Text Data Mining: Predictive and Exploratory Analysis of Text

> Jaime Arguello jarguell@email.unc.edu

January 10, 2018

Outline

Introductions

What is Text Data Mining?

Predictive Analysis of Text: The Big Picture

Exploratory Analysis of Text: The Big Picture

Applications

Introductions

- Hello, my name is _____.
- I'm in the _____ program.

•

• I'm taking this course because I'd like to learn how to



What is Text Data Mining?

 The science and practice of <u>building</u> and <u>evaluating</u> computer programs that automatically <u>detect</u> or <u>discover</u> <u>interesting</u> and <u>useful</u> things in collections of <u>natural</u> <u>language text</u>

Related Fields

- Machine Learning: developing computer programs that improve their performance with "experience"
- Data Mining: developing methods that discover patterns within large structured datasets
- Statistics: developing methods for the interpretation of data and experimental outcomes in reaching conclusions with a certain degree of confidence

Text Data Mining in this Course

- Predictive Analysis of Text
 - developing computer programs that automatically recognize or detect a particular concept within a span of text
- Exploratory Analysis of Text:
 - developing computer programs that automatically <u>discover</u> interesting and useful patterns or trends in text collections

Outline

Introductions

What is Text Data Mining?

Predictive Analysis of Text: The Big Picture

Exploratory Analysis of Text: The Big Picture

Applications



- We could imagine writing a "triangle detector" by hand:
 - if shape has three sides, then shape = triangle.
 - otherwise, shape = other
- Alternatively, we could use supervised machine learning!

training



labeled examples



predictions

new, unlabeled examples

training



labeled exampl

What is the part that is missing?

HINT: It's what most of this class will be about!

model

new, unlabeled examples

predictions

Predictive Analysis representation: features

color	size	# slides	equal sides	••••	label
red	big	3	no		yes
green	big	3	yes		yes
blue	small	inf	yes		no
blue	small	4	yes		no
i				•••••	
red	big	3	yes		yes



training

color	size	sides	equal sides		label
red	big	3	no		yes
green	big	3	yes		yes
blue	small	inf	yes		no
blue	small	4	yes		no
:	:	:		:	:
red	big	3	yes		yes

labeled examples

machine
learning
algorithm



color	size	sides	equal sides		label
red	big	3	no		???
green	big	3	yes		???
blue	small	inf	yes		???
blue	small	4	yes		???
	:	:			???
red	big	3	yes		???
n	new, unlabeled				



color	size	sides	equal sides	 label
red	big	3	no	 yes
green	big	3	yes	 yes
blue	small	inf	yes	 no
blue	small	4	yes	 no
:	:	:		
red	big	3	yes	 yes

predictions

examples

Predictive Analysis basic ingredients

- 1. Training data: a set of examples of the concept we want to automatically recognize
- 2. Representation: a set of features that we believe are useful in recognizing the desired concept
- 3. Learning algorithm: a computer program that uses the training data to learn a predictive model of the concept



Predictive Analysis

basic ingredients

Highly influential!

- 1. Training the concept we want to automatically recognize
- 2. Representation: a set of features that we believe are useful in recognizing the desired concept
- 3. Learning algorithm: a computer program that uses the training data to learn a predictive model of the concept

Predictive Analysis basic ingredients

- 4. Model: a (mathematical) function that describes a predictive relationship between the feature values and the presence/absence of the concept
- 5. Test data: a set of previously unseen examples used to estimate the model's effectiveness
- 6. Performance metrics: a set of statistics used measure the predictive effectiveness of the model

Predictive Analysis basic ingredients: the focus in this course

- 1. Training data: a set of examples of the concept we want to automatically recognize
- 2. Representation: a set of features that we believe are useful in recognizing the desired concept
 - 3. Learning algorithm: uses the training data to learn a predictive model of the "concept"



Predictive Analysis basic ingredients: the focus in this course

- 4. Model: describes a predictive relationship between feature values and the presence/ absence of the concept
- 5. Test data: a set of previously unseen examples used to estimate the model's effectiveness
- - 6. Performance metrics: a set of statistics used measure the predictive effectiveness of the model



Predictive Analysis applications

- Topic categorization
- Opinion mining
- Sentiment analysis
- Bias or viewpoint detection
- Discourse analysis (e.g., student retention)
- Forecasting and nowcasting
- Any other ideas?

What Could Possibly Go Wrong?

- 1. Bad feature representation
- 2. Bad data + misleading correlations
- 3. Noisy labels for training and testing
- 4. Bad learning algorithm
- 5. Misleading evaluation metric



color	size	90 deg. angle	equal sides	••••	label
red	big	yes	no		yes
green	big	no	yes		yes
blue	small	no	yes		no
blue	small	yes	yes		no
:	••••			••••	
red	big	no	yes		yes



color	size	90 deg. angle	equal sides	••••	label
red	big	yes	no		yes
green	big	no	yes		yes
blue	small	no	yes		no
blue	small	yes	yes		no
:	••••			••••	••••
red	big	no	yes		yes

1. bad feature representation!



color	size	# slides	equal sides	••••	label
blue	big	3	no		yes
blue	big	3	yes		yes
red	small	inf	yes		no
green	small	4	yes		no
:	••••		••••	••••	••••
blue	big	3	yes		yes



color	size	# slides	equal sides	••••	label
blue	big	3	no		yes
blue	big	3	yes		yes
red	small	inf	yes		no
green	small	4	yes		no
blue	big	3	yes		yes

2. bad data + misleading correlations



color	size	# slides	equal sides	••••	label
white	big	3	no		yes
white	big	3	no		no
white	small	inf	yes		yes
white	small	4	yes		no
				-	
white	big	3	yes		yes

3. noisy training data!

• Linear classifier

$$y = \begin{cases} 1 & \text{if } w_0 + \sum_{j=1}^n w_j x_j > 0 \\ 0 & \text{otherwise} \end{cases}$$

• Linear classifier

$$y = \begin{cases} 1 & \text{if } w_0 + \sum_{j=1}^n w_j x_j > 0 \\ 0 & \text{otherwise} \end{cases}$$

parameters learned by the model predicted value (e.g., I = positive, 0 = negative)

test instance

model parameters



output = $2.0 + (0.50 \times -5.0) + (1.0 \times 2.0) + (0.2 \times 1.0)$

output = 1.7

output prediction = positive





(source: http://en.wikipedia.org/wiki/File:Svm_separating_hyperplanes.png)



• Would a linear classifier do well on positive (black) and negative (white) data that looks like this?







Evaluation Metric what could possibly go wrong?

 Most evaluation metrics can be understood using a <u>contingency table</u>

- MI IO

		uuu				
		triangle	other			
lictec	triangle	Α	В			
pred	other	С	D			

- What number(s) do we want to maximize?
- What number(s) do we want to minimize?
predictec

- True positives (A): number of triangles <u>correctly</u> predicted as triangles
- False positives (B): number of "other" <u>incorrectly</u> predicted as triangles
- False negatives (C): number of triangles <u>incorrectly</u> predicted as "other"
- True negatives (D): number of "other" <u>correctly</u> predicted as "other"

		triangle	other		
	triangle	Α	В		
	other	С	D		

true

• Accuracy: percentage of predictions that are correct (i.e., true positives <u>and</u> true negatives)

$$(? + ?)$$

$$(? + ? + ? + ?)$$
true
$$triangle \quad Other$$

$$briangle \quad A \quad B$$

$$briangle \quad C \quad D$$

 Accuracy: percentage of predictions that are correct (i.e., true positives <u>and</u> true negatives)



• Accuracy: percentage of predictions that are correct (i.e., true positives <u>and</u> true negatives)



• What is the accuracy of this model?



• Interpreting the value of a metric on a particular data set requires some thinking ...



 On this dataset, what would be the expected accuracy of a model that does NO learning (degenerate baseline)?

• Interpreting the value of a metric on a particular data set requires some thinking ...



5. Misleading interpretation of a metric value!



What Could Possibly Go Wrong?

- 1. Bad feature representation
- 2. Bad data + misleading correlations
- 3. Noisy labels for training and testing
- 4. Bad learning algorithm
- 5. Misleading evaluation metric

Outline

Introductions

What is Text Data Mining?

Predictive Analysis of Text: The Big Picture

Exploratory Analysis of Text: The Big Picture

Applications

Text Data Mining in this Course

- Predictive Analysis of Text
 - developing computer programs that automatically recognize a particular concept within a span of text
- Exploratory Analysis of Text:
 - developing computer programs that automatically discover useful patterns or trends in text collections

Exploratory Analysis example: clustering shapes



Exploratory Analysis example: clustering shapes





Exploratory Analysis example: clustering shapes



Exploratory Analysis representation: features

color	size	# slides	equal sides	••••	shape
blue	big	3	no		triangle
blue	big	3	yes		triangle
red	small	inf	yes		circle
green	small	4	yes		square
••••					
blue	big	3	yes		triangle



Exploratory Analysis basic ingredients

- 1. Data: a set of examples that we want to automatically analyze in order to discover interesting trends
- 2. Representation: a set of features that we believe are useful in describing the data (i.e., its main attributes)
- 3. Similarity Metric: a measure of similarity between two examples that is based on their feature values
- 4. Clustering algorithm: an algorithm that assigns items with similar feature values to the same group



Representation what could possibly go wrong?





Representation what could possibly go wrong?



Exploratory Analysis basic ingredients: the focus in this course

- 1. Data: a set of examples that we want to automatically analyze in order to discover interesting trends
- 2. Representation: a set of features that we believe are useful in describing the data
- 3. Similarity Metric: a measure of similarity between two examples that is based on their feature values
- 4. Clustering algorithm: an algorithm that assigns items with similar feature values to the same group

Text Data Mining in this Course

- Predictive Analysis of Text
 - developing computer programs that automatically recognize or detect a particular concept within a span of text
- Exploratory Analysis of Text:
 - developing computer programs that automatically discover interesting and useful patterns or trends in text collections

Outline

Introductions

What is Text Data Mining?

Predictive Analysis of Text: The Big Picture

Exploratory Analysis of Text: The Big Picture

Applications

Predictive Analysis of Text examples we'll cover in class

- Topic Categorization
- Opinion Mining
- Sentiment/Affect Analysis
- Bias Detection
- Information Extraction and Relation Learning
- Text-driven Forecasting
- Temporal Summarization

• Topic Categorization: automatically assigning documents to a set of pre-defined topical categories

Topic Categorization

dmoz open directory project					In par AOI	rtnership with Search.
	about dmoz	dmoz blog	suggest URL	help	link	editor login
		Sear	rch <u>advanced</u>	!		

Arts Movies, Television, Music...

Games

Kids and Teens

Arts, School Time, Teen Life

Reference Maps, Education, Libraries...

Shopping Clothing, Food, Gifts...

Business Jobs, Real Estate, Investing...

Health Video Games, RPGs, Gambling... Fitness, Medicine, Alternative...

> News Media, Newspapers, Weather...

Regional US, Canada, UK, Europe...

Society People, Religion, Issues... Computers Internet, Software, Hardware...

Home Family, Consumers, Cooking ...

Recreation Travel, Food, Outdoors, Humor ...

Science Biology, Psychology, Physics ...

Sports Baseball, Soccer, Basketball

World

Català, Dansk, Deutsch, Español, Français, Italiano, 日本語, Nederlands, Polski, Русский, Svenska...

Become an Editor Help build the largest human-edited directory of the web



Copyright © 2012 Netscape

Topic Categorization





Català, Dansk, Deutsch, Español, Français, Italiano, 日本語, Nederlands, Polski, Русский, Svenska...

Become an Editor Help build the largest human-edited directory of the web



Copyright © 2012 Netscape

 Opinion Mining: automatically detecting whether a span of opinionated text expresses a positive or negative opinion about the item being judged

Opinion Mining movie reviews

- "Great movie! It kept me on the edge of my seat the whole time. I IMAX-ed it and have no positive regrets."
- "Waste of time! It sucked!"
- "This film should be brilliant. It sounds like a great plot, the actors are first grade, and the supporting cast is good as well, and Stallone is attempting to deliver a good performance. However, it can't hold up."
 - negative

negative

• "Trust me, this movie is a masterpiece after ??? you've seen it 4+ times."

• Sentiment/Affect Analysis: automatically detecting the emotional state of the author of a span of text (usually from a set of pre-defined emotional states).

Sentiment Analysis support group posts

- "[I] also found out that the radiologist is doing the biopsy, not a breast surgeon. I am more scared now than when I ..."
- "... My radiologist 'assured' me my scan was despair NOT going to be cancer...she was wrong."
- "… My radiologist did my core biopsy. Not a hope problem and he did a super job of it."
- "It's pretty standard for the radiologist to do the biopsy so I wouldn't be concerned on that score."

fear

hope

• Bias detection: automatically detecting whether the author of a span of text favors a particular viewpoint (usually from a set of pre-defined viewpoints)

Bias Detection

- "Coming [up] next, drug addicted pregnant women no longer have anything to fear from pro-policy the authorities thanks to the Supreme Court. (vs. anti-policy) Both sides on this in a moment." -- Bill O'Reilly
- "Nationalizing businesses, nationalizing banks, is not a solution for the democratic party, it's the objective." -- Rush Limbaugh
- "If you're keeping score at home, so far our war in Iraq has created a police state in that country and socialism in Spain. So, no democracies yet, but we're really getting close." -- Jon Stewart

conservative (vs. liberal)

against war in iraq (vs. in favor of)

- Information extraction: automatically detecting that a short sequence of words belongs to (or is an instance of) a particular entity type, for example:
 - Person(X)
 - Location(X)
 - TennisPlayer(X)
 - ...

- Relation Learning: automatically detecting pairs of entities that share a particular relation, for example:
 - CEO(<person>,<company>)
 - Capital(<city>,<country>)
 - Mother(<person>,<person>)
 - ConvictedFelon(<person>,<crime>)
 -

Relation Learning CEO(<person>,<company>)

Marissa Mayer Yahoo

Know Yahoo's Marissa Mayer in 11 facts - CNN.com

www.cnn.com/2012/07/17/...marissa-mayer/index.html



...

by John D. Sutter - in 846,411 Google+ circles - More by John D. Sutter

Jul 19, 2012 - Here's a quick guide to some of the most interesting and water-

cooler-worthy facts about Marissa Mayer, who was named CEO of Yahoo on

<person>, who was named CEO of <company>

Q

Relation Learning CEO(<person>,<company>)

",who was named CEO of"

DailyTech - Fisker Appoints New CEO, Eliminates Battery/Engine ... www.dailytech.com/article.aspx?newsid=25412 4 days ago – Tom LaSorda, who was named CEO of Fisker back in February 2012 when founder Henrik Fisker stepped down, is leaving the company, but ...

who was named CEO of Yahoo on Monday. Christian Science Monitor gtp123.com/.../who-was-named-ceo-of-yahoo-on-monday-christian-... Jul 17, 2012 – You are browsing the archive for who was named CEO of Yahoo on Monday. Christian Science Monitor. Avatar of Garland E. Harris ...

<u>CEO of renamed Sara Lee meat biz chooses Winnetka - Residential ...</u> www.chicagorealestatedaily.com > Home > Residential News Aug 7, 2012 – Sean Connolly, who was named CEO of Hillshire Brands Co. in January, declines to comment through a company spokesman. Records show ...

Who is the woman who was named CEO of Gilt Groupe in Septemb... askville.amazon.com > Miscellaneous > Popular News Askville Question: Who is the woman who was named CEO of Gilt Groupe in September? : Popular News.

Tom McKillop - Wikipedia, the free encyclopedia

en.wikipedia.org/wiki/Tom_McKillop Sir Thomas Fulton Wilson "Tom" McKillop, FRS (born 19 March 1943) is a Scottish chemist, who was named CEO of AstraZeneca PLC in 1999 (retired 1 January ...

Harrison adjusts to view from top at First Hawaiian - Pacific Business ... www.bizjournals.com/.../harrison-adjusts-to-view-from-top-at.html?... Jan 27, 2012 – Bob Harrison, who was named CEO of First Hawaiian Bank on Jan. 1, says he'll spend a lot of time focusing on his people and community ...

CEO(Sean Connolly, Hillshire Brands)

CEO(woman, Gilt Groupe)

Q

CEO(Tom LaSorda, Fisker)

CEO(scottish chemist, AztraZeneca)

CEO(Bob Harrison, First Hawaiian Bank)

- Text-based Forecasting: monitoring incoming text (e.g., tweets) and making predictions about external, real-world events or trends, for example:
 - a presidential candidate's poll rating
 - a company's stock value change
 - a movie's box office earnings
 - side-effects for a particular drug
 -

- Temporal Summarization: monitoring incoming text (e.g., tweets) about a news event and predicting whether a sentence should be included in an on-going summary of the event
- Updates to the summary should contain relevant, novel, and accurate information. S_{t-1}

output?input d_{t-5} d_{t-4} d_{t-3} d_{t-2} d_{t-1} d_t

 \mathcal{B}_I

(Diagram from Guo et al., ECIR 2013)

- Detecting other interesting properties of text: [insert your crazy idea here], for example, detecting humorous text:
 - "Beauty is in the eye of the beholder" not funny
 - "Beauty is in the eye of the beer holder" funny

(example from Mihalcea and Pulman, 2007)
Outline

Introductions

What is Text Data Mining?

Predictive Analysis of Text: The Big Picture

Exploratory Analysis of Text: The Big Picture

Applications

Course Overview

Road Map first half of the semester

- Predictive Analysis of Text
 - Supervised machine learning principles
 - Text representation
 - Feature selection
 - Basic machine learning algorithms
 - Tools for predictive analysis of text
 - Experimentation and evaluation
- Exploratory Analysis of Text
 - Clustering
 - Outlier detection (tentative)
 - Co-occurrence statistics

Road Map second half of the semester

- Applications
 - Text classification
 - Opinion mining
 - Sentiment analysis
 - Bias detection
 - Information extraction
 - Relation learning
 - Text-based forecasting
 - Temporal Summarization
- Is there anything that <u>you</u> would like to learn more about?

Grading

- 30% homework
 - ▶ 10% each
- 20% midterm
- 40% term project
 - ► 5% proposal
 - 10% presentation
 - > 25% paper
- 10% participation

Grading for Graduate Students

- H: 95-100%
- P: 80-94%
- L: 60-79%
- F: 0-59%

Grading for Undergraduate Students

- A+: 97-100%
- A: 94-96%
- A-: 90-93%
- B+: 87-89%
- B: 84-86%
- B-: 80-83%
- C+: 77-79%
- C: 74-76%
- C-: 70-73%

- D+: 67-69%
- D: 64-66%
- D-: 60-63%
- F: <= 59%

General Outline of Homework

- Given a dataset (i.e., a training and test set), run experiments where you try to predict the target class using different feature representations
- Do error analysis
- Report on what worked, what didn't, and why!
- Answer essay questions about the assignment
 - These will be associated with the course material

Homework vs. Midterm



• The homework will be more challenging than the midterm. It should be, you have more time.

Course Tips

- Work hard
- Do the assigned readings
- Do <u>other</u> readings
- Be patient and have reasonable expectations
 - you're not supposed to understand everything we cover in class <u>during</u> class
- Seek help sooner rather than later
 - office hours: by appointment
 - questions via email
- Remember the golden rule: no pain, no gain

Questions?