Classification: Big Picture

Computational Linguistics: Jordan Boyd-Graber
University of Maryland
SEPTEMBER 23, 2018

Slides adapted from Jason Eisner

Classification as Hammer

- Huge number of papers
- What are features?
- Where do data / labels come from?
- How do you evaluate?

 **An extremely versatile machine!**, November 22, 2006

By **Dr. Nickolas E. Jorgensen "njorgens3"**

This review is from: **Cuisinart DGB-600BC Grind & Brew, Brushed Chrome (Kitchen)**

This coffee-maker does so much! It makes weak, watery coffee! It grinds beans if you want it to! It inexplicably floods the entire counter with half-brewed coffee when you aren't looking! Perhaps it could be used to irrigate crops... It is time-consuming to clean, but in fairness I should also point out that the stainless-steel thermal carafe is a durable item that has withstood being hurled onto the floor in rage several times. And if all these features weren't enough, it's pretty expensive too. If faced with the choice between having a car door repeatedly slamming into my genitalia and buying this coffee-maker, I'd unhesitatingly choose the Cuisinart! The coffee would be lousy, but at least I could still have children...

Positive or Negative?

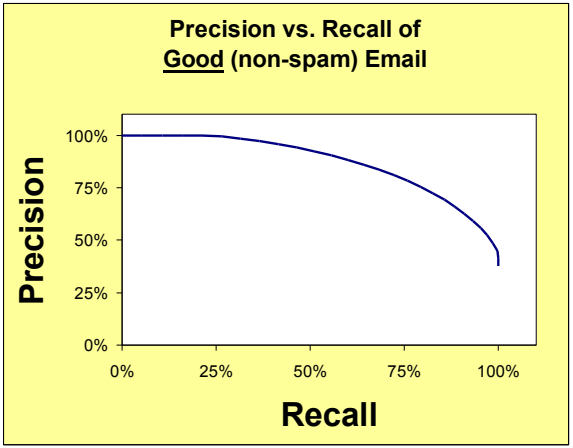
Other document classification tasks

- Is it spam? (see features)
- What medical billing code for this visit?
- What grade, as an answer to this essay question?
- Is it interesting to this user?
- News filtering; helpdesk routing
- Where should it be filed?

Measuring Classification

		Truth: Positive	Truth: Negative	
Test: Positive		True Positive	False Positive Type I α	Total Testing Positive
Test: Negative		False Negative Type II β	True Negative	Total Testing Negative
		Total Truly Positive	Total Truly Negative	Total

Measuring Classification



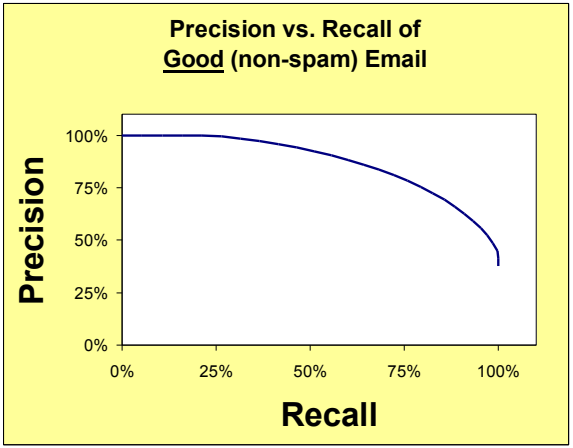
- Precision: Of what you returned, how much was right?

$$P = \frac{|\text{TP}|}{|\text{TP}| + |\text{FP}|} \tag{1}$$

- Recall: Of what could be right, how much did you find?

$$P = \frac{|\text{TP}|}{|\text{TP}| + |\text{FN}|} \tag{2}$$

Measuring Classification



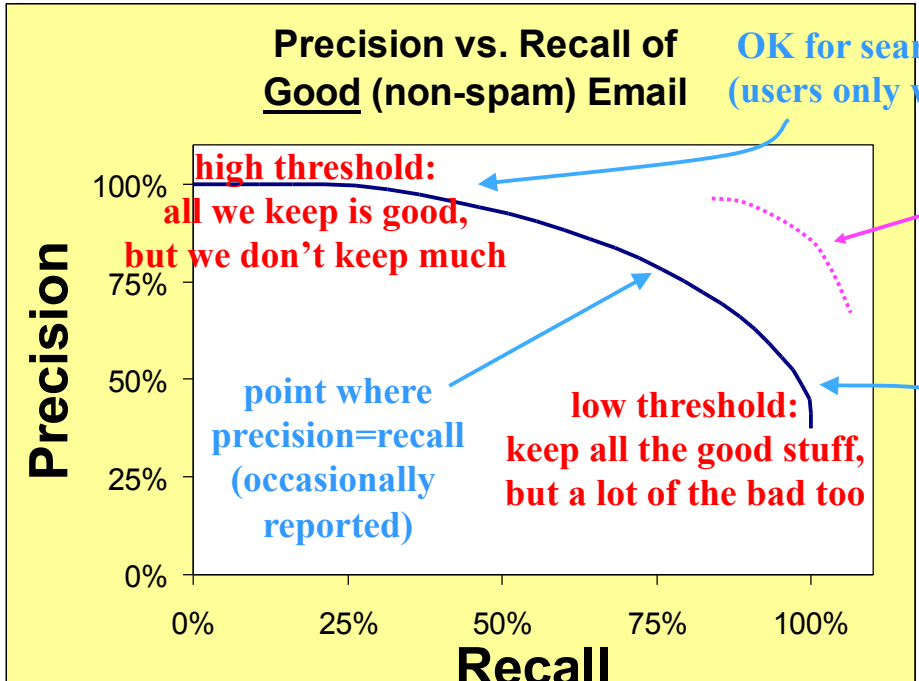
- Precision: Of what you returned, how much was right?

$$P = \frac{|\text{TP}|}{|\text{TP}| + |\text{FP}|} \quad (1)$$

- Recall: Of what could be right, how much did you find?

$$P = \frac{|\text{TP}|}{|\text{TP}| + |\text{FN}|} \quad (2)$$

- F-measure: geometric mean


Precision vs. Recall of Good (non-spam) Email
OK for search engines (users only want top 10)
would prefer to be here!
high threshold: all we keep is good, but we don't keep much
point where precision=recall (occasionally reported)
low threshold: keep all the good stuff, but a lot of the bad too
OK for spam filtering and legal search
Precision
Recall
100%
75%
50%
25%
0%
0%
25%
50%
75%
100%

Classifying Words

Training Data:

Sense	Context
(1) Manufacturing	... union responses to *plant* closures . ...
” ”	... computer disk drive *plant* located in ...
” ”	company manufacturing *plant* is in Orlando ...
(2) Living	... animal rather than *plant* tissues can be ...
” ”	... to strain microscopic *plant* life from the ...
” ”	and Golgi apparatus of *plant* and animal cells

Test Data:

Sense	Context
???	... vinyl chloride monomer *plant* , which is ...
???	... molecules found in *plant* tissue from the ...

Word sense disambiguation

Classifying Words

Problem:

Input: ... deja travaille cote a cote ...

⇓

Output: ... déjà travaillé côte à côte ...

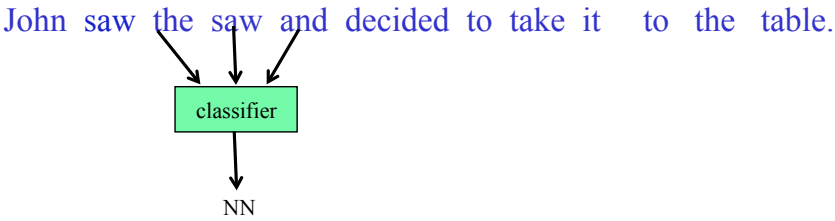
Examples:

... appeler l'autre **cote** de l'atlantique ...

⇒ *côté* (meaning side) or

⇒ *côte* (meaning coast)

Classifying Words



Part of speech tagging (more later!)

Where do labeled data come from?

- For supervised classification, we've assumed that our data are already available
- Not always the case
- This comes from **annotation** (e.g., all of the previous examples)

Why do we annotate?

We manually annotate texts for several reasons

- to understand the nature of text (e.g., what % of sentences in news articles are opinions?)
- to establish the level of human performance (e.g., how well can people assign POS tags?)
- to evaluate a computer model for some phenomenon (e.g., how often does my tagger or parser find the correct answer?)

The process of annotation

- Develop a set of annotations
- Define each of the annotations
- Have annotations annotate the **same** data
- See if they agree (more on this later)
  - If not, go back to Step 1
  - Why not?
    - Bad annotators?
    - Bad definitions?
    - Unexpected data?

Who does the annotation?

- Undergrads
- Grad students
- Crowdsourcing
  - Scammers
  - Diverse population
    - Worldwide
    - Bored office workers
    - Individuals at home
  - Equity issues
- Users
  - Reviews
  - Blog categories
  - Metadata
  - Often noisy

Why is it important to have agreement?

- Think about what happens to a classifier if it has inconsistent data (same data, different annotations)

Why is it important to have agreement?

- Think about what happens to a classifier if it has inconsistent data (same data, different annotations)
  - For an SVM: there's separating hyperplane
  - For a decision tree: decreases information gain of all the features
- Your classifier is only as good as the data it gets
- If your annotators only agree on 40% of the data, your accuracy will be less than 40%
- Common problem: disagreement is undetected because each item is only annotated once
- Resulting complaint: machine learning sucks

What does agreement mean?

- Simple answer: how often do two annotators give the same answer
- More complicated: above, **adjusting for chance agreement**
- Most important for class-imbalanced data

Computing Agreement

$$\kappa = \frac{P_a - P_c}{1 - P_c} \tag{3}$$

- P_a: Probability of coders agreeing
- P_c: Probability of coders agreeing by chance

Agreement example

	Annotator B		
Annotator A	Y	N	
Y	20	5	25
N	10	15	25
	30	20	

Agreement example

	Annotator B		
Annotator A	Y	N	
Y	20	5	25
N	10	15	25
	30	20	

Agreement example

	Annotator B		
Annotator A	Y	N	
Y	20	5	25
N	10	15	25
	30	20	

Probability of agreement

$P_a = \frac{15+20}{50} = 0.7$

Agreement example

	Annotator B		
Annotator A	Y	N	
Y	20	5	25
N	10	15	25
	30	20	

Probability of agreement

$P_a = \frac{15+20}{50} = 0.7$

Agreement example

	Annotator B		
Annotator A	Y	N	
Y	20	5	25
N	10	15	25
	30	20	

Chance agreement

- *A* says yes with probability .5
- *B* says yes with probability .6
- The probability that both of them say yes (assuming independence) is .3; the probability both say no is .2. The probability of chance agreement is then $P_c = 0.2 + 0.3$.

Agreement example

	Annotator B		
Annotator A	Y	N	
Y	20	5	25
N	10	15	25
	30	20	

Chance agreement

- *A* says yes with probability .5
- *B* says yes with probability .6
- The probability that both of them say yes (assuming independence) is .3; the probability both say no is .2. The probability of chance agreement is then $P_c = 0.2 + 0.3$.

Agreement example

	Annotator B		
Annotator A	Y	N	
Y	20	5	25
N	10	15	25
	30	20	

Chance agreement

- *A* says yes with probability .5
- *B* says yes with probability .6
- The probability that both of them say yes (assuming independence) is .3; the probability both say no is .2. The probability of chance agreement is then $P_c = 0.2 + 0.3$.

Agreement example

	Annotator B		
Annotator A	Y	N	
Y	20	5	25
N	10	15	25
	30	20	

Agreement:

$$\kappa = \frac{.7 - .5}{1 - .5} = .4 \tag{4}$$

Typically, you want above 0.7 agreement.

Recap

- We've talked about some classification algorithms (and you'll see others!)
- Important to keep in keep in mind why we're doing it
- How well they work
- How to set them up correctly: data often more important that algorithm
- Professionals argue about algorithm conditioned on data; better data always wins