Evaluating Systems: precision-recall

Prof. Chris McCool

Overview

- Evaluating systems
  - Precision-Recall

Evaluating Systems

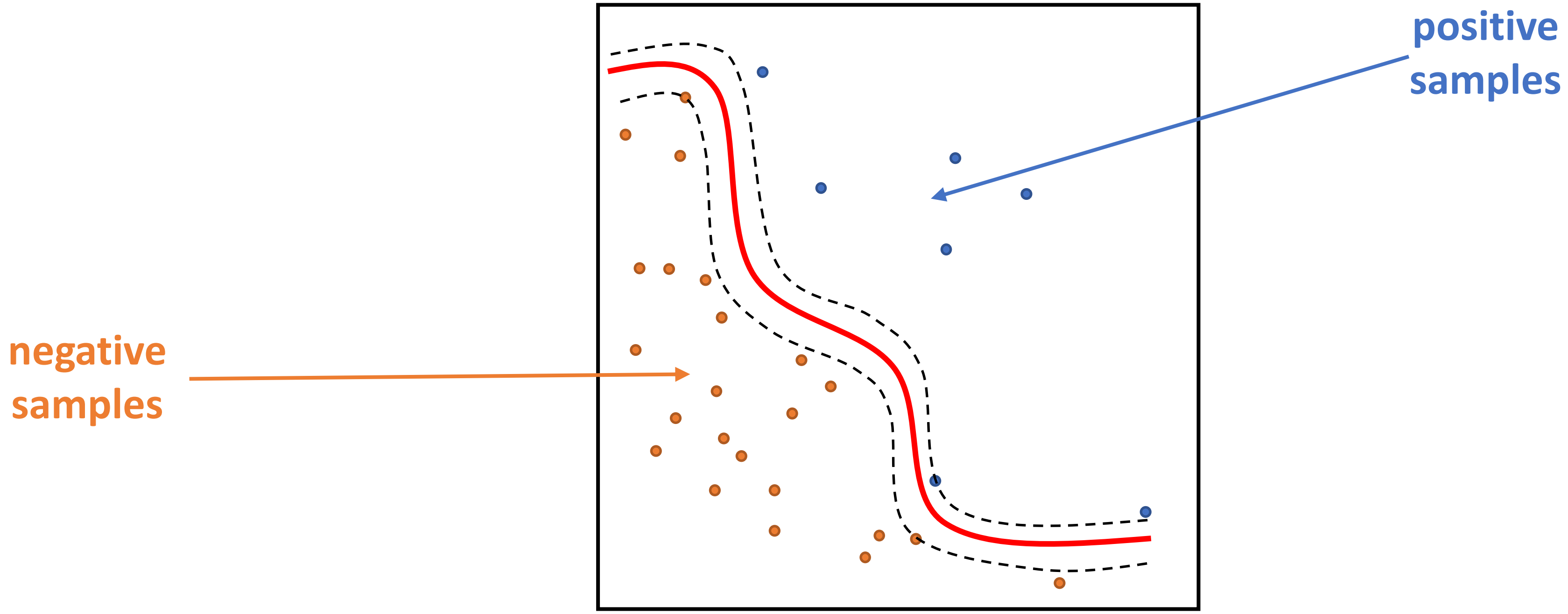
- Essential to understanding how well the system works
  - Does it meet requirements

- Almost all classifiers produce a score
- The score is “thresholded” to make a decision
- Consider a two-class classifer with parameters $\boldsymbol{\theta}$ that produces a score for the i-th sample

$$s_i = h(\boldsymbol{x}_i|\boldsymbol{\theta})$$

Evaluating Systems

When we evaluate we have **positive samples** and **negative samples**



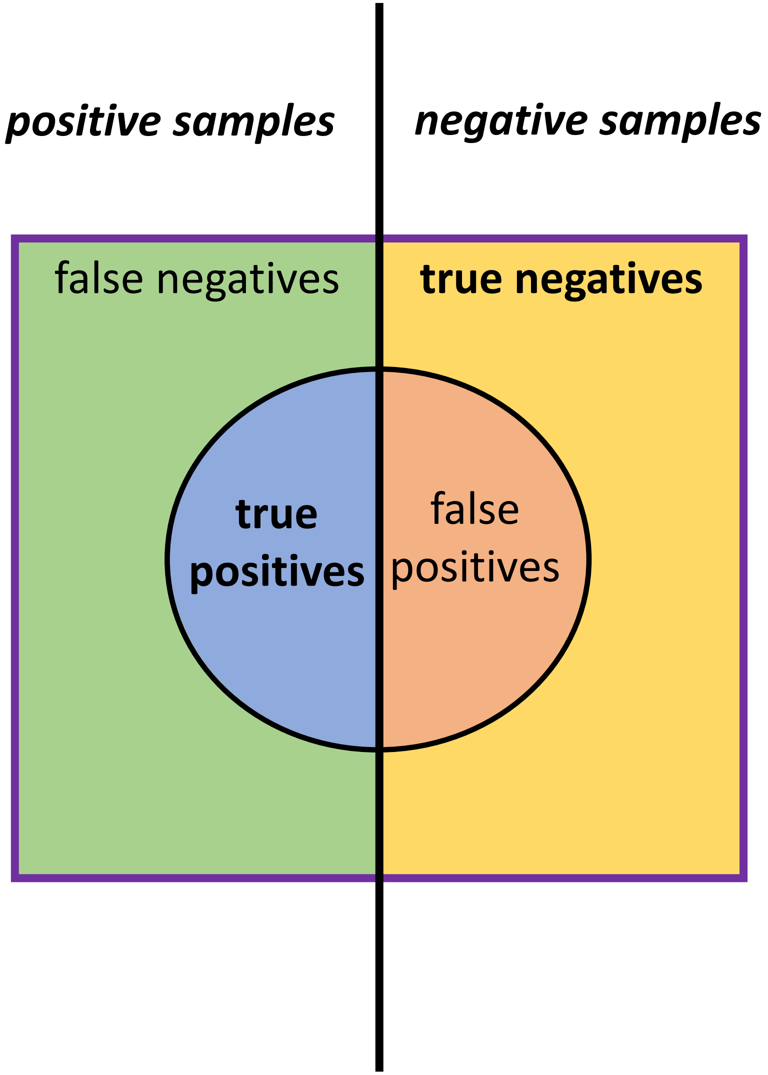
Precision-Recall Curves

Precision

- the ratio of samples found that are relevant
- true positives (TP)
- true positives (TP) + false Positives (FP)
- $P = \frac{TP}{TP+FP}$

Recall

- the ratio of relevant samples found
- true positives (TP)
- true positives (TP) + False negatives (FN)
- $R = \frac{TP}{TP+FN}$

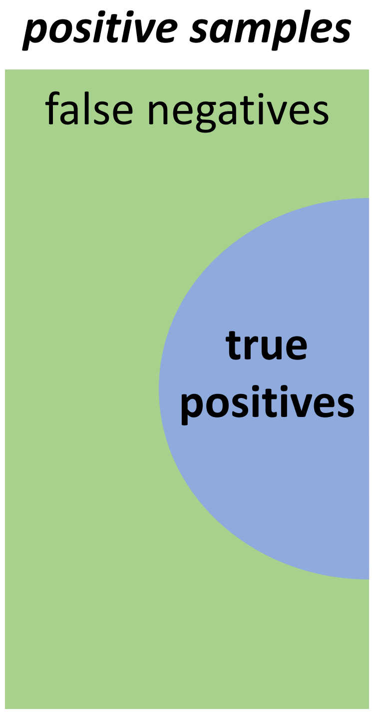


Precision-Recall Curves

Relevant Samples (Positive Samples)

- true positives
  - the ones that we labelled as being *true* and are *true samples*
- false negatives
  - the ones that we labelled as being *false* and are *true samples*

Want to maximise true positives and minimise false negative

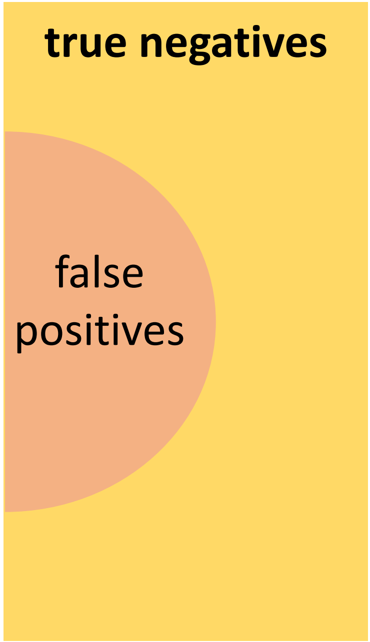


Precision-Recall Curves

Not Relevant Samples (Negative Samples)

- true negatives
  - the ones that we labelled as being *false* and are *false samples*
- false positives
  - the ones that we labelled as being *true* and are *false samples*

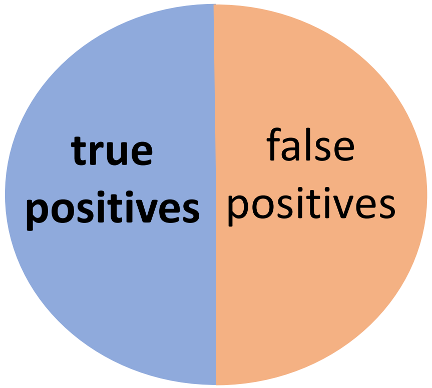
Want to maximise true negatives and minimise false positives



Precision-Recall Curves

Precision

- the ratio of samples found that are relevant
- true positives (TP)
- true positives (TP) + false Positives (FP)
- $P = \frac{TP}{TP+FP}$

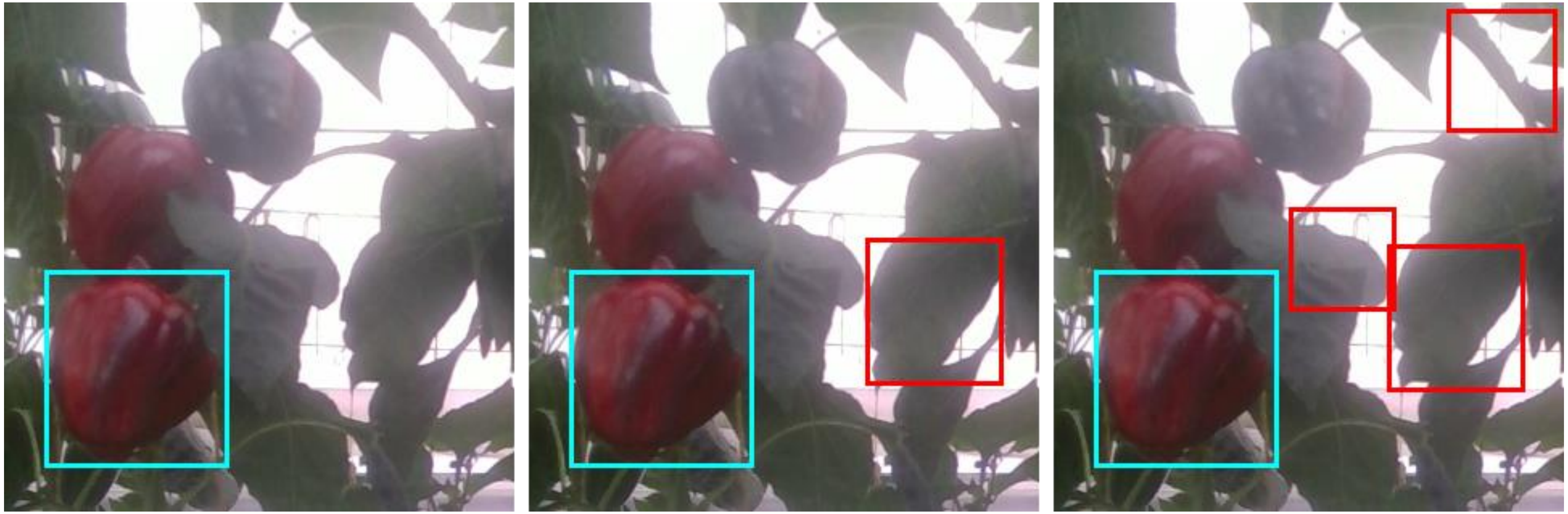


Precision-Recall: practical example

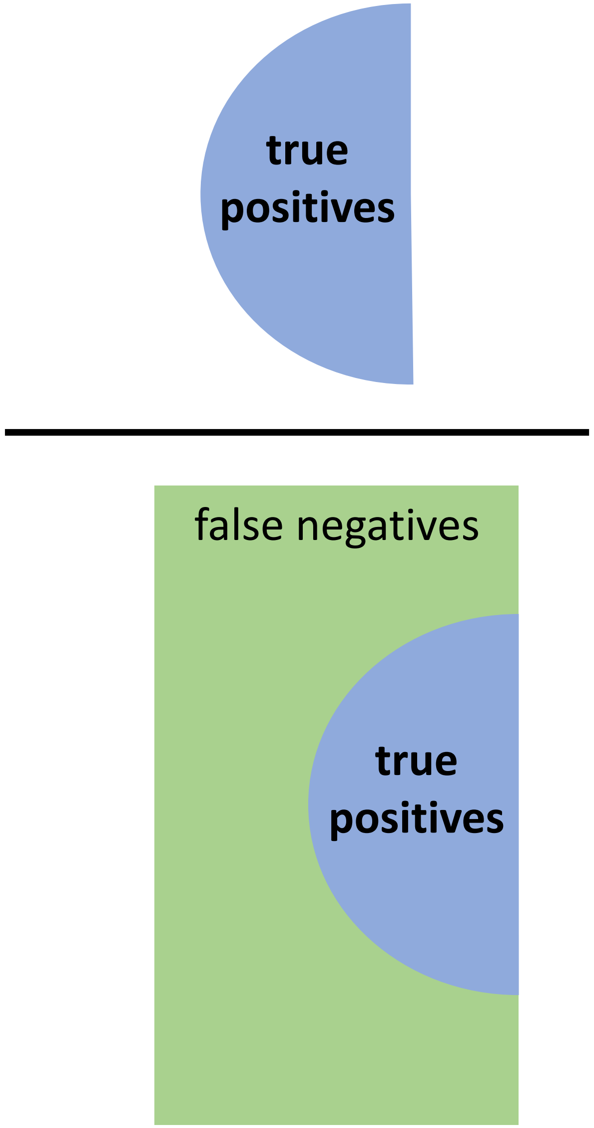
- How accurately do we detect objects in the frame.

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

- Red are detections that are not the class of interest (FP).



Precision-Recall Curves



Recall

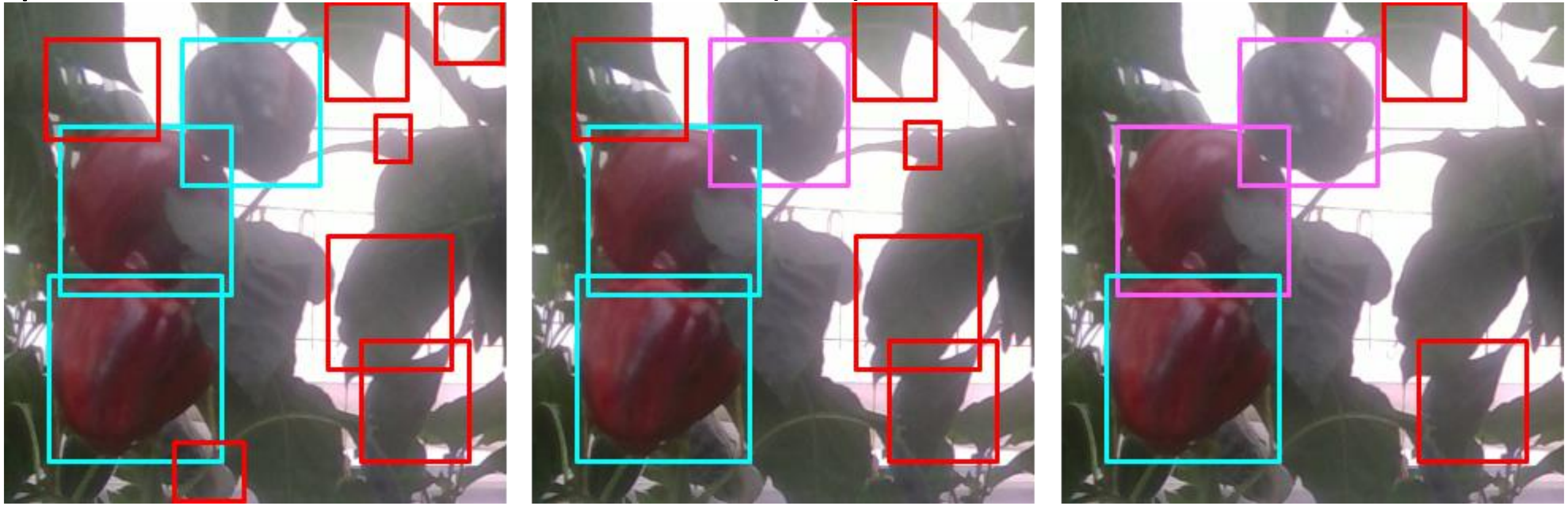
- the ratio of relevant samples found
- true positives (TP)
- true positives (TP) + False negatives (FN)
- $R = \frac{TP}{TP+FN}$

Precision-Recall: practical example

- How many of the objects in the frame do we find.

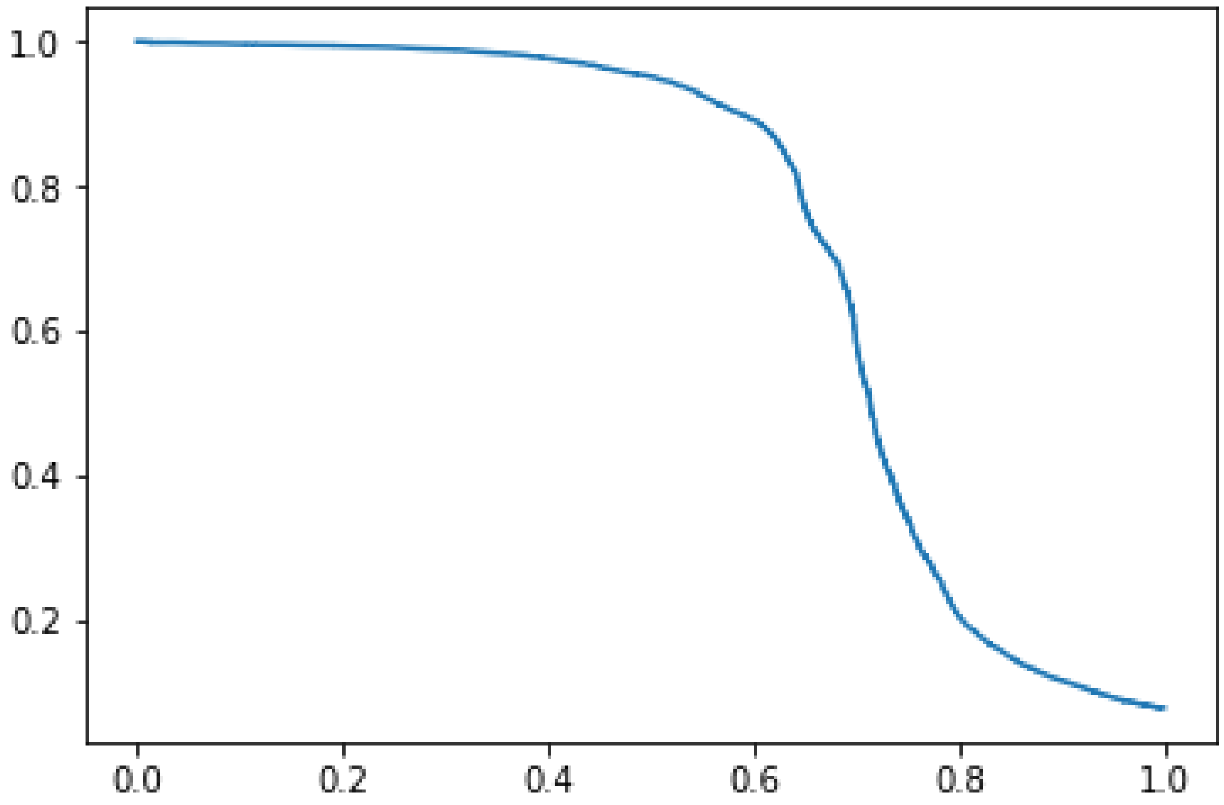
$$Recall = \frac{TP}{TP + FN}$$

- Limitation include higher false detections.
- Purple are detections that were missed (FN).



Precision-Recall Curves

- At different thresholds $[\tau_1, \tau_2, \ldots, \tau_T]$ the system produces different results
  - false positives
  - false negatives
  - true positives
  - true negatives
- Could we summarise this with 1 number?



Precision-Recall Curves

- Mixing precision and recall to produce a single number
- F-score

$$\boxed{F_1 = 2 * \frac{P * R}{P + R}}$$

- Generalised as the F_β-score

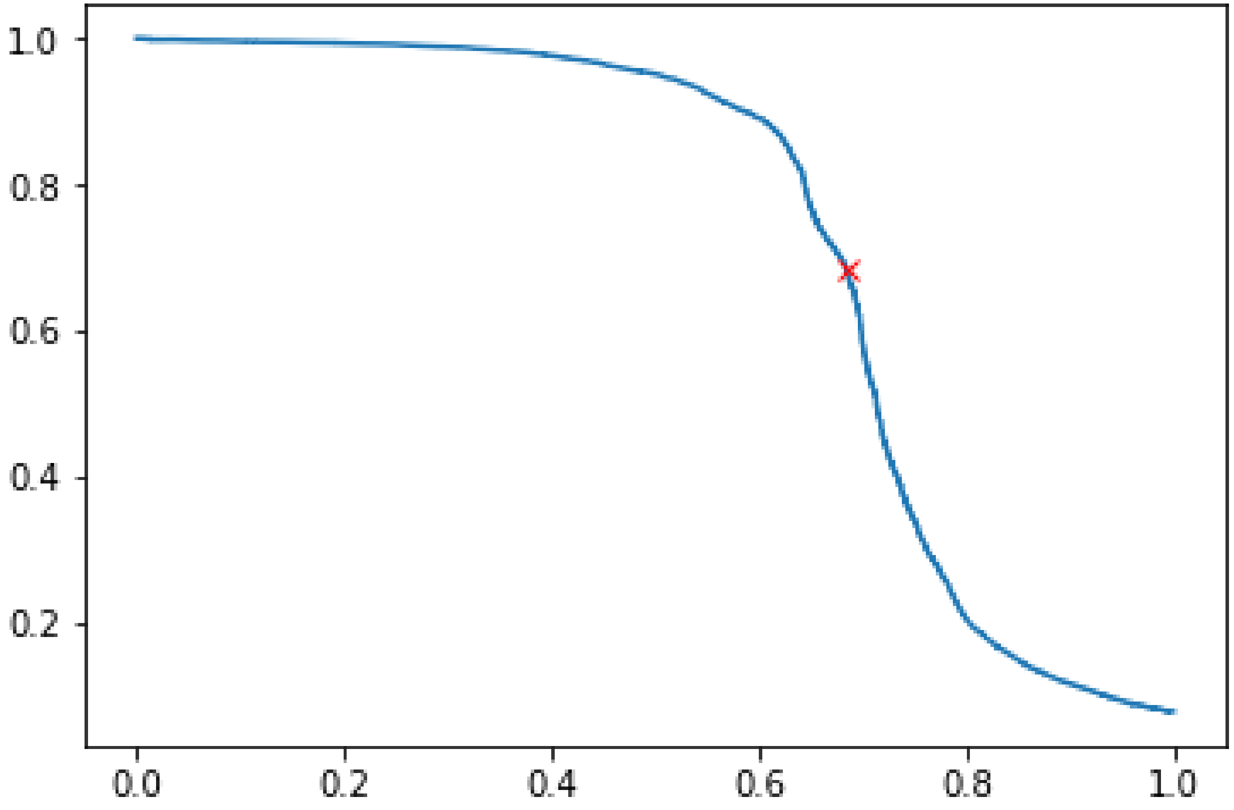
$$F_\beta = (1 + \beta^2) * \frac{P * R}{\beta^2 P + R}$$

Precision-Recall Curves

- F-Score

$$F_1 = 2 * \frac{P * R}{P + R}$$

- At the point where precision equals recall



Evaluating Systems

Splitting your data

- Training Set
  - Used to train the model/system

- Validation Set
  - Used to validate the model, choose the best one (from the training)
  - Use this to get a threshold

- Evaluation Set
  - How well did my system go, with a threshold!
  - Use it once!

- Simply way to split is 2:1:1 ratio

Evaluating Systems

Splitting your data

- Validation Set vs Evaluation Set
  - How well did my system go, with a threshold!
  - Will the threshold now represent the same point as on the validation set?
    - That is, will it be the point that precision equals recall on the evaluation set?

  - For a deployed system, why is it important to have a good threshold?

Evaluating Our Outlier Detector

- Now we know about precision, recall and how we might get a threshold
  - From the data!!
- Let’s show an active example of this for our Gaussian class.

Overview

- Evaluating systems
  - Precision-Recall