

CMPT 825: Natural Language Processing

Text Classification - Evaluation

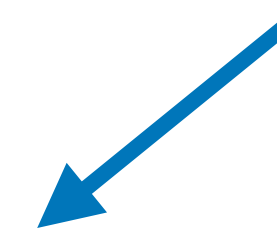
Fall 2020
2020-09-21

Adapted from slides from Danqi Chen and Karthik Narasimhan

Evaluation

- Consider binary classification
- Table of predictions

Confusion matrix



		<i>Truth</i>	
		Positive	Negative
<i>Predicted</i>	Positive	100	5
	Negative	45	100

- Ideally, we want:

		<i>Truth</i>	
		Positive	Negative
<i>Predicted</i>	Positive	145	0
	Negative	0	105

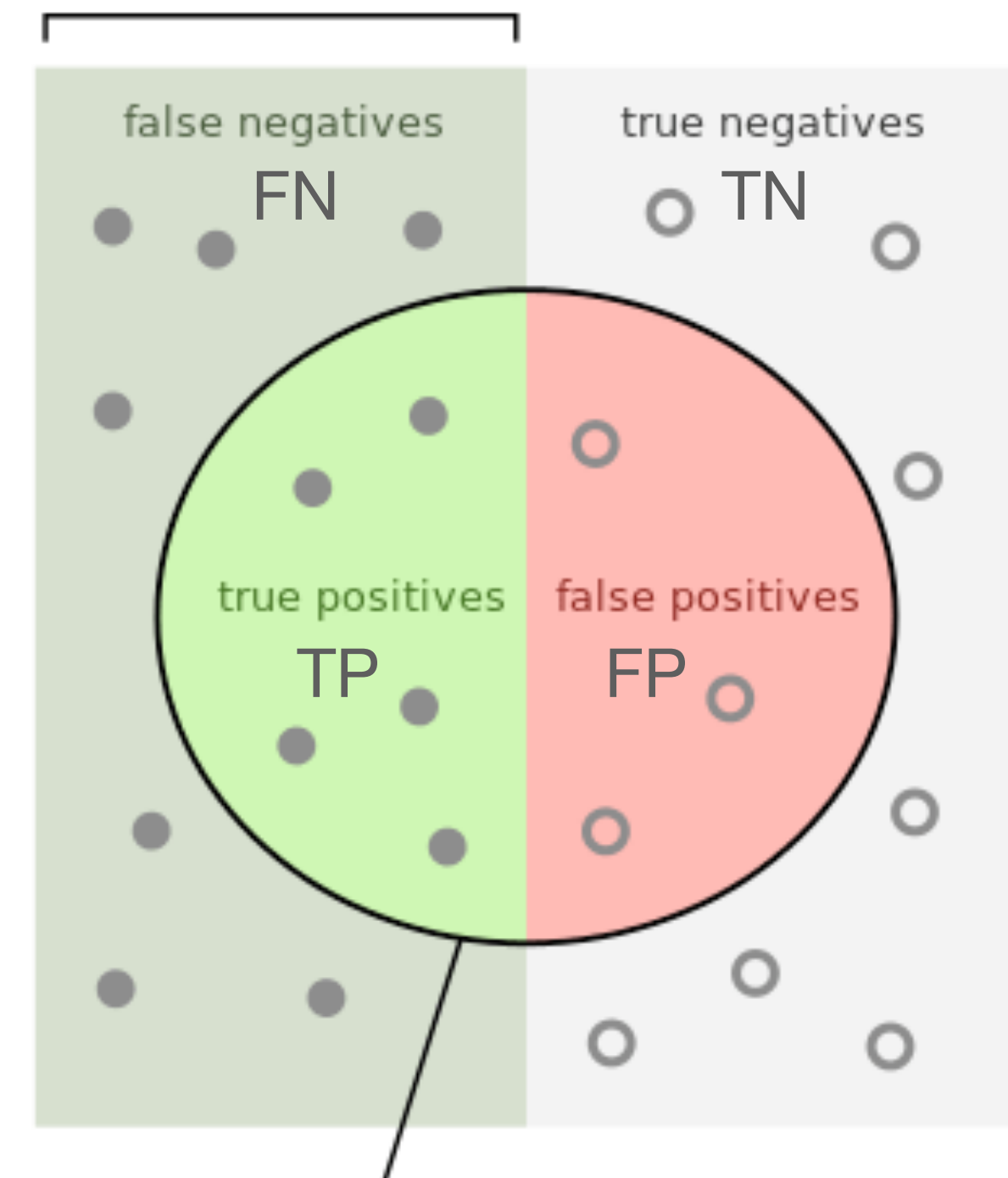
Evaluation Metrics

		Truth	
		Positive	Negative
Predicted	Positive	100 TP	5 FP
	Negative	45 FN	100 TN

- True positive (TP): Predicted + and actual +
- True negative (TN): Predicted - and actual -
- False positive (FP): Predicted + and actual -
- False negative (FN): Predicted - and actual +

$$\text{Accuracy} = \frac{TP + TN}{\text{Total}} = \frac{200}{250} = 80\%$$

Actual positives



Predicted positives

(image credit: wikipedia)

Evaluation Metrics

</

- True positive (TP): Predicted + and actual +
- True negative (TN): Predicted - and actual -
- False positive (FP): Predicted + and actual -
- False negative (FN): Predicted - and actual +

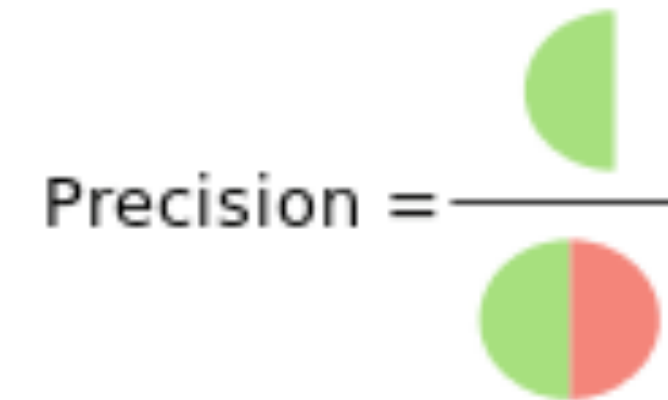
Coarse metric

$$\text{Accuracy} = \frac{TP + TN}{\text{Total}} = \frac{200}{250} = 80 \%$$

Precision and Recall

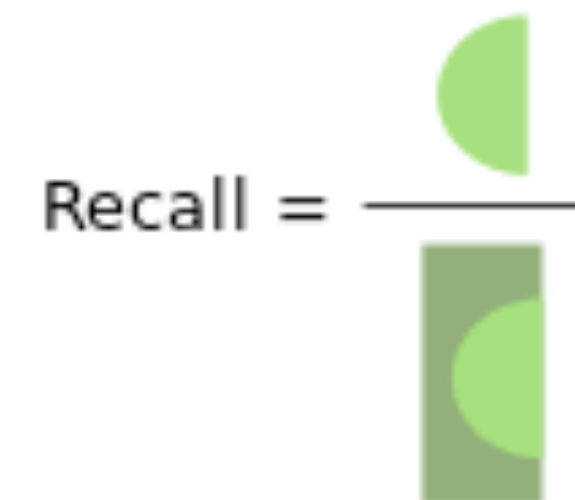
- Precision: % of selected classes that are correct

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

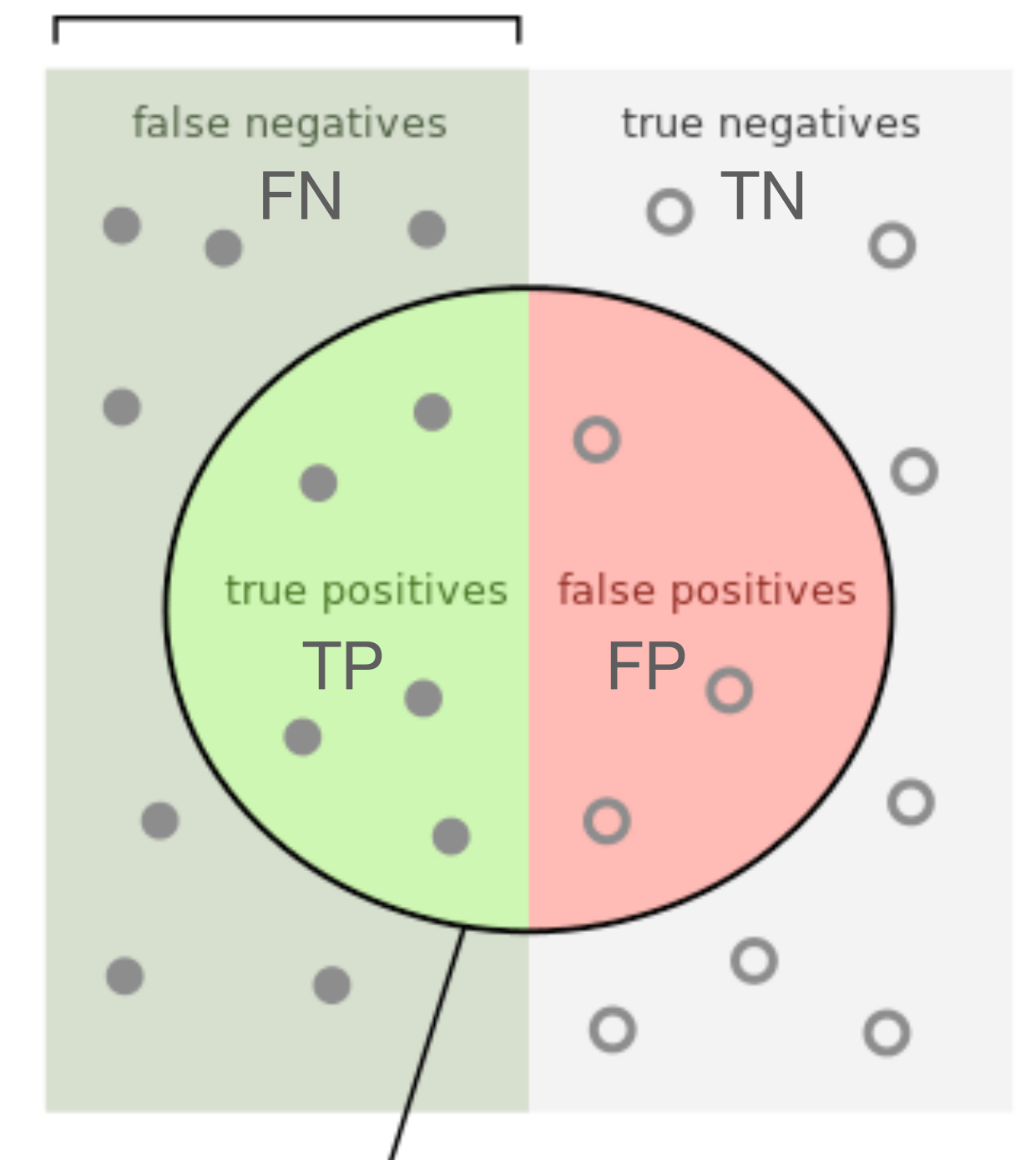


- Recall: % of correct items selected

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



Actual positives (relevant)



(image credit: wikipedia)

Precision and Recall

- Precision: % of selected classes that are correct

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

$$\text{Precision}(-) = \frac{TN}{TN + FN}$$

- Recall: % of correct items selected

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

$$\text{Recall}(-) = \frac{TN}{TN + FP}$$

Evaluation Metrics

		<i>Truth</i>	
		Positive	Negative
<i>Predicted</i>	Positive	100	5
	Negative	45	100

		Positive	Negative
	Positive	50	25
	Negative	25	150

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

$$\frac{100}{100 + 5} = 0.95$$

$$\frac{50}{50 + 25} = 0.75$$

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

$$\frac{100}{100 + 45} = 0.69$$

$$\frac{50}{50 + 25} = 0.75$$

Two metrics - which one to use?

F-Score

- Combined measure
- Harmonic mean of Precision and Recall

$$F_1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

- Or more generally,

$$F_\beta = \frac{(1 + \beta^2) \cdot \text{Precision} \cdot \text{Recall}}{\beta^2 \cdot \text{Precision} + \text{Recall}}$$

Evaluation Metrics

		Truth	
		Positive	Negative
Predicted	Positive	100	5
	Negative	45	100

		Truth	
		Positive	Negative
Predicted	Positive	50	25
	Negative	25	150

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

$$\frac{100}{100 + 5} = 0.95$$

$$\frac{50}{50 + 25} = 0.75$$

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

$$\frac{100}{100 + 45} = 0.69$$

$$\frac{50}{50 + 25} = 0.75$$

$$F_1(+) = \frac{2 \cdot P(+)R(+)}{P(+) + R(+)}$$

$$0.8$$

$$0.75$$

Evaluation Metrics

Q: What happens to $F_1(+)$ if $FN = 25$ and $FP = 25$?

		Truth	
		Positive	Negative
Predicted	Positive	100	5
	Negative	45	100

	Positive	Negative
Positive	100	25
Negative	25	100

$$\text{Precision}(+) = \frac{TP}{TP + FP} = \frac{100}{100 + 5} = 0.95$$

$$\text{Recall}(+) = \frac{TP}{TP + FN} = \frac{100}{100 + 45} = 0.69$$

$$F_1(+) = \frac{2 \cdot P(+)R(+)}{P(+) + R(+)} = 0.8$$

Evaluation Metrics

		<i>Truth</i>	
		Positive	Negative
<i>Predicted</i>	Positive	TP	FP
	Negative	FN	TN

Use a simple rule, can you design a classifier with

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

Q. perfect precision?

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

Q. perfect recall?

Choosing Beta

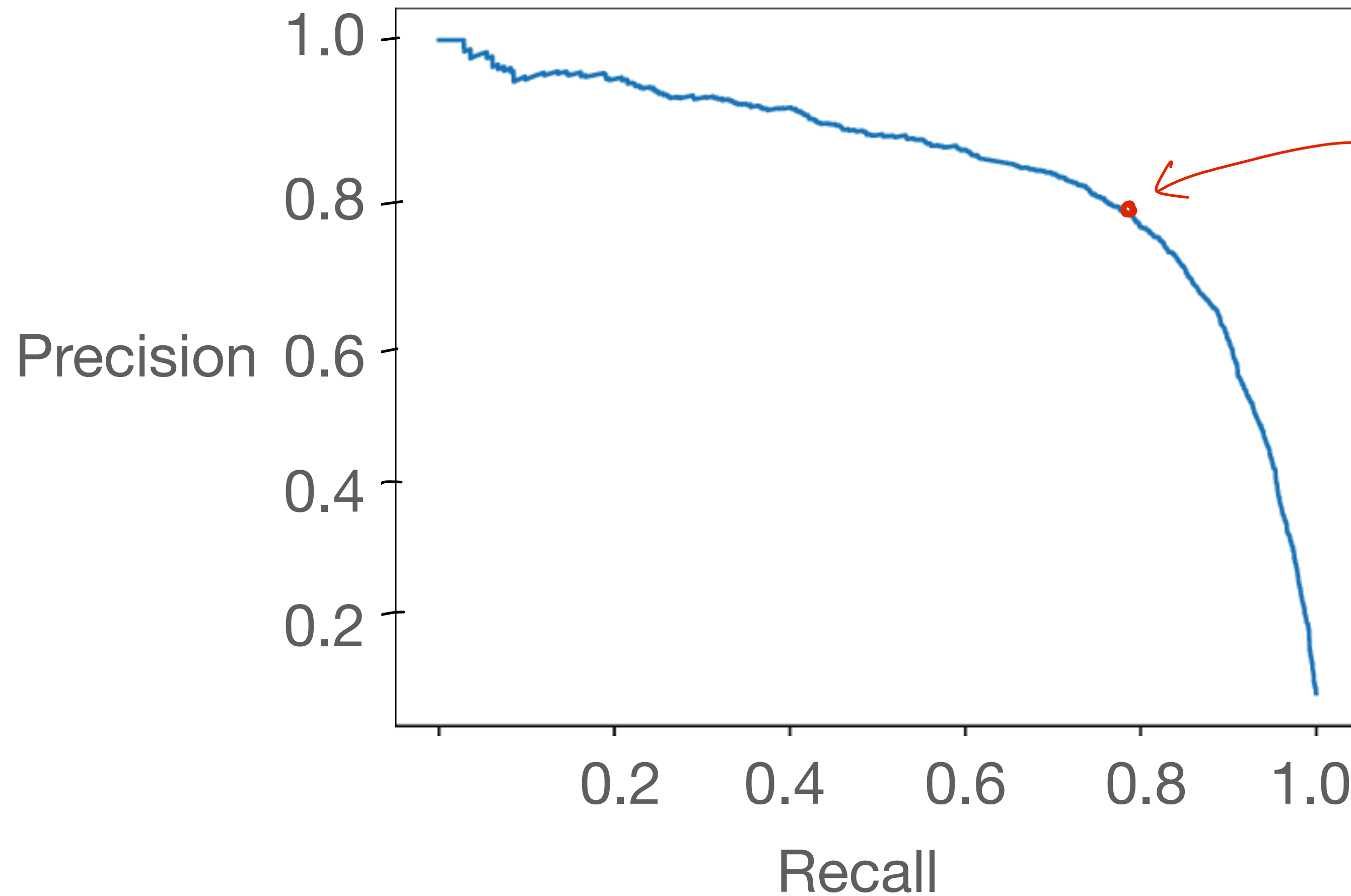
		<i>Truth</i>	
		Positive	Negative
<i>Predicted</i>	Positive	200	100
	Negative	50	100

$$F_{\beta} = \frac{(1 + \beta^2) \cdot \text{Precision} \cdot \text{Recall}}{\beta^2 \cdot \text{Precision} + \text{Recall}}$$

Q. Which value of Beta maximizes F_{β} for positive class?

- A. $\beta = 0.5$
- B. $\beta = 1$
- C. $\beta = 2$

Precision Recall tradeoff



Maximum F1

Vary

- Smoothing α
- Threshold T

$$\frac{P(+|d)}{P(-|d)} > T$$

Tune on validation set

Validation

Train

Validation

Test

- Choose a metric: Precision/Recall/F1
- Optimize for metric on **Validation** (aka Development) set
- Finally evaluate on ‘unseen’ **test** set
- Cross-validation:
 - Repeatedly sample several train-val splits
 - Reduces sampling bias due to sampling errors

Train Valid

Train Valid

⋮

Valid Train

Aggregating scores

- We have Precision, Recall, F1 for each class
- How to combine them for an overall score?
- **Macro-average**: Compute for each class, then average
- **Micro-average**: Collect predictions for all classes and jointly evaluate

Macro vs Micro average

Class 1

	Truth: yes	Truth: no
Classifier: yes	10	10
Classifier: no	10	970

Class 2

	Truth: yes	Truth: no
Classifier: yes	90	10
Classifier: no	10	890

Micro Ave. Table

	Truth: yes	Truth: no
Classifier: yes	100	20
Classifier: no	20	1860

Precision

- Macro-averaged precision: $(0.5 + 0.9) / 2 = 0.7$
- Micro-averaged precision: $100/120 = 0.85$
- Micro-averaged score is dominated by score on **common** classes

Summary

- Evaluation Metrics
 - **Accuracy** - coarse metric
 - **Precision**, **Recall**, **F1** for each class
- Aggregated scores
 - **Macro-average**: Compute for each class, then average
 - **Micro-average**: Collect predictions for all classes and jointly evaluate (dominated by common classes)
- Precision-Recall curve: pick threshold for maximum F1
 - Use **validation** set to tune hyperparameters, **test** set should remain “unseen”

