Evaluation

SLP 4.7

https://web.stanford.edu/~jurafsky/slp3/4.pdf

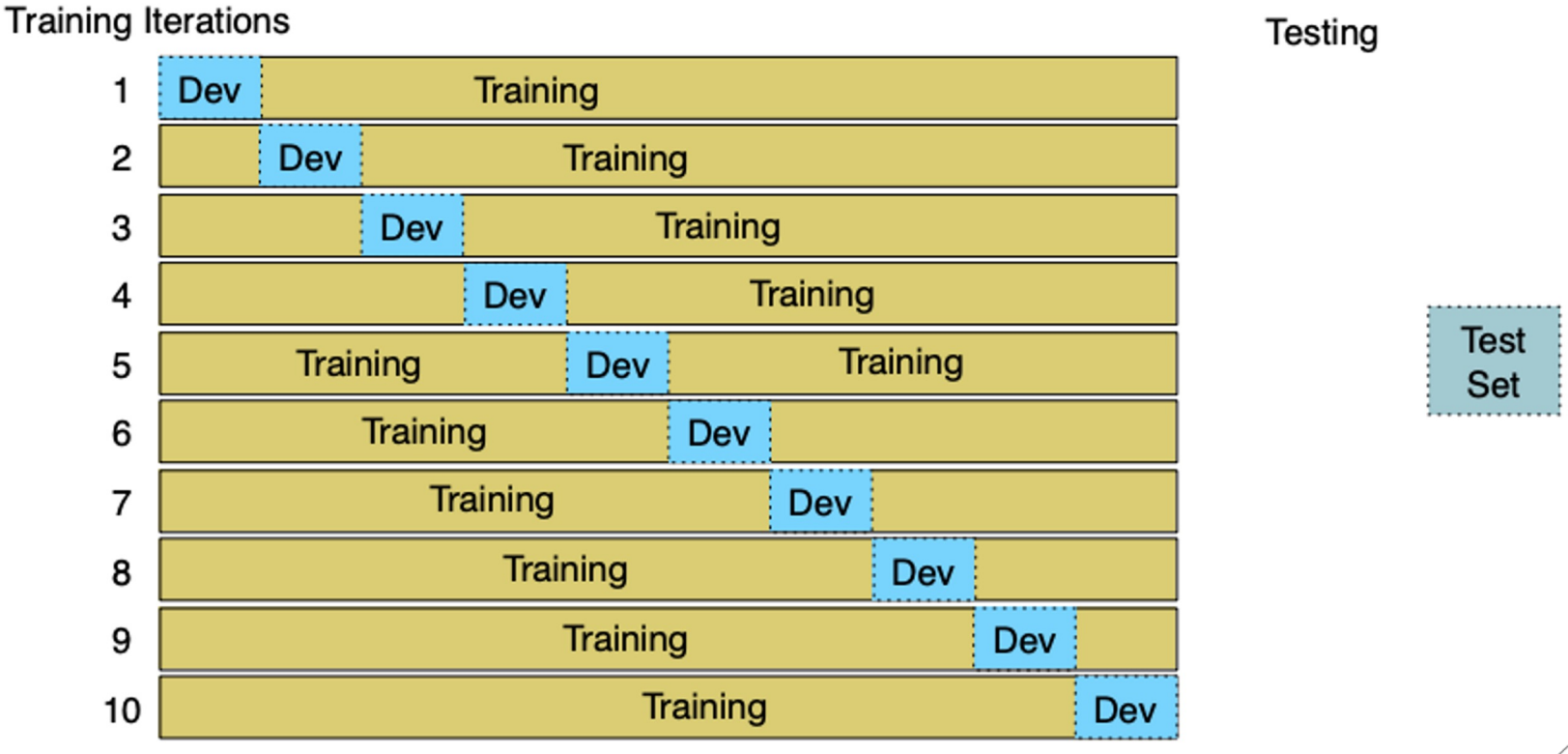
NLP 4.4

https://github.com/jacobeisenstein/gt-nlp-class/blob/master/notes/eisenstein-nlp-notes.pdf

Development Test Sets

- **Dev** test set for parameter tuning
- **Unseen test set**
  - avoid overfitting (‘tuning to the test set’)
  - more conservative estimate of performance
- **Cross-validation** over multiple splits
  - Handle sampling errors from different datasets

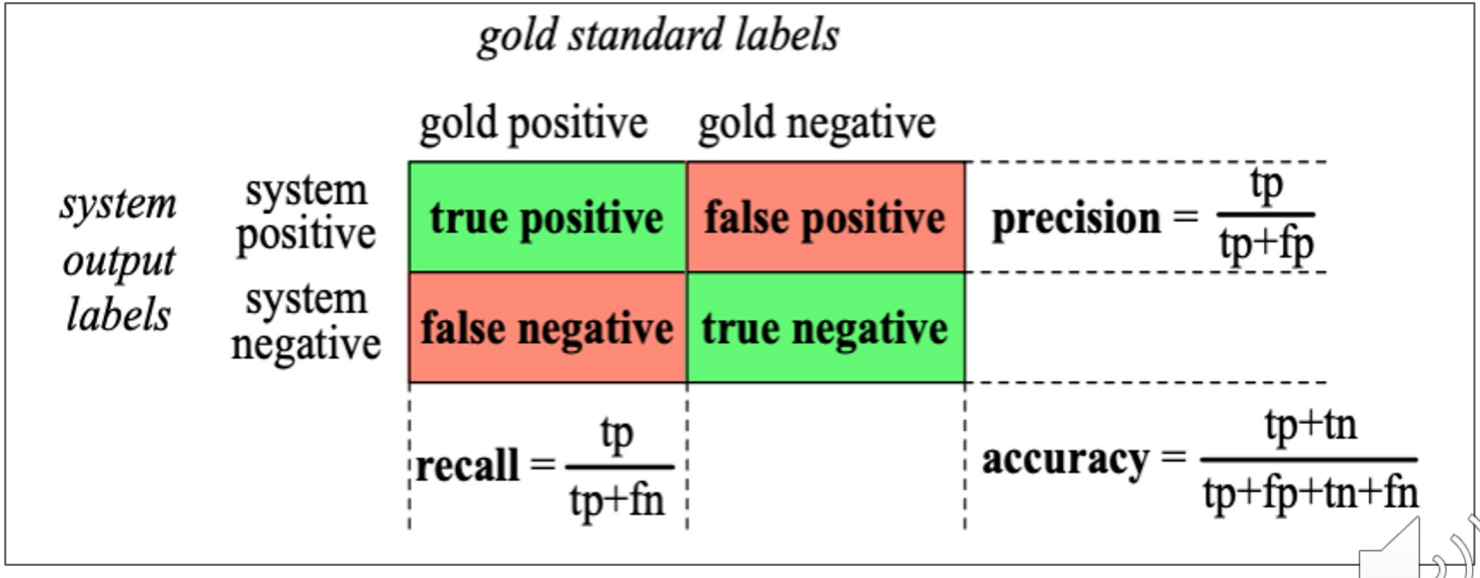
N-fold -cross validation



Metrics: Precision and Recall

- **Accuracy**
  - Class imbalance problem → balanced test set
- **Precision** and **recall** for each class

Contingency Table



Metrics: F-score

- A combined measure that assesses the P/R tradeoff is F measure (weighted harmonic mean):

$$F = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}$$

F_2 weights recall higher while $F_{0.5}$ weights precision higher

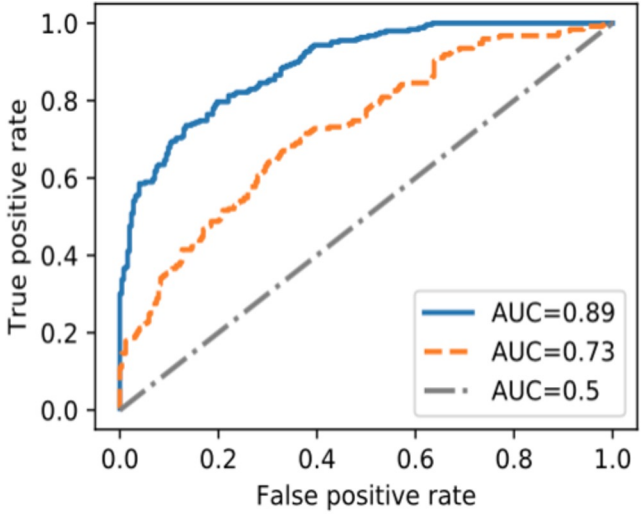
- Usually use

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

Metrics: AuC

- ROC (receiver operating characteristic) curve
- AuC(area under the curve): the probability that a randomly- selected positive example will be assigned a higher score by the classifier than a randomly-selected negative example

For AuC, refer to https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc

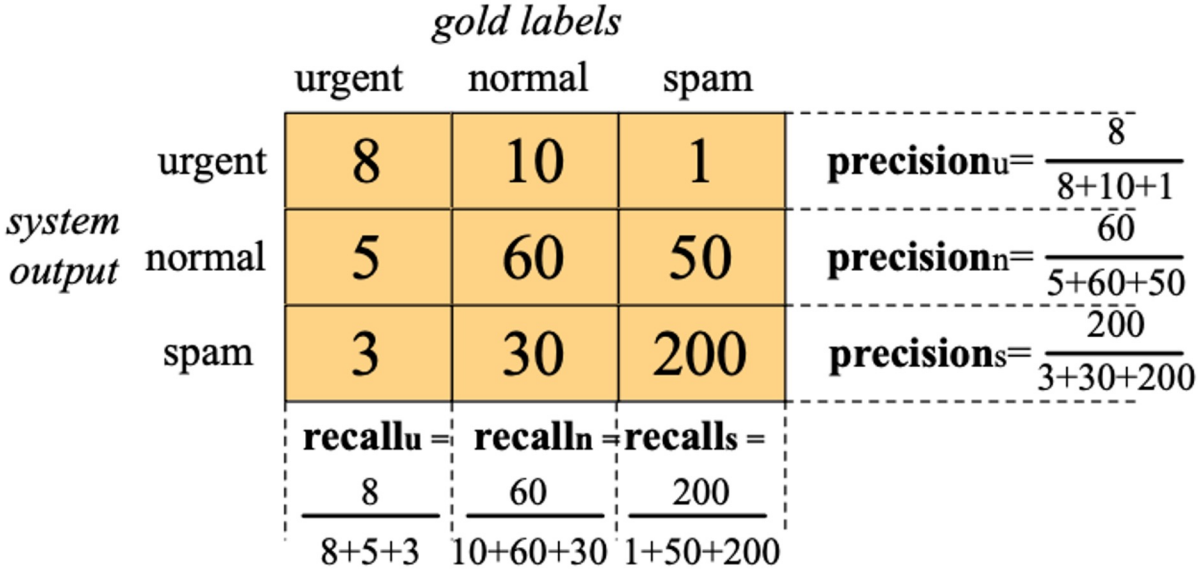


Metrics for multinomial classification

- **Macroaveraging** computes the performance for each class, and then average over classes
- **Microaveraging** collects the decisions for all classes into a single contingency table, and then compute precision and recall from that table
- **Microaveraged score is dominated by score on common classes**

Example: email categorization decision

Classify email as one of urgent, normal, spam



Micro vs. Macro averaging

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

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

Pooled

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

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

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

Confusion Matrix for Error analysis

- Classic Reuters-21578 Data Set: 21,578 docs (each 90 types, 200 tokens)
- 9603 training, 3299 test articles
- 118 categories
  - an article can be in more than one category
  - learn 118 binary category distinctions
- Average document has 1.24 classes
- Only about 10 out of 118 categories are large

Common categories
(#train, #test)

- Earn (2877, 1087)
- Acquisitions (1650, 179)
- Money-fx (538, 179)
- Grain (433, 149)
- Crude (389, 189)

- Trade (369,119)
- Interest (347, 131)
- Ship (197, 89)
- Wheat (212, 71)
- Corn (182, 56)

Reuters Text Categorization data

<REUTERS TOPICS="YES" LEWISSPLIT="TRAIN" CGISPLIT="TRAINING-SET" OLDID="12981" NEWID="798">
<DATE> 2-MAR-1987 16:51:43.42</DATE>
<TOPICS><D>livestock</D><D>hog</D></TOPICS>
<TITLE>AMERICAN PORK CONGRESS KICKS OFF TOMORROW</TITLE>
<DATELINE> CHICAGO, March 2 - </DATELINE><BODY>The American Pork Congress kicks off tomorrow, March 3, in Indianapolis with 160 of the nations pork producers from 44 member states determining industry positions on a number of issues, according to the National Pork Producers Council, NPPC.

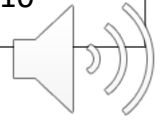
Delegates to the three day Congress will be considering 26 resolutions concerning various issues, including the future direction of farm policy and the tax law as it applies to the agriculture sector. The delegates will also debate whether to endorse concepts of a national PRV (pseudorabies virus) control and eradication program, the NPPC said.

A large trade show, in conjunction with the congress, will feature the latest in technology in all areas of the industry, the NPPC added. Reuter
&#3;</BODY></TEXT></REUTERS>

Confusion matrix

- For each pair of classes $<c_1,c_2>$ how many documents from c_1 were incorrectly assigned to c_2?
  - $c_{3,2}$: 90 wheat documents incorrectly assigned to poultry

Docs in test set	Assigned UK	Assigned poultry	Assigned wheat	Assigned coffee	Assigned interest	Assigned trade
True UK	95	1	13	0	1	0
True poultry	0	1	0	0	0	0
True wheat	10	**90**	0	1	0	0
True coffee	0	0	0	34	3	7
True interest	-	1	2	13	26	5
True trade	0	0	2	14	5	10



Summary

- Classifiers are trained using distinct training, dev, and test sets, including the use of cross-validation in the training set
- Classifiers are evaluated based on precision, recall, and F-score
- Macro vs. micro averaging for multinomial classification
- Confusion matrix to seek performance improvement