# Annotation and Evaluation

Digging into Data: Jordan Boyd-Graber

University of Maryland

April 15, 2013


COLLEGE OF
INFORMATION
STUDIES

# Exam

- Solutions posted
- Most missed question: “One coin in a collection of 9 has two heads. The rest are fair. If a coin, chosen at random from the lot and then tossed, turns up heads 3 times in a row, what is the probability that it is the two-headed coin?”

# Exam

- Solutions posted
- Most missed question: “One coin in a collection of 9 has two heads. The rest are fair. If a coin, chosen at random from the lot and then tossed, turns up heads 3 times in a row, what is the probability that it is the two-headed coin?”

**Solution**

Let $A$ be the event that the two-headed coin was selected. Let $B$ be the event that three heads were observed.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{1 \cdot 1/9}{1/9 + 8/9 \cdot (1/2)^3} = \frac{1/9}{2/9} = 1/2 \tag{1}$$

# Homework 2

- Some very imaginative
- Too open ended answers: more next time
- Need to be very careful about what data you collect features from and which dataset you report on!

# Free-response Questions

- SVM
  - Diagonal line does better than axis-aligned
  - Most who attempted it did fine
- Suicide
  - Not enough training data
  - Baseline: always say no
- Decision Tree
  - Wrong base
  - Math errors

# Rest of this course

- Less time with me lecturing
- Use the class time to connect with your project teammates to work on project
- But don't forget about HW3

# Outline

# 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 Þnd 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)

# Outline

# 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

$$\kappa = \frac{P_a - P_c}{1 - P_c} \tag{2}$$

- $P_a$: 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 | |

## Probability of agreement

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

## 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 | |

## Probability of agreement

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

## 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 | |

## Probability of agreement

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

## 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 | |

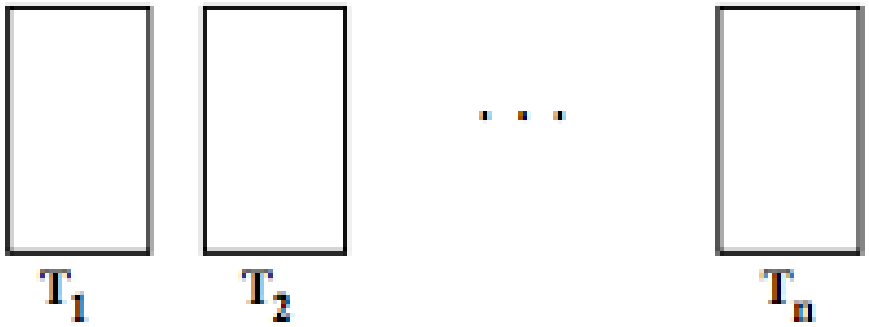
Agreement:

$$\kappa = \frac{.7 - .5}{1 - .5} = .4 \tag{3}$$

Typically, you want above 0.7 agreement.

# Outline

# Cross Validation




- Split your data into *N* folds
- For each fold:
  - Train your data on all the **other** folds
  - Compute accuracy/precision/recall on this fold
- Average over all folds
- Uses all available data

# Intrinsic vs. Extrinsic Evaluation

- We've focused on **intrinsic** evaluation
  - Correctly predicting spoilers
  - Assigning words/documents to correct category
  - Detecting whether an image has a cow in it
- More realistic: **extrinsic** evaluation
  - Number of spoilers seen by social media user
  - Number of relevant documents returned by IR system
  - Throughput of automatic cow milking system
- Bottom line: extrinsic evaluations are harder, but they're more often the thing you care about.

# Convincing Results

- Give baseline performance
  - Most frequent class
  - Random guessing
  - Current “best practice”
- Give qualitative results
  - Examples that were right / wrong
  - Error analysis
  - Tell a story
- Give “blue sky” bounds
  - Oracle results for pipeline systems
  - Human ability

# Takeaways

- Effective data science depends not just on algorithms, but also having good data
- Agreement measures the quality of the inputs into algorithms
- Effective evaluation is important for communicating your results