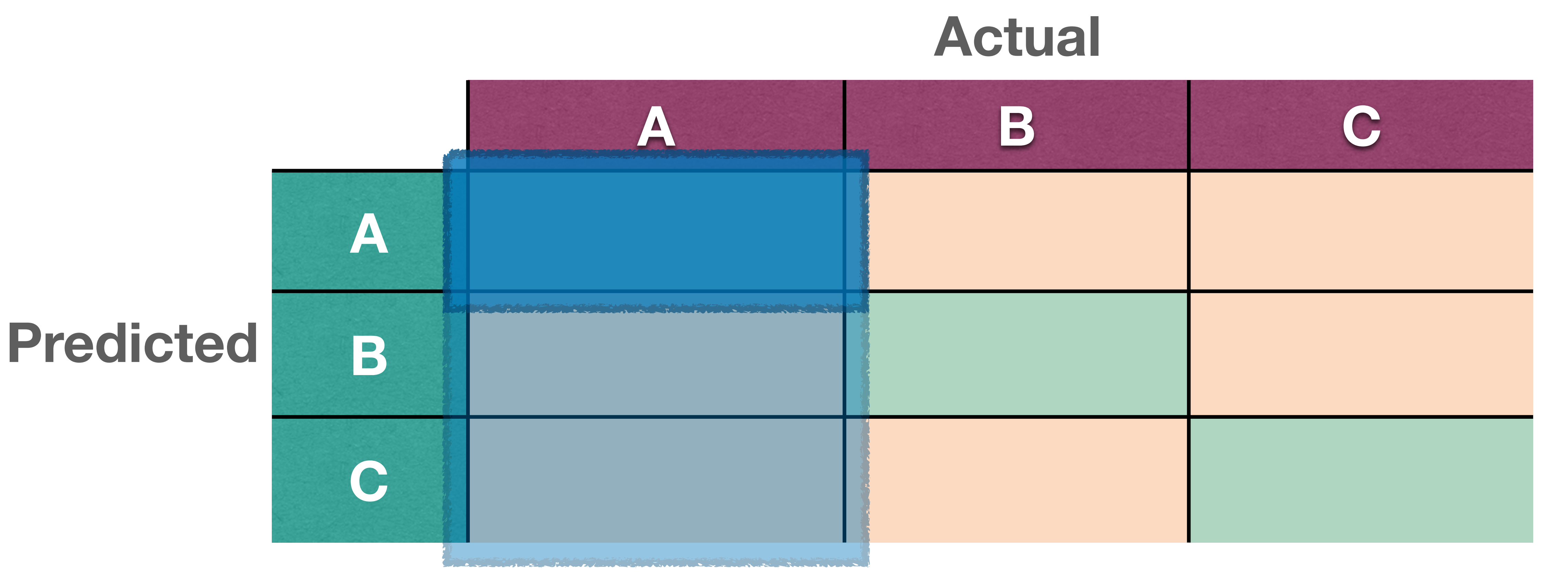
Precision for a class A:

$$\frac{\text{no. objects correctly classified } A}{\text{no. objects classified } A}$$



Recall for class A:

$$\frac{\text{no. objects correctly classified } A}{\text{no. objects that belong to class } A}$$



Evaluation measures for multiple classes

F-score for class A

$$\text{F-score}_A = \frac{2 \times \text{Precision}_A \times \text{Recall}_A}{\text{Precision}_A \; + \; \text{Recall}_A}$$

$$\textbf{Macro F-score} = \frac{1}{C}\sum_{i=1}^{C} \text{F-score}_i,$$

where C is the number of classes and F-score_i is the F-score for class i