Feature Engineering

Digging into Data: Jordan Boyd-Graber

University of Maryland

April 7, 2014

COLLEGE OF
INFORMATION
STUDIES

Roadmap

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

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

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

- 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:

$$\kappa = \frac{.7 - .5}{1 - .5} = .4 \tag{2}$$

Typically, you want above 0.7 agreement.

Outline

Humans doing Incremental Classification

- Game called “quiz bowl”
- Two teams play each other
  - Moderator reads a question
  - When a team knows the answer, they signal (“buzz” in)
  - If right, they get points; otherwise, rest of the question is read to the other team
- Hundreds of teams in the US alone

Humans doing Incremental Classification

- Game called “quiz bowl”
- Two teams play each other
  - Moderator reads a question
  - When a team knows the answer, they signal (“buzz” in)
  - If right, they get points; otherwise, rest of the question is read to the other team
- Hundreds of teams in the US alone
- Example . . .

Sample Question 1

With Leo Szilard, he invented a doubly-eponymous

Sample Question 1

With Leo Szilard, he invented a doubly-eponymous refrigerator with no moving parts. He did not take interaction with neighbors into account when formulating his theory of

Sample Question 1

With Leo Szilard, he invented a doubly-eponymous refrigerator with no moving parts. He did not take interaction with neighbors into account when formulating his theory of heat capacity, so

Sample Question 1

With Leo Szilard, he invented a doubly-eponymous refrigerator with no moving parts. He did not take interaction with neighbors into account when formulating his theory of heat capacity, so Debye adjusted the theory for low temperatures. His summation convention automatically sums repeated indices in tensor products. His name is attached to the A and B coefficients

Sample Question 1

With Leo Szilard, he invented a doubly-eponymous refrigerator with no moving parts. He did not take interaction with neighbors into account when formulating his theory of heat capacity, so Debye adjusted the theory for low temperatures. His summation convention automatically sums repeated indices in tensor products. His name is attached to the A and B coefficients for spontaneous and stimulated emission, the subject of one of his multiple groundbreaking 1905 papers. He further developed the model of statistics sent to him by

Sample Question 1

With Leo Szilard, he invented a doubly-eponymous refrigerator with no moving parts. He did not take interaction with neighbors into account when formulating his theory of heat capacity, so Debye adjusted the theory for low temperatures. His summation convention automatically sums repeated indices in tensor products. His name is attached to the A and B coefficients for spontaneous and stimulated emission, the subject of one of his multiple groundbreaking 1905 papers. He further developed the model of statistics sent to him by Bose to describe particles with integer spin. For 10 points, who is this German physicist best known for formulating the

Sample Question 1

With Leo Szilard, he invented a doubly-eponymous refrigerator with no moving parts. He did not take interaction with neighbors into account when formulating his theory of heat capacity, so Debye adjusted the theory for low temperatures. His summation convention automatically sums repeated indices in tensor products. His name is attached to the A and B coefficients for spontaneous and stimulated emission, the subject of one of his multiple groundbreaking 1905 papers. He further developed the model of statistics sent to him by Bose to describe particles with integer spin. For 10 points, who is this German physicist best known for formulating the special and general theories of relativity?

Sample Question 1

With Leo Szilard, he invented a doubly-eponymous refrigerator with no moving parts. He did not take interaction with neighbors into account when formulating his theory of heat capacity, so Debye adjusted the theory for low temperatures. His summation convention automatically sums repeated indices in tensor products. His name is attached to the A and B coefficients for spontaneous and stimulated emission, the subject of one of his multiple groundbreaking 1905 papers. He further developed the model of statistics sent to him by Bose to describe particles with integer spin. For 10 points, who is this German physicist best known for formulating the special and general theories of relativity?

Albert Einstein

Humans doing Incremental Classification

- This is **not** Jeopardy
- There are buzzers, but players can only buzz at the end of a question
- Doesn't discriminate knowledge
- Quiz bowl questions are pyramidal

Research Question: How do we know if a guess is correct?

- Turn (question, guess) into features
- Treat it as a binary classification problem
- What features help us do this well?

Research Question: How do we know if a guess is correct?

- Turn (question, guess) into features
- Treat it as a binary classification problem
- What features help us do this well?
- Subject of HW3

Provided Dataset

- **text**: the clues revealed so far
- **page**: a guess at the answer
- **answer**: the actual answer (closest Wikipedia page)
- **body_score**: IR measure of how good a match the text is

Outline

Baseline

- What if we always say that the answer is wrong?
- Performance: 0.54
- Every feature should do better than this (otherwise, it's useless)

Page Name

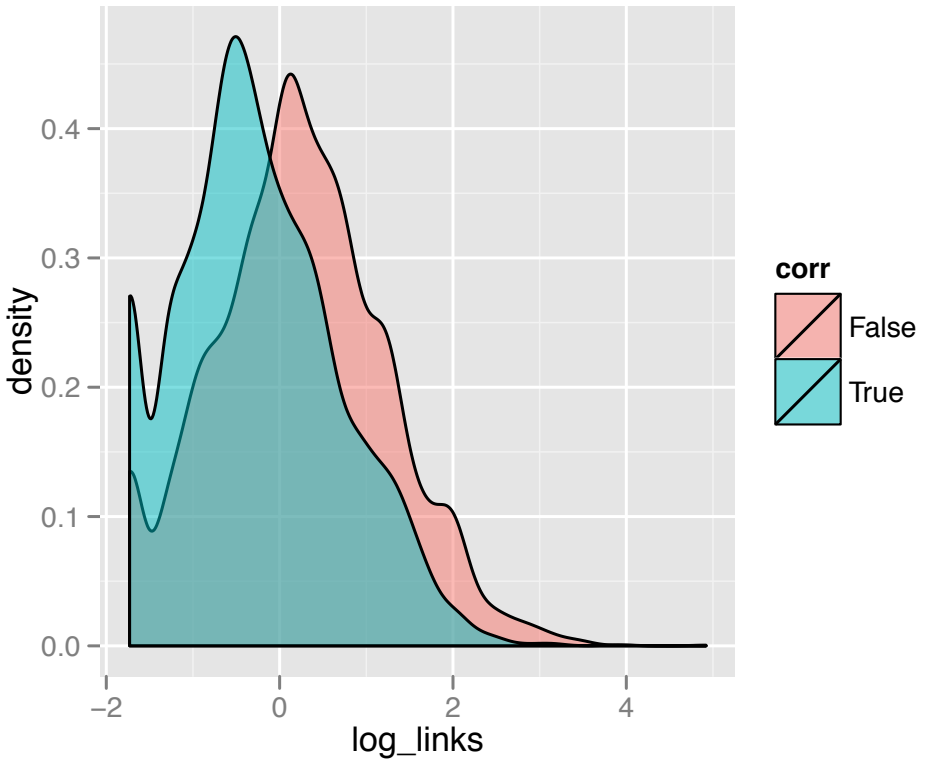
- The title of wikipedia pages often have disambiguation in parentheses
  1. Paris (mythology)
  2. Paris (song)
  3. Paris (genus)
  4. Paris (band)

Page Name

- The title of wikipedia pages often have disambiguation in parentheses
  1. Paris (mythology)
  2. Paris (song)
  3. Paris (genus)
  4. Paris (band)
- Feature is 1 if the page has disambiguator in the text
  - "This band performed . . . ", Paris (band) → `True`
  - "This band performed . . . ", Paris (mythology) → `False`
- Slight improvement: 0.58

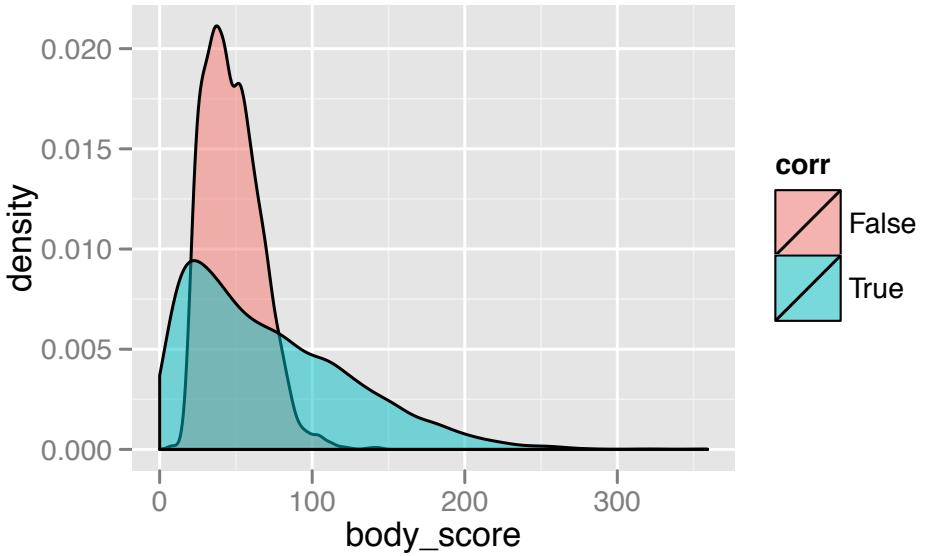
Links

- The more more links a Wikipedia page has, the more popular it is
- Popularity is often a sign of a **wrong answer**
- By itself, doesn't do so well: 0.56
- But improves if we take the log of the value: 0.61



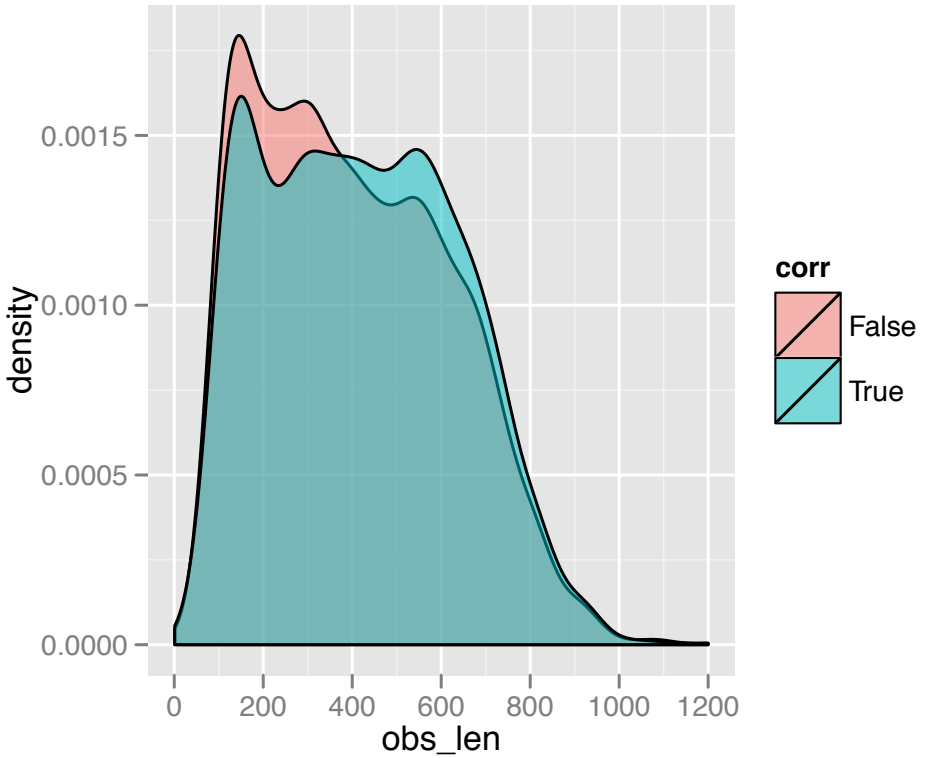
Score

- We can see how similar the text of a Wikipedia page is
- Higher, the better
- This feature alone gives accuracy of 0.75



Length

- The more text we see, the more confident we should be
- By itself, doesn't do so well: 0.56
- But when combined with the IR score, does great: 0.82 (best so far)



Others . . .

- Tournament the question was used in
- The type of thing the answer is
- Try your own, be creative!
- Last year's feature engineering assignment

Outline

TV Tropes

- Social media site
- Catalog of “tropes”
- Functionally like Wikipedia, but . . .
  - Less formal
  - No notability requirement
  - Focused on popular culture

Absent-Minded Professor

- “Doc” Emmett Brown from *Back to the Future*.
- The drunk mathematician in *Strangers on a Train* becomes a plot point, because of his forgetfulness, Guy is suspected of a murder he didn’t commit.
- *The Muppet Show*: Dr. Bunsen Honeydew.

Spoilers

- What makes neat is that the dataset is annotated by users for **spoilers**.
- A spoiler: “A published piece of information that divulges a surprise, such as a plot twist in a movie.”

Spoiler

- Han Solo arriving just in time to save Luke from Vader and buy Luke the vital seconds needed to send the proton torpedos into the Death Star’s thermal exhaust port.
- Leia, after finding out that despite her (feigned) cooperation, Tarkin intends to destroy Alderaan anyway.
- Luke rushes to the farm, only to find it already raided and his relatives dead harkens to an equally distressing scene in The Searchers.

Not a spoiler

- Diving into the garbage chute gets them out of the firefight, but the droids have to save them from the compacter.
- They do some pretty evil things with that Death Star, but we never hear much of how they affect the rest of the Galaxy. A deleted scene between Luke and Biggs explores this somewhat.
- Luke enters Leia’s cell in a Stormtrooper uniform, and she calmly starts some banter.

The dataset

- Downloaded the pages associated with a **show**. Took complete sentences from the text and split them into ones with spoilers and those without
- Created a balanced dataset (50% spoilers, 50% not)
- Split into training, development, and test **shows**

The dataset

- Downloaded the pages associated with a **show**. Took complete sentences from the text and split them into ones with spoilers and those without
- Created a balanced dataset (50% spoilers, 50% not)
- Split into training, development, and test **shows**
  - Why is this important?

The dataset

- Downloaded the pages associated with a **show**. Took complete sentences from the text and split them into ones with spoilers and those without
- Created a balanced dataset (50% spoilers, 50% not)
- Split into training, development, and test **shows**
  - Why is this important?
- I'll show results using SVM; similar results apply to other classifiers

Outline

Step 1: The obvious

- Take every sentence, and split on on-characters.
- Input: “These aren’t the droids you’re looking for.”

Step 1: The obvious

- Take every sentence, and split on on-characters.
- Input: “These aren’t the droids you’re looking for.”

Features

These:1 aren:1 t:1 the:1 droids:1 you:1 re:1 looking:1 for:1

	False	True
False	56	34
True	583	605

Accuracy: 0.517

Step 1: The obvious

- Take every sentence, and split on on-characters.
- Input: “These aren’t the droids you’re looking for.”

Features

These:1 aren:1 t:1 the:1 droids:1 you:1 re:1 looking:1 for:1

What’s wrong with this?

	False	True
False	56	34
True	583	605

Accuracy: 0.517

Step 2: Normalization

- Normalize the words
  - Lowercase everything
  - Stem the words (not always a good idea!)
- Input: “These aren’t the droids you’re looking for.”

Step 2: Normalization

- Normalize the words
  - Lowercase everything
  - Stem the words (not always a good idea!)
- Input: “These aren’t the droids you’re looking for.”

Features

these:1 are:1 t:1 the:1 droid:1
you:1 re:1 look:1 for:1

	False	True
False	52	27
True	587	612

Accuracy: 0.520

Step 3: Remove Usless Features

- Use a “stoplist”
- Remove features that appear in > 10% of observations (and aren’t correlated with label)
- Input: “These aren’t the droids you’re looking for.”

Step 3: Remove Usless Features

- Use a “stoplist”
- Remove features that appear in > 10% of observations (and aren’t correlated with label)
- Input: “These aren’t the droids you’re looking for.”

Features

droid:1 look:1

	False	True
False	59	20
True	578	621

Accuracy: 0.532

Step 4: Add Useful Features

- Use bigrams (“these_are”) instead of unigrams (“these”, “are”)
- Creates a lot of features!
- Input: “These aren’t the droids you’re looking for.”

Step 4: Add Useful Features

- Use bigrams (“these_are”) instead of unigrams (“these”, “are”)
- Creates a lot of features!
- Input: “These aren’t the droids you’re looking for.”

Features

these_are:1 aren_t:1 t_the:1 the_droids:1 you_re:1 re_looking:1 looking_for:1

	False	True
False	203	104
True	436	535

Accuracy: 0.578

Step 5: Prune (Again)

- Not all bigrams appear often
- SVM has to search a long time and might not get to the right answer
- Helps to prune features
- Input: “These aren’t the droids you’re looking for.”

Step 5: Prune (Again)

- Not all bigrams appear often
- SVM has to search a long time and might not get to the right answer
- Helps to prune features
- Input: “These aren’t the droids you’re looking for.”

Features

these_are:1 the_droids:1
re_looking:1 looking_for:1

	False	True
False	410	276
True	229	363

Accuracy: 0.605

How do you find new features?

- Make predictions on the development set.
- Look at contingency table; where are the errors?
- What do you miss?

How do you find new features?

- Make predictions on the development set.
- Look at contingency table; where are the errors?
- What do you miss? **Error analysis!**
- What feature would the classifier need to get this right?
- What features are confusing the classifier?
  - If it never appears in the development set, it isn't useful
  - If it doesn't appear often, it isn't useful

How do you know something is a good feature?

- Make a contingency table for that feature (should give you good information gain)
- Throw it into your classifier (accuracy should improve)

Homework 3

- I've given you quiz bowl questions
- And test data (no labels)
- Only have small number of features (should get you around 81%)
  - For these features, it doesn't matter (much) which classifier you use
- Your job: add additional features and see how they do
- Be creative! Find new and interesting data, extract useful things from these data.

Homework 3

- I've given you quiz bowl questions
- And test data (no labels)
- Only have small number of features (should get you around 81%)
  - For these features, it doesn't matter (much) which classifier you use
- Your job: add additional features and see how they do
- Be creative! Find new and interesting data, extract useful things from these data.
- Last year: best students wrote a paper with me:
  Jordan Boyd-Graber, Kimberly Glasgow, and Jackie Sauter Zajac. **Spoiler Alert: Machine Learning Approaches to Detect Social Media Posts with Revelatory Information**. *ASIST 2013: The 76th Annual Meeting of the American Society for Information Science and Technology*, 2013.

Outline

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

Outline

Lifecycle of Project

- Starting with no labels
- Building classification scheme
- Feature engineering
- Evaluation

RTextTools

```
library(RTextTools)

train.df <- read.csv("train/train.csv")
train.df$sentence <- as.character(train.df$sentence)

dev.df <- read.csv("dev/dev.csv")
dev.df$sentence <- as.character(dev.df$sentence)

train.df <- train.df[1:1000,]
dev.df <- dev.df[1:100,]

data <- rbind(train.df, dev.df)
dev_size <- dim(dev.df)[1]
total_size <- dim(data)[1]

matrix <- create_matrix(cbind(data$sentence, data$trope),
                language="english", removeNumbers=TRUE, stemWords=FALSE,
                weighting=weightTfIdf)

container <- create_container(matrix, data$spoiler, trainSize=1:dev_size,
                      testSize=(1+dev_size):total_size, virgin=FALSE)

models <- train_models(container, algorithms=c("MAXENT","SVM"))
results <- classify_models(container, models)
```