- Next week: On Data!

  Then: Classification Case Studies & Project Discussion

- Today
  - Model Pros/Cons
  - Evaluation
  - Regularization + Overfitting
  - Annotations (see Exercise!)

9/23/21
UMass CS490A

Text Classif Models

$z_k = \vec{\beta}_k^T x$

Naive Bayes ⟷ Log. Reg

(ML) weights β learned from training data

Lexicon Counting

↑

β: made up by human choosing words
weights set manually

NB: $\beta_{k,w}$ learned indep. from other words

LR: $\beta_{k,w}$ learned jointly

$\Rightarrow$ both $p(y|x) = \sigma(\beta^T x)$

y: gold labels

"social"
"security"

NB: $p(\text{social}|k)\, p(\text{sec.}|k)$

Bad indep. assump

ML vs. Lexicons

⊕ Automatically detect human-like patterns

⊕ More "objective"? follow the data (good if data "unbiased")

⊖ Black box, not transparent

⊖ Need labeled data!!!

Lecture 7
Evaluation & Annotations

CS 490A, Fall 2021
https://people.cs.umass.edu/~brenocon/cs490a_f21/

Laure Thompson and Brendan O'Connor
College of Information and Computer Sciences
University of Massachusetts Amherst

[Many slides from Ari Kobren

- If you have labels, we know how to do:
  - Train a ML model
  - Evaluation metrics
  - Avoid overfitting

- But
  - Where do we get the labels ("annotations")?
  - Are these "gold standard" labels any good?

Evaluation

y_d "gold standard" label for doc d

$\hat{y}_d$ predicted labels

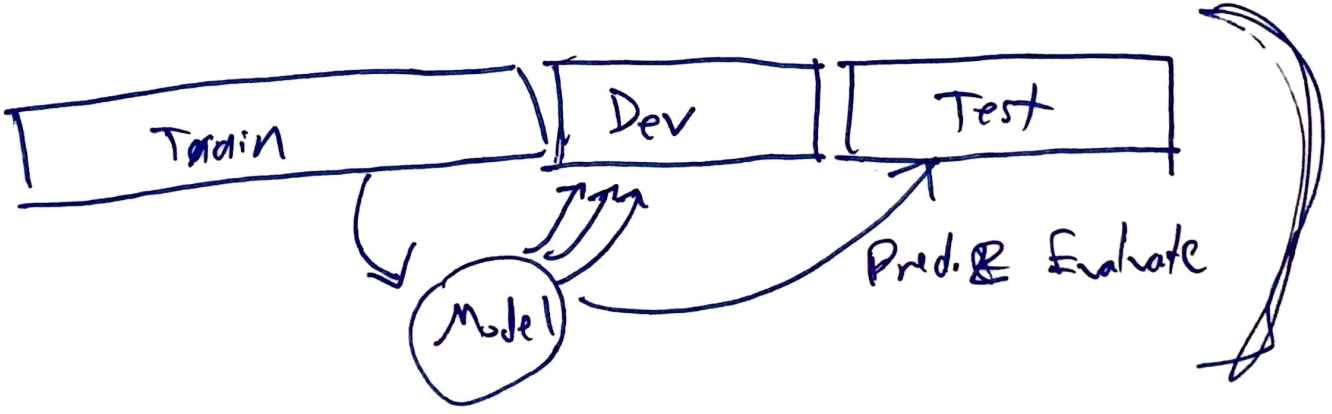
$$\text{Accuracy} = \frac{\text{Count}(\hat{y}_d = y_d)}{\text{Num. docs in test set}} = \frac{\sum_{d=1}^{N_{doc}} 1\{\hat{y}_d = y_d\}}{N_{doc}}$$

$\mathbb{I}(y_d = \hat{y}_d)$

Want generalization

on train data: overfitting

Split

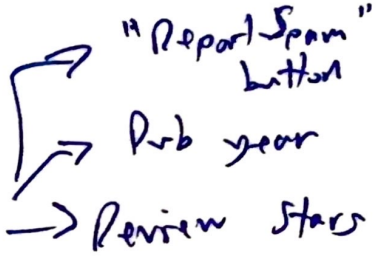


Regularization - α in NB "pseudocount" smoothing

L_2 regularization

Where to get labels?

- Natural annotations
  - Metadata - information associated with text document, but not in text itself
  - Clever patterns from text itself
- New human annotations
  - Yourself
  - Your friends
  - Hire people locally
  - Hire people online
    - Mechanical Turk — most commonly used crowdsourcing site
    - (For larger/more expensive tasks: Upwork/ODesk)



Annotation process

- To pilot a new task, requires an iterative process
  - Look at data to see what's possible
  - Conceptualize the task, try it yourself
  - Write annotation guidelines
  - Have annotators try to do it. Where do they disagree? What feedback do they have?
  - Revise guidelines and repeat
- If you don't do this, your labeled data will have lots of unclear, arbitrary, and implicit decisions inside of it

Name: _____

Interannotator Agreement for Sarcasm Detection
UMass CS 490A in-class exercise, 9/23/21

These are Reddit comments (from https://arxiv.org/pdf/1704.05579.pdf). Their authors tagged them as "sarcastic" or "not sarcastic". These tags are hidden. The task is to predict whether the message was tagged as sarcasm.

Step 1: On your own, label each 1=sarcastic, 0=not sarcastic, to the best of your ability.

1. but CNN told me the leaks are all faked by Russia!

2. Is this some sort of mug shot mash up?

3. there isn't... the website says specifically there is no way to change it once placed, you must cancel the order and place a new one to make any modifications.

4. All men are handsome!

5. Yeah but he paid his dues so it's his turn.

6. .....what?

7. Maybe you should stop reading it

8. It's okay, it's just his opinion!

9. I know that's what she's doing now, I'm saying this isn't the first time I've seen her.

10. honestly i'd have turned around and sold it to buy the K. But at the end of the day, if you aren't really one to overclock, then what difference does it make to you anyway?

Step 2: Once you are done, compare your answers to a neighbor who has also finished. (For makeup: use Piazza to find someone to share their answers with you.) Write down their answers above so you can easily do calculations as necessary. Calculate the following.

Agreement Rate: $\frac{\text{Count (items agree)}}{10}$

Cohen's Kappa:

Interannotator agreement

- How "real" is a task? Replicable? Reliability of annotations?
- How much do two humans *agree* on labels?
  - Difficulty of task. Human training? Human motivation/effort?
  - Goal: get the human performance upper bound
- If some classes predominate, raw agreement rate may be misleading
  - Chance-adjusted agreement: Cohen kappa for a pair of human annotators (see also Fleiss kappa, Krippendorff alpha…)

Cohen's kappa

p_o: observed agreement rate

p_e: agreement rate by chance

$$\frac{p_o - p_e}{1 - p_e}$$

- Reliability analysis: from the social sciences, especially psychology, content analysis, communications, etc.