CMPT 413/713: Natural Language Processing

Classification - Evaluation

Spring 2024
2024-01-17

Adapted from slides from Danqi Chen, Karthik Narasimhan

Evaluation

Evaluation

- Consider binary classification
- Table of predictions

Confusion matrix

Truth

Predicted		Positive	Negative
	Positive	100	5
	Negative	45	100

- Ideally, we want:

Truth

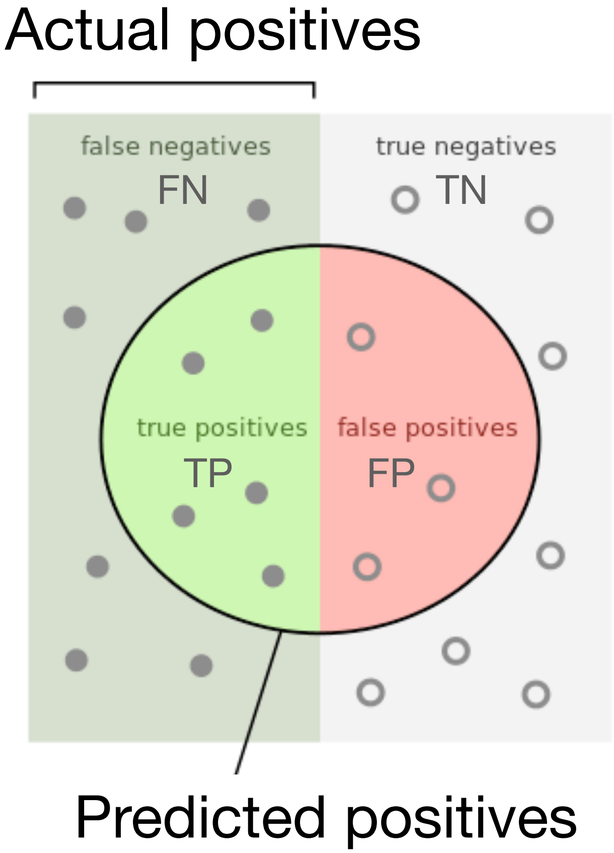
Predicted		Positive	Negative
	Positive	145	0
	Negative	0	105

Evaluation Metrics

Confusion matrix

Truth

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



(image credit: wikipedia)

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

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

Coarse metric

Evaluation Metrics

Truth

Predicted		Positive	Negative
	Positive	100	5
	Negative	45	100

	Positive	Negative
Positive	50	25
Negative	25	150

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

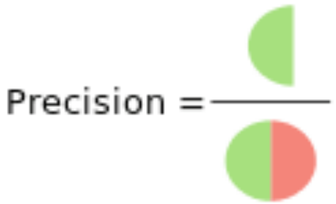
Accuracy cannot distinguish between the two models!

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

Precision and Recall

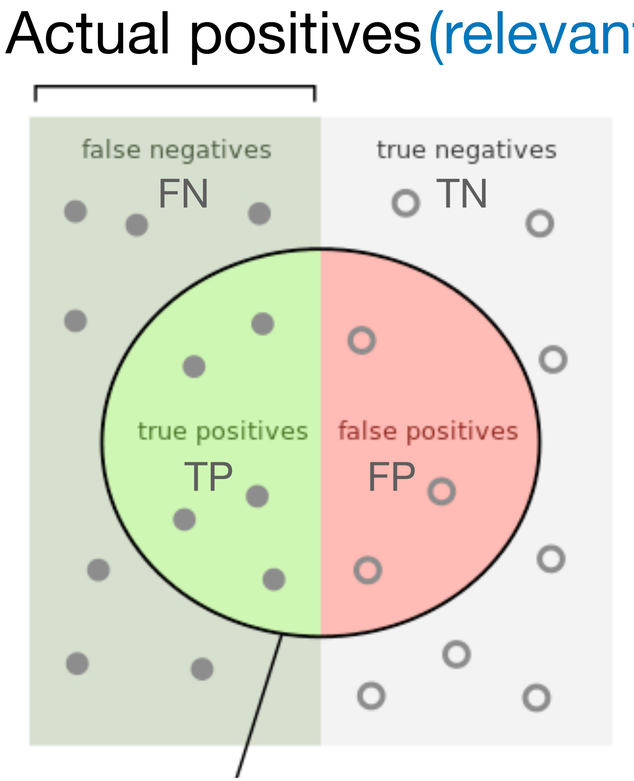
- Precision: % of selected classes that are correct

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



- Recall: % of correct items selected

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



(image credit: wikipedia)

Evaluation Metrics

Truth

Predicted	Positive	Negative
Positive	100	5
Negative	45	100

	Positive	Negative
Positive	50	25
Negative	25	150

$$\text{Precision}(+) = \frac{TP}{TP + FP} \qquad \frac{100}{100 + 5} = 0.95 \qquad \frac{50}{50 + 25} = 0.75$$

$$\text{Recall}(+) = \frac{TP}{TP + FN} \qquad \frac{100}{100 + 45} = 0.69 \qquad \frac{50}{50 + 25} = 0.75$$

Two metrics - which one to use when comparing the two models?

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,

Use β to control importance of Precision vs Recall

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

Evaluation Metrics

Truth

Predicted		Positive	Negative
	Positive	100	5
	Negative	45	100

	Positive	Negative
Positive	50	25
Negative	25	150

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

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

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

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

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

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

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

$$\boxed{0.8}$$

$$0.75$$

Evaluation Metrics

Truth

Predicted	Positive	Negative
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?

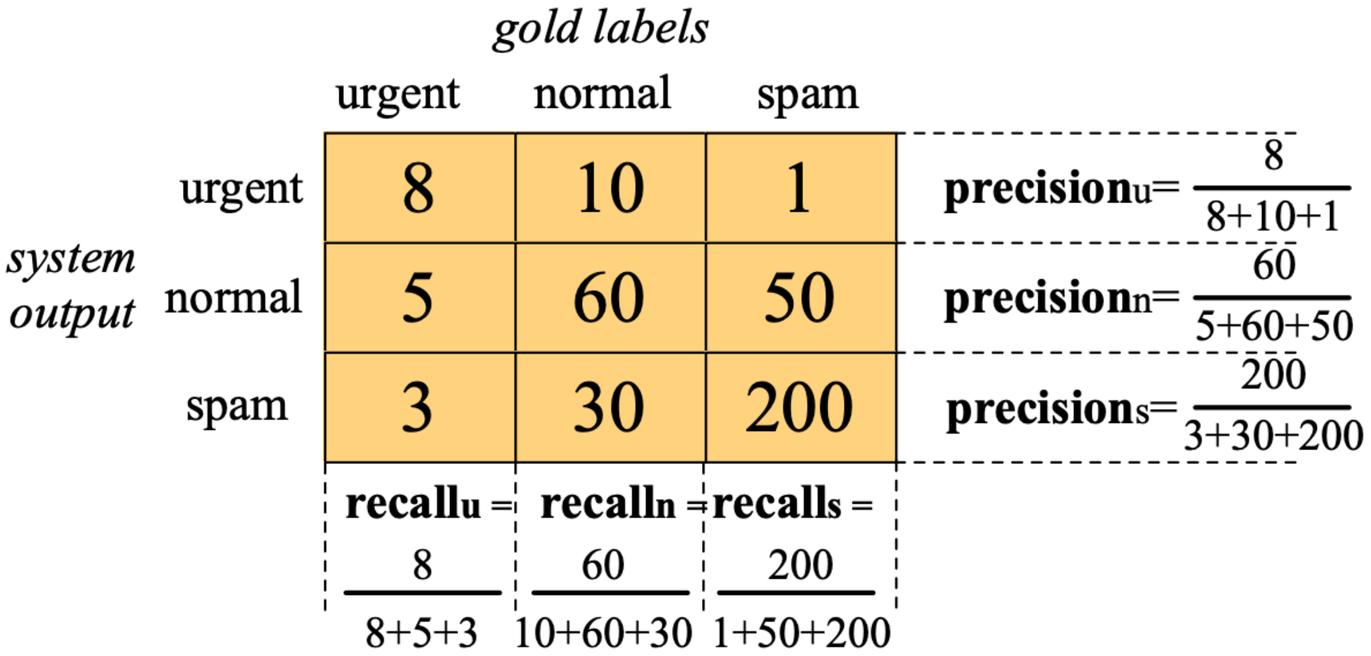
$$\text{Recall}(+) = \frac{TP}{TP + FN}$$

Q. perfect recall?

Aggregating scores

- How to handle more than 2 classes?

- We have Precision, Recall, F1 for each class



(Credits: Dan Jurafsky)

Aggregating scores

- How to handle more than 2 classes?
- 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

- Micro-averaged score is dominated by score on common classes

Class 1: Urgent

	true urgent	true not
system urgent	8	11
system not	8	340

$$\text{precision} = \frac{8}{8+11} = .42$$

Class 2: Normal

	true normal	true not
system normal	60	55
system not	40	212

$$\text{precision} = \frac{60}{60+55} = .52$$

Class 3: Spam

	true spam	true not
system spam	200	33
system not	51	83

$$\text{precision} = \frac{200}{200+33} = .86$$

$$\text{macroaverage precision} = \frac{.42+.52+.86}{3} = \mathbf{.60}$$

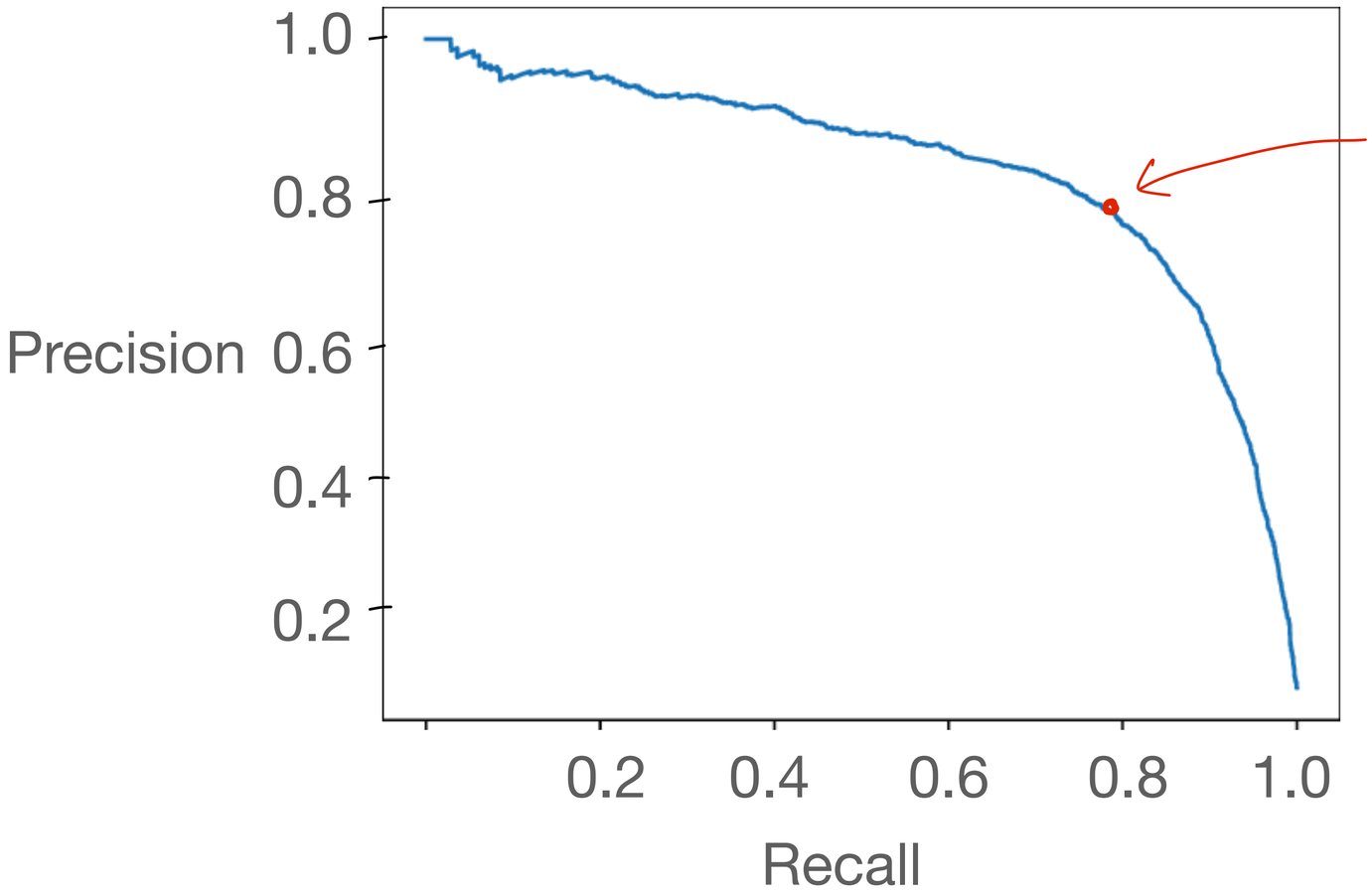
Pooled

	true yes	true no
system yes	268	99
system no	99	635

$$\text{microaverage precision} = \frac{268}{268+99} = \mathbf{.73}$$

(Credits: Dan Jurafsky)

Precision Recall tradeoff



Maximum F1

Vary hyperparameters

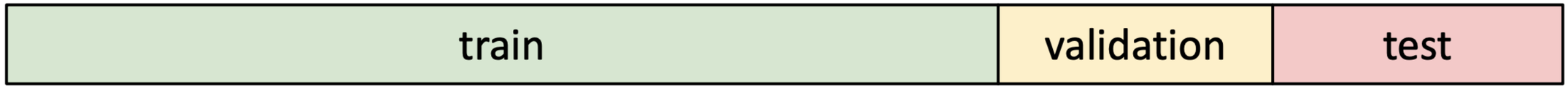
- Smoothing α
- Threshold T

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

Tune on validation set

Train, val, test split

- Train model on training set
- Tune hyperparameters on validation set
- Evaluate performance on unseen test set



Why do we do this?

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”