Department of Computer Science
UNIVERSITY OF COLORADO **BOULDER**

Annotation and Feature Engineering

Introduction to Data Science Algorithms

Jordan Boyd-Graber and Michael Paul

HOUSES, SPOILERS, AND TRIVIA

Roadmap

- Getting good labels
- Feature engineering
    - Quiz Bowl Dataset
    - House Prices
    - TV Tropes Dataset
- How to split your dataset

Outline

1 Annotation: Getting Labels

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**

Examples of annotation

- Whether an e-mail is spam or not
- Whether a document is relevant to a court case (e-Discovery)
- Which meaning the noun “break” has
    - A time where you’re not working
    - A stroke of luck
    - A fracture or other discontinuity
    - A change in how things are done
- Whether an image has a van or not

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

Annotation Tools

- WordFreak (for text)
- LabelMe (for images)
- OpenAnnotation (an XML framework)
- Bamboo (visualization and annotation for humanists)