How to Read 10,000 Blogs During This Talk

Daniel J. Hopkins[1]
Assistant Professor
Department of Government
Georgetown University
Presentation at the University of Kentucky

March 9, 2011

[1] This talk is based on co-authored work with Gary King.

Automated Content Analysis

- Classification of documents by hand $\longrightarrow$ central tool in political science

Automated Content Analysis

- Classification of documents by hand → central tool in political science
- New applications: explosive increase in web pages, blogs, emails, digitized books and articles, audio recordings (automatically converted to text), government reports, legislative hearings and records, electronic medical records, etc.

Automated Content Analysis

- Classification of documents by hand → central tool in political science
- New applications: explosive increase in web pages, blogs, emails, digitized books and articles, audio recordings (automatically converted to text), government reports, legislative hearings and records, electronic medical records, etc.
- Rutherford D. Roger: "We are drowning in information and starving for knowledge" (Hastie et al. 2001:vii)

Content Analysis

Content Analysis

- Automated methods are essential

Content Analysis

- Automated methods are essential
- Possibilities include:

Content Analysis

- Automated methods are essential
- Possibilities include:
  - Agenda-setting (Quinn et al. 2010)

Content Analysis

- Automated methods are essential
- Possibilities include:
  - Agenda-setting (Quinn et al. 2010)
  - Campaigns (Leskovec et al. 2009)

Content Analysis

- Automated methods are essential
- Possibilities include:
  - Agenda-setting (Quinn et al. 2010)
  - Campaigns (Leskovec et al. 2009)
  - Party positions (Laver et al. 2003)

Content Analysis

- Automated methods are essential
- Possibilities include:
  - Agenda-setting (Quinn et al. 2010)
  - Campaigns (Leskovec et al. 2009)
  - Party positions (Laver et al. 2003)
  - Legislative behavior (Monroe et al. 2008; Grimmer 2010)

Content Analysis

- Automated methods are essential
- Possibilities include:
  - Agenda-setting (Quinn et al. 2010)
  - Campaigns (Leskovec et al. 2009)
  - Party positions (Laver et al. 2003)
  - Legislative behavior (Monroe et al. 2008; Grimmer 2010)
  - Measuring public opinion (e.g. Pang et al. 2002; Hopkins and King 2010, O'Connor et al. 2010)

Content Analysis

- Dates to the 1600s: The Church tracked nonreligious texts by classifying newspaper stories

Content Analysis

- Dates to the 1600s: The Church tracked nonreligious texts by classifying newspaper stories
- Early approaches: dictionary-based

Content Analysis

- Dates to the 1600s: The Church tracked nonreligious texts by classifying newspaper stories
- Early approaches: dictionary-based
- e.g. deterministic mapping from words $\rightarrow$ categories

Content Analysis

- Dates to the 1600s: The Church tracked nonreligious texts by classifying newspaper stories
- Early approaches: dictionary-based
- e.g. deterministic mapping from words → categories
- e.g. Lehman Brothers oversight → searches for 23 phrases like "stupid," "huge mistake," etc. (Goldstein 2010)

Content Analysis

- Dates to the 1600s: The Church tracked nonreligious texts by classifying newspaper stories
- Early approaches: dictionary-based
- e.g. deterministic mapping from words $\rightarrow$ categories
- e.g. Lehman Brothers oversight $\rightarrow$ searches for 23 phrases like "stupid," "huge mistake," etc. (Goldstein 2010)
- Dictionary-based methods: inflexible; heavy reliance on user knowledge

Beyond the Dictionary

- Core conceptual distinction: unsupervised learning vs. supervised learning

Beyond the Dictionary

- Core conceptual distinction: unsupervised learning vs. supervised learning
- Discovery vs. Data Extension

Beyond the Dictionary

- Core conceptual distinction: unsupervised learning vs. supervised learning
- Discovery vs. Data Extension
- Separate distinction between sentiment analysis (e.g. Pang et al. 2002) and topic classification (Quinn et al. 2010)

Beyond the Dictionary

- Core conceptual distinction: unsupervised learning vs. supervised learning
- Discovery vs. Data Extension
- Separate distinction between sentiment analysis (e.g. Pang et al. 2002) and topic classification (Quinn et al. 2010)
- This talk: provides supervised technique for data extension, main application is to sentiment analysis

Outline

- Introduction: Why Automated Content Analysis

Outline

- Introduction: Why Automated Content Analysis
- Motivating Example: Opinions in Blogs

Outline

- Introduction: Why Automated Content Analysis
- Motivating Example: Opinions in Blogs
- Background on Supervised Learning

Outline

- Introduction: Why Automated Content Analysis
- Motivating Example: Opinions in Blogs
- Background on Supervised Learning
- Goals for the Estimator

Outline

- Introduction: Why Automated Content Analysis
- Motivating Example: Opinions in Blogs
- Background on Supervised Learning
- Goals for the Estimator
- Preprocessing: From Text to Data

Outline

- Introduction: Why Automated Content Analysis
- Motivating Example: Opinions in Blogs
- Background on Supervised Learning
- Goals for the Estimator
- Preprocessing: From Text to Data
- The Nonparametric Estimator (the math)

Outline

- Introduction: Why Automated Content Analysis
- Motivating Example: Opinions in Blogs
- Background on Supervised Learning
- Goals for the Estimator
- Preprocessing: From Text to Data
- The Nonparametric Estimator (the math)
- Empirical Tests: Blogs, Editorials, etc.

Goals

- Introduction to possibilities of automated content analysis

Goals

How to Read 10,000 Blogs During This Talk

Daniel J. Hopkins

Introduction

Opinions in Blogs

Background

Goals for Estimator

Preprocessing

Nonparametric Estimator

Conclusion

Additional Material

- Introduction to possibilities of automated content analysis
- Argument that most computer science techniques $\rightarrow$ optimize for a different goal

Goals

- Introduction to possibilities of automated content analysis
- Argument that most computer science techniques $\longrightarrow$ optimize for a different goal
- Outline our nonparametric estimator to estimate category proportions

Ex. Public Opinion

- `http://www.youtube.com/watch?v=dRjUubkhmv4`

Ex. Public Opinion

- `http://www.youtube.com/watch?v=dRjUubkhmv4`
- Hand-code 442 blog posts in early November 2006 about John Kerry

Ex. Public Opinion

- http://www.youtube.com/watch?v=dRjUubkhmv4
- Hand-code 442 blog posts in early November 2006 about John Kerry
- Identify pro-, anti-Kerry sentiment

Ex. Public Opinion

- `http://www.youtube.com/watch?v=dRjUubkhmv4`
- Hand-code 442 blog posts in early November 2006 about John Kerry
- Identify pro-, anti-Kerry sentiment
- Apply model to 10,000 blog posts

Ex. Public Opinion

- http://www.youtube.com/watch?v=dRjUubkhmv4
- Hand-code 442 blog posts in early November 2006 about John Kerry
- Identify pro-, anti-Kerry sentiment
- Apply model to 10,000 blog posts
- Retrospective measure of opinion

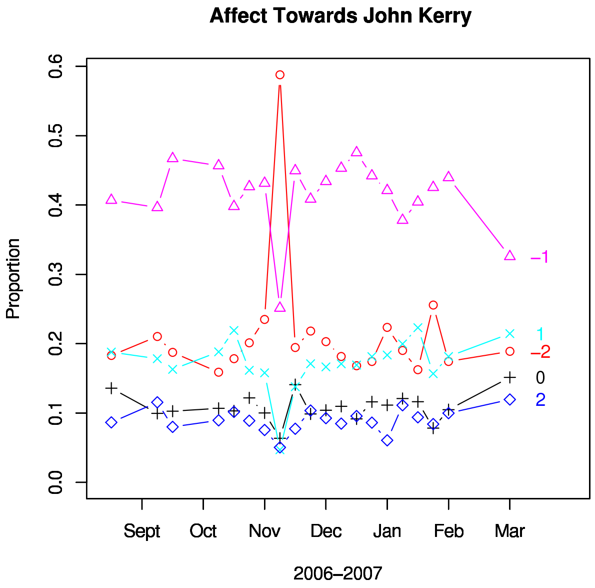


Figure: From Hopkins and King (2010)

Supervised Learning

- Supervised learning: analyze subset of texts to identify mapping from features (typically words) to categories

Supervised Learning

- Supervised learning: analyze subset of texts to identify mapping from features (typically words) to categories
- Tool for data extension

Supervised Learning

- Advantages:

Supervised Learning

- Advantages:
  - Allows for extensions beyond limits of hand-coding (e.g. 10,000 blogs)

Supervised Learning

- Advantages:
  - Allows for extensions beyond limits of hand-coding (e.g. 10,000 blogs)
  - Requires little interpretation after analysis

Supervised Learning

- Advantages:
  - Allows for extensions beyond limits of hand-coding (e.g. 10,000 blogs)
  - Requires little interpretation after analysis
- Disadvantages:

Supervised Learning

- Advantages:
  - Allows for extensions beyond limits of hand-coding (e.g. 10,000 blogs)
  - Requires little interpretation after analysis
- Disadvantages:
  - Not necessary if random sample is sufficient

Supervised Learning

- Advantages:
  - Allows for extensions beyond limits of hand-coding (e.g. 10,000 blogs)
  - Requires little interpretation after analysis
- Disadvantages:
  - Not necessary if random sample is sufficient
  - Significant pre-analysis costs

Coding Scheme

- Question: affect about President Bush and 2008 candidates

Coding Scheme

- Question: affect about President Bush and 2008 candidates
- Specific categories:

Label	Category
−2	extremely negative
−1	negative
0	neutral
1	positive
2	extremely positive
NA	no opinion expressed
NB	not a blog

Coding Scheme

- Question: affect about President Bush and 2008 candidates
- Specific categories:

Label	Category
-2	extremely negative
-1	negative
0	neutral
1	positive
2	extremely positive
NA	no opinion expressed
NB	not a blog

- Hard case:

Coding Scheme

- Question: affect about President Bush and 2008 candidates
- Specific categories:

Label	Category
-2	extremely negative
-1	negative
0	neutral
1	positive
2	extremely positive
NA	no opinion expressed
NB	not a blog

- Hard case:
  - Part ordinal, part nominal categorization

Coding Scheme

- Question: affect about President Bush and 2008 candidates
- Specific categories:

Label	Category
-2	extremely negative
-1	negative
0	neutral
1	positive
2	extremely positive
NA	no opinion expressed
NB	not a blog

- Hard case:
  - Part ordinal, part nominal categorization
  - "Sentiment categorization is more difficult than topic classification"

Coding Scheme

- Question: affect about President Bush and 2008 candidates
- Specific categories:

Label	Category
-2	extremely negative
-1	negative
0	neutral
1	positive
2	extremely positive
NA	no opinion expressed
NB	not a blog

- Hard case:
  - Part ordinal, part nominal categorization
  - "Sentiment categorization is more difficult than topic classification"
  - Language ranges from "my crunchy gf thinks dubya hid the wmd's, :)!" to the Queen's English

Inter-coder reliability

	-2	-1	0	1	2	NA	NB
-2	**.70**	.10	.01	.01	.00	.02	.16
-1	.33	**.25**	.04	.02	.01	.01	.35
0	.13	.17	**.13**	.11	.05	.02	.40
1	.07	.06	.08	**.20**	.25	.01	.34
2	.03	.03	.03	.22	**.43**	.01	.25
NA	.04	.01	.00	.00	.00	**.81**	.14
NB	.10	.07	.02	.02	.02	.04	**.75**

Inputs

Available Inputs:

- Large set of text documents

Inputs

Available Inputs:

- Large set of text documents
- A set of mutually exclusive and exhaustive categories

Inputs

Available Inputs:

- Large set of text documents
- A set of mutually exclusive and exhaustive categories
- A small subset of documents hand-coded into the categories

Quantities of Interest

Common Quantities of interest

- individual document classification

Maximizing one goal won't get you the other: high classification accuracy can coexist with huge biases in category proportions

Quantities of Interest

Common Quantities of interest

- individual document classification
- proportion of documents in each category

Maximizing one goal won't get you the other: high classification accuracy can coexist with huge biases in category proportions

Quantities of Interest

Common Quantities of interest

- individual document classification
- proportion of documents in each category
- *Can* get the 2nd by aggregating the 1st (but not necessary)

Maximizing one goal won't get you the other: high classification accuracy can coexist with huge biases in category proportions

Quantities of Interest

Common Quantities of interest

- individual document classification
- proportion of documents in each category
- *Can* get the 2nd by aggregating the 1st (but not necessary)
- E.g., classify constituents' letters to a member of congress by policy area, or estimate proportion of letters in each policy area

Maximizing one goal won't get you the other: high classification accuracy can coexist with huge biases in category proportions

Quantities of Interest

Common Quantities of interest

- individual document classification
- proportion of documents in each category
- *Can* get the 2nd by aggregating the 1st (but not necessary)
- E.g., classify constituents' letters to a member of congress by policy area, or estimate proportion of letters in each policy area
- E.g., classify emails as spam or not, or estimate proportion of email that is spam

Maximizing one goal won't get you the other: high classification accuracy can coexist with huge biases in category proportions

Core Idea

- Survey researchers: care about population parameters, *not* any specific person's approval of President

Core Idea

- Survey researchers: care about population parameters, *not* any specific person's approval of President
- Social scientists typically interested in characterizing populations, not individual texts

Core Idea

- Survey researchers: care about population parameters, *not* any specific person's approval of President
- Social scientists typically interested in characterizing populations, not individual texts
- Can estimate population proportions without estimating individual document categories

This Nonparametric Approach

- Gives unbiased estimates of population proportions

This Nonparametric Approach

- Gives unbiased estimates of population proportions
- Works better than aggregating imperfect classification methods

This Nonparametric Approach

- Gives unbiased estimates of population proportions
- Works better than aggregating imperfect classification methods
- No problem if classification accuracy is low

This Nonparametric Approach

- Gives unbiased estimates of population proportions
- Works better than aggregating imperfect classification methods
- No problem if classification accuracy is low
- No parametric modeling assumptions

This Nonparametric Approach

- Gives unbiased estimates of population proportions
- Works better than aggregating imperfect classification methods
- No problem if classification accuracy is low
- No parametric modeling assumptions
- The hand coded subset need not be a random sample

This Nonparametric Approach

- Gives unbiased estimates of population proportions
- Works better than aggregating imperfect classification methods
- No problem if classification accuracy is low
- No parametric modeling assumptions
- The hand coded subset need not be a random sample
- Scales to large numbers of documents

This Nonparametric Approach

- Gives unbiased estimates of population proportions
- Works better than aggregating imperfect classification methods
- No problem if classification accuracy is low
- No parametric modeling assumptions
- The hand coded subset need not be a random sample
- Scales to large numbers of documents
- Software available: readme() function in ReadMe

This Nonparametric Approach

- Gives unbiased estimates of population proportions
- Works better than aggregating imperfect classification methods
- No problem if classification accuracy is low
- No parametric modeling assumptions
- The hand coded subset need not be a random sample
- Scales to large numbers of documents
- Software available: readme() function in ReadMe
- Our core assumption: relationship between words, categories constant between labeled, unlabeled sets

From Text to Data

- You have millions of blog posts. Now what? Dimension reduction

From Text to Data

- You have millions of blog posts. Now what? Dimension reduction
- Goal for *both* supervised, unsupervised analyses: transform articles into term-frequency matrix

From Text to Data

- You have millions of blog posts. Now what? Dimension reduction
- Goal for *both* supervised, unsupervised analyses: transform articles into term-frequency matrix
- Rows: documents

From Text to Data

- You have millions of blog posts. Now what? Dimension reduction
- Goal for *both* supervised, unsupervised analyses: transform articles into term-frequency matrix
- Rows: documents
- Columns: unique word strings

Representing Text as Numbers

- Filter: choose English language blogs that mention Bush ("Bush", "George W.", "Dubya", "King George", etc.), Hillary Clinton ("Senator Clinton", "Hillary", "Hitlery", "Mrs. Clinton"), etc.

Representing Text as Numbers

- Filter: choose English language blogs that mention Bush ("Bush", "George W.", "Dubya", "King George", etc.), Hillary Clinton ("Senator Clinton", "Hillary", "Hitlery", "Mrs. Clinton"), etc.
- Preprocess: convert to lower case, remove punctuation, perform stemming (reduce "consist", "consisted", "consistency", "consistent", "consistently", "consisting", and "consists", to their stem: "consist")

Representing Text as Numbers

- Filter: choose English language blogs that mention Bush ("Bush", "George W.", "Dubya", "King George", etc.), Hillary Clinton ("Senator Clinton", "Hillary", "Hitlery", "Mrs. Clinton"), etc.
- Preprocess: convert to lower case, remove punctuation, perform stemming (reduce "consist", "consisted", "consistency", "consistent", "consistently", "consisting", and "consists", to their stem: "consist")
- Stop words: some analyses remove very common words (e.g. "the," "almost")

Representing Text as Numbers

- Filter: choose English language blogs that mention Bush ("Bush", "George W.", "Dubya", "King George", etc.), Hillary Clinton ("Senator Clinton", "Hillary", "Hitlery", "Mrs. Clinton"), etc.
- Preprocess: convert to lower case, remove punctuation, perform stemming (reduce "consist", "consisted", "consistency", "consistent", "consistently", "consisting", and "consists", to their stem: "consist")
- Stop words: some analyses remove very common words (e.g. "the," "almost")
- Code variables as number/presence of unique unigrams, bigrams, trigrams, etc.

An Example

- Our 10,771 blog posts about Bush and Clinton: 201,676 unigrams, 2,392,027 bigrams, 5,761,979 trigrams.

An Example

- Our 10,771 blog posts about Bush and Clinton: 201,676 unigrams, 2,392,027 bigrams, 5,761,979 trigrams.
- Unigrams in $> 1\%$ or $< 99\%$ of documents: 3,672 variables

Bag of Words

- This = “bag of words” approach

Bag of Words

- This = "bag of words" approach
- Word order is discarded (but can tag each word with its part of speech)

Bag of Words

- This = "bag of words" approach
- Word order is discarded (but can tag each word with its part of speech)
- Negation ignored (although that can be fixed by making "not good" one string)

Bag of Words

- This = "bag of words" approach
- Word order is discarded (but can tag each word with its part of speech)
- Negation ignored (although that can be fixed by making "not good" one string)
- Be (2), Not (1), Or (1), To (2)

Bag of Words

- This = "bag of words" approach
- Word order is discarded (but can tag each word with its part of speech)
- Negation ignored (although that can be fixed by making "not good" one string)
- Be (2), Not (1), Or (1), To (2)
- Typically provides reasonable predictive power (e.g. Pang et al. 2002)

Notation

- Document Category

$$D_i = \begin{cases} \text{-2} & \text{extremely negative} \\ \text{-1} & \text{negative} \\ 0 & \text{neutral} \\ 1 & \text{positive} \\ 2 & \text{extremely positive} \\ \text{NA} & \text{no opinion expressed} \\ \text{NB} & \text{not a blog} \end{cases}$$

Notation

- Word Stem Profile:

$$S_i = \begin{cases} S_{i1} = 1 & \text{if "awful" is used, 0 if not} \\ S_{i2} = 1 & \text{if "good" is used, 0 if not} \\ \vdots & \vdots \\ S_{iK} = 1 & \text{if "zoo" is used, 0 if not} \end{cases}$$

Quantities of Interest

- Individual document classifications

$$D_1, D_2 \ldots, D_L$$

Quantities of Interest

- Individual document classifications

$$D_1, D_2 \ldots, D_L$$

- proportions in each category

$$P(D) = \begin{pmatrix} P(D=-2) \\ P(D=-1) \\ P(D=0) \\ P(D=1) \\ P(D=2) \\ P(D=\mathrm{NA}) \\ P(D=\mathrm{NB}) \end{pmatrix}$$

Quantities of Interest

- Sensitivity, sens $\equiv P(\hat{D} = 1 | D = 1)$

Quantities of Interest

- Sensitivity, sens $\equiv P(\hat{D} = 1|D = 1)$
- Specificity, spec $\equiv P(\hat{D} = 2|D = 2)$

Quantities of Interest

- Sensitivity, sens $\equiv P(\hat{D} = 1 | D = 1)$
- Specificity, spec $\equiv P(\hat{D} = 2 | D = 2)$
- Core intuition: if we know misclassification rates, we can adjust any estimator to produce unbiased category proportions

Quantities of Interest

- Sensitivity, sens $\equiv P(\hat{D} = 1 | D = 1)$
- Specificity, spec $\equiv P(\hat{D} = 2 | D = 2)$
- Core intuition: if we know misclassification rates, we can adjust any estimator to produce unbiased category proportions
- To know overall population parameters, we don't need to know *which* we mis-classified

Formalization from Epidemiology (Levy and Kass, 1970)

- Accounting identity for 2 categories:

$$P(\hat{D}=1) = (\text{sens})P(D=1) + (1-\text{spec})P(D=2)$$

Formalization from Epidemiology (Levy and Kass, 1970)

- Accounting identity for 2 categories:

$$P(\hat{D}=1) = (\text{sens})P(D=1) + (1-\text{spec})P(D=2)$$

- Solve:

$$P(D=1) = \frac{P(\hat{D}=1) - (1-\text{spec})}{\text{sens} - (1-\text{spec})}$$

Formalization from Epidemiology (Levy and Kass, 1970)

- Accounting identity for 2 categories:

$$P(\hat{D}=1) = (\text{sens})P(D=1) + (1-\text{spec})P(D=2)$$

- Solve:

$$P(D=1) = \frac{P(\hat{D}=1) - (1-\text{spec})}{\text{sens} - (1-\text{spec})}$$

- Use this equation to correct $P(\hat{D})$

Generalizations: J Categories, No Individual

- From King and Lu (2007)

Generalizations: J Categories, No Individual

- From King and Lu (2007)
- Accounting identity for J categories

$$P(\hat{D} = j) = \sum_{j'=1}^{J} P(\hat{D} = j | D = j')P(D = j')$$

Generalizations: J Categories, No Individual

- From King and Lu (2007)
- Accounting identity for J categories

$$P(\hat{D} = j) = \sum_{j'=1}^{J} P(\hat{D} = j | D = j')P(D = j')$$

- Drop $\hat{D}$ calculation, since $\hat{D} = f(S)$:

$$P(S = s) = \sum_{j'=1}^{J} P(S = s | D = j')P(D = j')$$

Generalizations: J Categories, No Individual

- From King and Lu (2007)
- Accounting identity for J categories

$$P(\hat{D} = j) = \sum_{j'=1}^{J} P(\hat{D} = j | D = j') P(D = j')$$

- Drop $\hat{D}$ calculation, since $\hat{D} = f(S)$:

$$P(S = s) = \sum_{j'=1}^{J} P(S = s | D = j') P(D = j')$$

- Simplify to an equivalent matrix expression:

$$P(S) = P(S|D)P(D)$$

Estimation

The matrix expression again:

$$\underset{2^K \times 1}{P(S)} = \underset{2^K \times J}{P(S|D)} \underset{J \times 1}{P(D)}$$

Estimation

The matrix expression again:

$$\underset{2^K \times 1}{P(S)} = \underset{2^K \times J}{P(S|D)} \underset{J \times 1}{P(D)}$$

Document category proportions (quantity of interest)

Estimation

The matrix expression again:

$$\underset{2^K \times 1}{P(S)} = \underset{2^K \times J}{P(S|D)} \underset{J \times 1}{P(D)}$$

Word stem profile proportions (estimate in unlabeled set by tabulation)

Estimation

The matrix expression again:

$$\underset{2^K \times 1}{P(S)} = \underset{2^K \times J}{P(S|D)} \underset{J \times 1}{P(D)}$$

Word stem profiles, by category (estimate in *labeled* set by tabulation)

Estimation

The matrix expression again:

$$\underset{2^K \times 1}{P(S)} = \underset{2^K \times J}{P(S|D)} \underset{J \times 1}{P(D)}$$

$$\Longrightarrow \; Y = X\beta$$

Alternative symbols (to emphasize the linear equation)

Estimation

The matrix expression again:

$$\underset{2^K\times 1}{P(S)} = \underset{2^K\times J}{P(S|D)}\underset{J\times 1}{P(D)}$$

$$\implies Y = X\beta \quad \implies \quad \beta = (X'X)^{-1}X'y$$

Solve for quantity of interest (with no error term)

Estimation

The matrix expression again:

$$\underset{2^K\times 1}{P(S)} = \underset{2^K\times J}{P(S|D)}\underset{J\times 1}{P(D)}$$

$$\Longrightarrow \; Y = X\beta \quad \Longrightarrow \quad \beta = (X'X)^{-1}X'y$$

- Technical estimation issues:

Estimation

The matrix expression again:

$$\underset{2^K\times 1}{P(S)} = \underset{2^K\times J}{P(S|D)}\underset{J\times 1}{P(D)}$$

$$\Longrightarrow\ Y = X\beta \quad \Longrightarrow \quad \beta = (X'X)^{-1}X'y$$

- Technical estimation issues:
  - 2^K is enormous, far larger than any existing computer

Estimation

The matrix expression again:

$$\underset{2^K\times 1}{P(S)} = \underset{2^K\times J}{P(S|D)}\underset{J\times 1}{P(D)}$$

$$\Longrightarrow \; Y = X\beta \quad \Longrightarrow \quad \beta = (X'X)^{-1}X'y$$

- Technical estimation issues:
  - 2^K is enormous, far larger than any existing computer
  - $P(S)$ and $P(S|D)$ will be too sparse

Estimation

The matrix expression again:

$$\underset{2^K\times 1}{P(S)} = \underset{2^K\times J}{P(S|D)}\underset{J\times 1}{P(D)}$$

$$\Longrightarrow \quad Y = X\beta \quad \Longrightarrow \quad \beta = (X'X)^{-1}X'y$$

- Technical estimation issues:
  - 2^K is enormous, far larger than any existing computer
  - $P(S)$ and $P(S|D)$ will be too sparse
  - Elements of $P(D)$ must be between 0 and 1 and sum to 1

Estimation

The matrix expression again:

$$\underset{2^K \times 1}{P(S)} = \underset{2^K \times J}{P(S|D)} \underset{J \times 1}{P(D)}$$

$$\implies Y = X\beta \implies \beta = (X'X)^{-1}X'y$$

- Technical estimation issues:
  - 2^K is enormous, far larger than any existing computer
  - $P(S)$ and $P(S|D)$ will be too sparse
  - Elements of $P(D)$ must be between 0 and 1 and sum to 1
- Solutions

Estimation

The matrix expression again:

$$\underset{2^K \times 1}{P(S)} = \underset{2^K \times J}{P(S|D)} \underset{J \times 1}{P(D)}$$

$$\implies Y = X\beta \quad \implies \quad \beta = (X'X)^{-1}X'y$$

- Technical estimation issues:
  - 2^K is enormous, far larger than any existing computer
  - $P(S)$ and $P(S|D)$ will be too sparse
  - Elements of $P(D)$ must be between 0 and 1 and sum to 1
- Solutions
  - Use subsets of S; average results

Estimation

The matrix expression again:

$$\underset{2^K \times 1}{P(S)} = \underset{2^K \times J}{P(S|D)}\,\underset{J \times 1}{P(D)}$$

$$\implies Y = X\beta \quad \implies \quad \beta = (X'X)^{-1}X'y$$

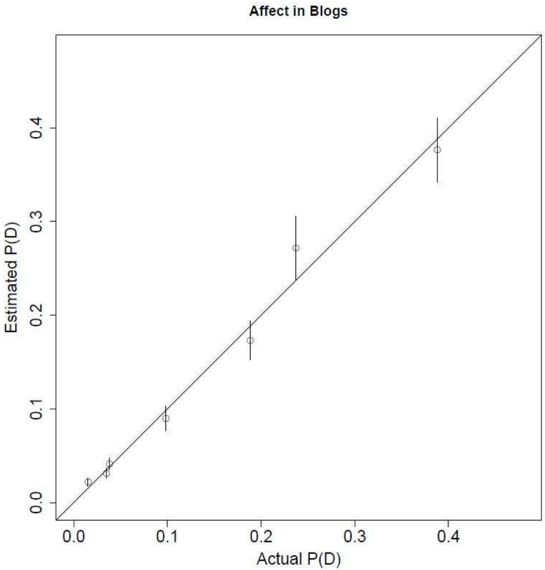
- Technical estimation issues:
  - 2^K is enormous, far larger than any existing computer
  - $P(S)$ and $P(S|D)$ will be too sparse
  - Elements of $P(D)$ must be between 0 and 1 and sum to 1
- Solutions
  - Use subsets of S; average results
  - Use constrained LS to constrain $P(D)$ to simplex

Comparing Performance

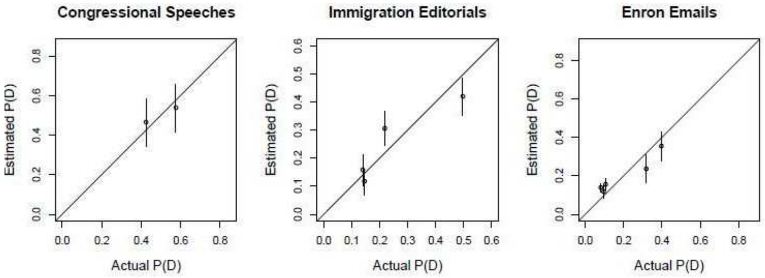
	Percent of Blog Posts Correctly Classified			
	In-Sample Fit	In-Sample Cross-Validation	Out-of-Sample Prediction	Mean Absolute Proportion Error
Nonparametric	—	—	—	1.2
Linear	67.6	55.2	49.3	7.7
Radial	67.6	54.2	49.1	7.7
Polynomial	99.7	48.9	47.8	5.3
Sigmoid	15.6	15.6	18.2	23.2

Table: Performance of our Nonparametric Approach and Four Support Vector Machine Analyses.

Out of Sample Validation: Blogs



Out of Sample Validation: Other Arenas



What can go wrong?

What can go wrong?

- We assume $P^h(S|D) = P(S|D)$

What can go wrong?

- We assume $P^h(S|D) = P(S|D)$
- Must choose word stem subset size (a smoothing parameter)

What can go wrong?

- We assume $P^h(S|D) = P(S|D)$
- Must choose word stem subset size (a smoothing parameter)
- Need enough labeled documents in each category (can hand code more if CI's are too large)

What can go wrong?

- We assume $P^h(S|D) = P(S|D)$
- Must choose word stem subset size (a smoothing parameter)
- Need enough labeled documents in each category (can hand code more if CI's are too large)
- Need sufficient information in: documents, categorization scheme, numerical summaries of the documents, and hand-codings

What can go wrong?

- We assume $P^h(S|D) = P(S|D)$
- Must choose word stem subset size (a smoothing parameter)
- Need enough labeled documents in each category (can hand code more if CI's are too large)
- Need sufficient information in: documents, categorization scheme, numerical summaries of the documents, and hand-codings
- Use additional hand coding to verify assumptions

Conclusion

- Computer science $\rightarrow$ developed huge range of supervised, unsupervised techniques (e.g. SVM, LDA)

Conclusion

- Computer science $\longrightarrow$ developed huge range of supervised, unsupervised techniques (e.g. SVM, LDA)
- Automated techniques $\longrightarrow$ open many areas of inquiry for political scientists of all stripes

Conclusion

- Computer science → developed huge range of supervised, unsupervised techniques (e.g. SVM, LDA)
- Automated techniques → open many areas of inquiry for political scientists of all stripes
- Core distinction: supervised vs. unsupervised techniques

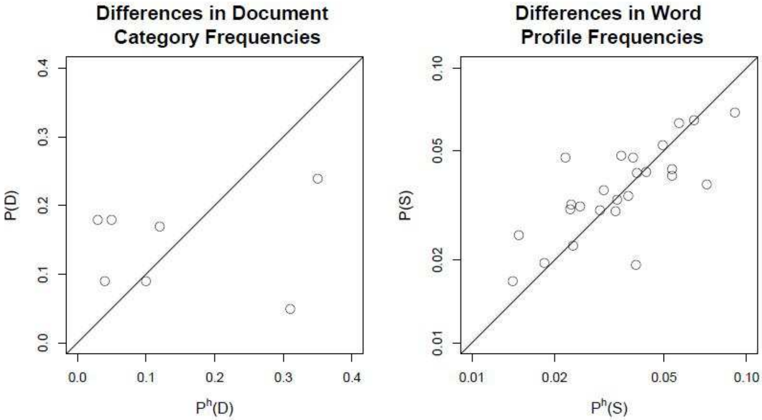
Conclusion

- Computer science $\rightarrow$ developed huge range of supervised, unsupervised techniques (e.g. SVM, LDA)
- Automated techniques $\rightarrow$ open many areas of inquiry for political scientists of all stripes
- Core distinction: supervised vs. unsupervised techniques
- For supervised learning, computer scientists' typical goal $\neq$ political scientists'

Conclusion

- Computer science $\rightarrow$ developed huge range of supervised, unsupervised techniques (e.g. SVM, LDA)
- Automated techniques $\rightarrow$ open many areas of inquiry for political scientists of all stripes
- Core distinction: supervised vs. unsupervised techniques
- For supervised learning, computer scientists' typical goal $\neq$ political scientists'
- ReadMe designed to return unbiased estimates of category proportions

A Nonrandom Hand-coded Sample



Coding categories

Effective coding categories:

- Mutually exclusive and exhaustive

Coding categories

Effective coding categories:

- Mutually exclusive and exhaustive
- Simple is better: project began with 20 unordered categories (e.g. "hopeful"), ended with five (e.g. "strongly positive")

Coding categories

Effective coding categories:

- Mutually exclusive and exhaustive
- Simple is better: project began with 20 unordered categories (e.g. "hopeful"), ended with five (e.g. "strongly positive")
- Problem of "character" vs. "policy" distinction

Coding categories

Effective coding categories:

- Mutually exclusive and exhaustive
- Simple is better: project began with 20 unordered categories (e.g. "hopeful"), ended with five (e.g. "strongly positive")
- Problem of "character" vs. "policy" distinction
- Produce coding manual clear enough that it is sufficient for accurate coding

Coding categories

Effective coding categories:

- Mutually exclusive and exhaustive
- Simple is better: project began with 20 unordered categories (e.g. "hopeful"), ended with five (e.g. "strongly positive")
- Problem of "character" vs. "policy" distinction
- Produce coding manual clear enough that it is sufficient for accurate coding
- Burn-in period for project, coders (one major project: six months!)

Coding categories

Effective coding categories:

- Mutually exclusive and exhaustive
- Simple is better: project began with 20 unordered categories (e.g. "hopeful"), ended with five (e.g. "strongly positive")
- Problem of "character" vs. "policy" distinction
- Produce coding manual clear enough that it is sufficient for accurate coding
- Burn-in period for project, coders (one major project: six months!)
- Measure inter-coder reliability

An Unsupervised Example: LDA

	1	2	3	4	5	6	7
1	don't	know	they	illeg	law	they	about
2	you	you	english	here	the	here	church
3	peopl	differ	languag	fine	we're	and	would
4	there	communiti	speak	pay	enforc	want	you
5	are	american	them	they're	that	their	immigr
6	job	veri	they're	legal	togeth	get	that
7	mani	like	their	tax	about	money	cathol
8	know	more	know	who	was	back	like
9	problem	your	learn	you	this	work	say
10	there	and	our	should	down	lot	i'm
Prop.	0.144	0.122	0.139	0.157	0.142	0.141	0.155

Table: Clustering 836 comments from focus groups on immigration using 165 word stems, LDA.

Campaign Quotations

Source: Leskovec, Backstrom, and Klineberg (2009)

- Cull through 90 million new articles from 1.6 million websites

Campaign Quotations

Source: Leskovec, Backstrom, and Klineberg (2009)

- Cull through 90 million new articles from 1.6 million websites
- Identify variants of text strings from 2008 U.S. Presidential campaign

Campaign Quotations

Source: Leskovec, Backstrom, and Klineberg (2009)

- Cull through 90 million new articles from 1.6 million websites
- Identify variants of text strings from 2008 U.S. Presidential campaign
- 94,700 distinct phrases

Campaign Quotations

Source: Leskovec, Backstrom, and Klineberg (2009)

- Cull through 90 million new articles from 1.6 million websites
- Identify variants of text strings from 2008 U.S. Presidential campaign
- 94,700 distinct phrases
- Many research opportunities: study campaign dynamics, back-and-forth of campaign rhetoric

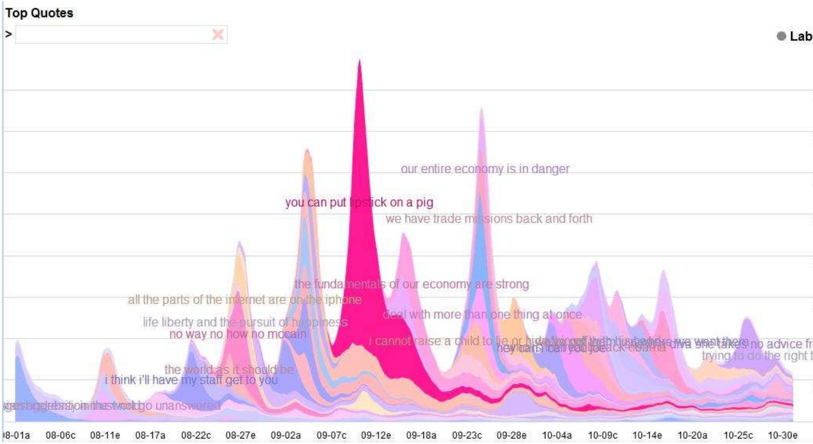


Figure: From Leskovec et al. (2009)

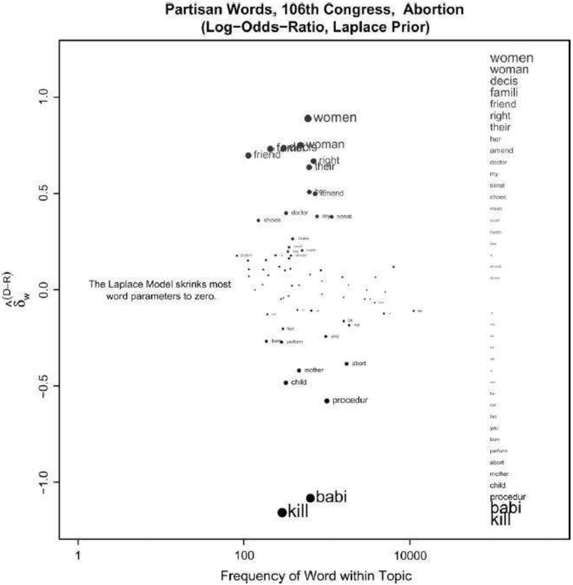


Figure: From Monroe et al. (2008)

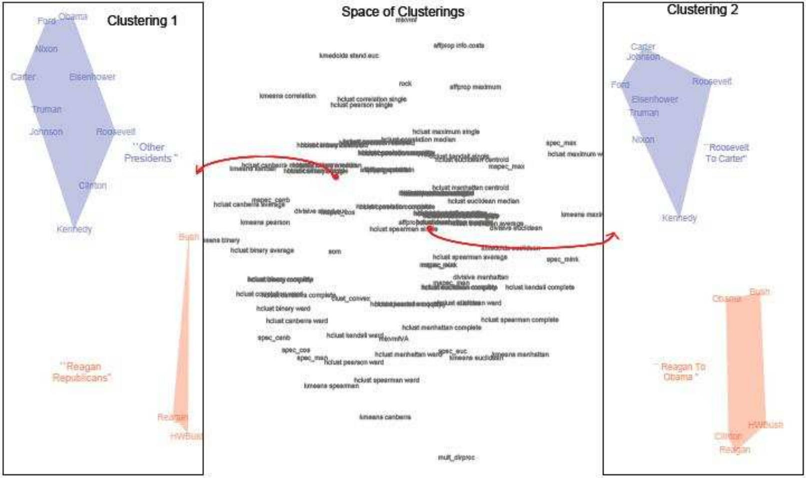


Figure: From Grimmer and King (2010)

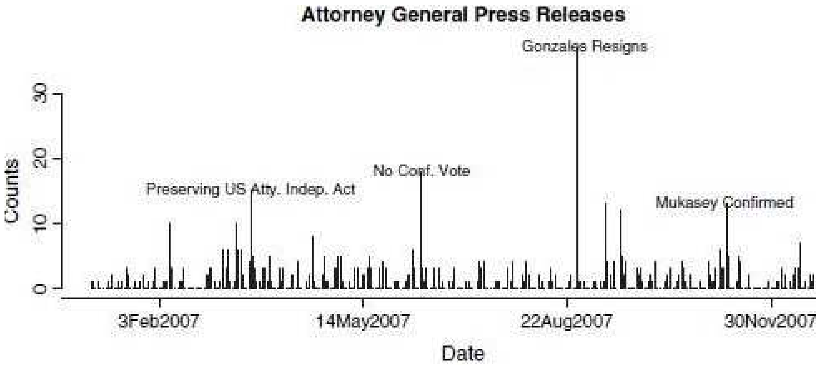


Figure: From Grimmer 2010; press releases

Code (1): Loading Textual Data

How to do this at home?

- Most computer scientists use Perl, Python, other programming languages

Code (1): Loading Textual Data

How to do this at home?

- Most computer scientists use Perl, Python, other programming languages
- R $\rightarrow$ increasing tools for automated content analysis

Code (1): Loading Textual Data

How to do this at home?

- Most computer scientists use Perl, Python, other programming languages
- R $\rightarrow$ increasing tools for automated content analysis
- e.g. tm, ReadMe

Code (1): Loading Textual Data

How to do this at home?

- Most computer scientists use Perl, Python, other programming languages
- R $\rightarrow$ increasing tools for automated content analysis
- e.g. tm, ReadMe
- Commercial software $\rightarrow$ increasing tools as well (e.g. Clementine for Stata)

Loading Textual Data

- Example here: from ReadMe

Loading Textual Data

- Example here: from ReadMe
- Must specify control file telling ReadMe where documents are

Loading Textual Data

- Example here: from ReadMe
- Must specify control file telling ReadMe where documents are
- Control file: lists each document location, category, whether it is in training set (vs. test set)

Control File

C:/Users/114715-berk.txt,None,1
C:/Users/62815-berk.txt,None,1
C:/Users/118871-berk.txt,California,1
C:/Users/106588-berk.txt,California,1
C:/Users/122973-berk.txt,None,1
C:/Users/106590-berk.txt,California,1
C:/Users/54635-berk.txt,Regulation,1
C:/Users/136556-berk.txt,Regulation-Politics,1

Input Command

- Input data from R using following function

Input Command

- Input data from R using following function
- setwd("C:/Users/Dan/")

Input Command

- Input data from R using following function
- setwd("C:/Users/Dan/")
- library(ReadMe)

Input Command

- Input data from R using following function
- setwd("C:/Users/Dan/")
- library(ReadMe)
- underg ← undergrad(control="C:/Users/Dan/control1.txt",sep=",")

Input Command

- Input data from R using following function
- setwd("C:/Users/Dan/")
- library(ReadMe)
- underg ← undergrad(control="C:/Users/Dan/control1.txt",sep=",")
- Need to use fullfreq=T argument to get number of words (not occurrence)

Code (2): Estimating an SVM

- library(e1071)

Code (2): Estimating an SVM

- library(e1071)
- svout ← svm(as.factor(TRUTH2) ., data=underg2$trainingset2,cross=5,probability=T,kernel=”ra

Code (2): Estimating an SVM

- library(e1071)
- svout ← svm(as.factor(TRUTH2) ., data=underg2$trainingset2,cross=5,probability=T,kernel="ra
- p1 ← predict(svout,newdata=underg2$testset2,probability=T)

Code (2): Estimating an SVM

- library(e1071)
- svout ← svm(as.factor(TRUTH2) ., data=underg2$trainingset2,cross=5,probability=T,kernel="ra
- p1 ← predict(svout,newdata=underg2$testset2,probability=T)
- table(underg2$testset2$TRUTH2,$p1 > .5$)