

Advanced Quantitative Research Methodology, Lecture

Notes: Text Analysis: Supervised Learning

Gary King
Institute for Quantitative Social Science
Harvard University

April 22, 2012

- Daniel Hopkins and Gary King. “Extracting Systematic Social Science Meaning from Text” *American Journal of Political Science*, ~→ commercialized via:



- Gary King and Ying Lu. “Verbal Autopsy Methods with Multiple Causes of Death,” *Statistical Science*
- Copies at <http://gking.harvard.edu>

Inputs and Target Quantities of Interest

- Input Data:
 - Large set of text documents (blogs, web pages, emails, etc.)
 - A set of (mutually exclusive and exhaustive) categories
 - A small set of documents hand-coded into the categories
- Quantities of interest
 - individual document classifications (spam filters)
 - proportion in each category (proportion email which is spam)
- Estimation
 - Can get the 2nd by counting the 1st (turns out not to be necessary!)
 - High classification accuracy $\nRightarrow$ unbiased category proportions
 - $\Rightarrow$ Different methods optimize estimation of the different quantities

One specific quantity of interest

- Daily opinion about President Bush and 2008 candidates among all English language blog posts

- Specific categories:

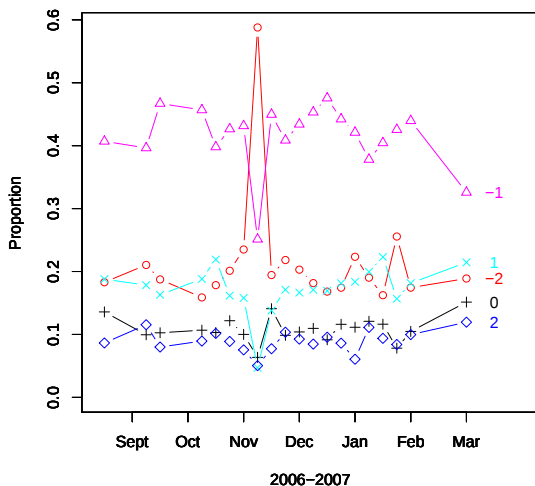
<u>Label</u>	<u>Category</u>
-2	extremely negative
-1	negative
0	neutral
1	positive
2	extremely positive
NA	no opinion expressed
NB	not a blog

- Hard case:
 - Part ordinal, part nominal categorization
 - “Sentiment categorization is more difficult than topic classification”
 - Informal language: “my crunchy gf thinks dubya hid the wmd's, :)!”
 - Little common internal structure (no inverted pyramid)

The Conversation about John Kerry's Botched Joke

You know, education — if you make the most of it . . . you can do well. If you don't, you get stuck in Iraq.

Affect Towards John Kerry



Representing Text as Numbers

- **Filter**: choose English language blogs that mention Bush
- **Preprocess**: convert to lower case, remove punctuation, keep only word stems (“consist”, “consisted”, “consistency” $\rightsquigarrow$ “consist”)
- **Code variables**: presence/absence of unique unigrams, bigrams, trigrams
- **Our Example**:
 - Our 10,771 blog posts about Bush and Clinton: 201,676 unigrams, 2,392,027 bigrams, 5,761,979 trigrams.
 - **keep only** unigrams in $> 1\%$ or $< 99\%$ of documents: 3,672 variables
 - Groups infinite possible posts into “only” $2^{3,672}$ distinct types
- **More sophisticated summaries**: we’ve used, but they’re not necessary

- Document Category

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

- Word Stem Profile:

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

Quantities of Interest

- Computer Science: individual document **classifications**

$$D_1, D_2, \dots, D_L$$

- Social Science: **proportions** in each category

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

Issues with Existing Statistical Approaches

① Direct Sampling

- Biased without a random sample
- nonrandomness common due to population drift, data subdivisions, etc.
- (Classification of population documents not necessary)

② Aggregation of model-based individual classifications

- Biased without a random sample
- Models $P(D|\mathbf{S})$, but the world works as $P(\mathbf{S}|D)$
- Bias unless
 - $P(D|\mathbf{S})$ encompasses the “true” model.
 - $\mathbf{S}$ spans the space of all predictors of D (i.e., all information in the document)
- Bias even with optimal classification and high % correctly classified

Using Misclassification Rates to Correct Proportions

- Use some method to **classify unlabeled documents**
- **Aggregate classifications** to category proportions
- Use labeled set to **estimate misclassification rates** (by cross-validation)
- **Use misclassification rates to correct proportions**
- **Result:** vastly improved estimates of category proportions
- (No new assumptions beyond that of the classifier)
- (still requires random samples, individual classification, etc)

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 $P(\hat{D} = 1)$

Generalizations: J Categories, No Individual Classification

(King and Lu, 2008, in press)

- 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 $\hat{D} = f(\mathbf{S})$:

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

- Simplify to an equivalent matrix expression:

$$P(\mathbf{S}) = P(\mathbf{S} | D) P(D)$$

Estimation

The matrix expression again:

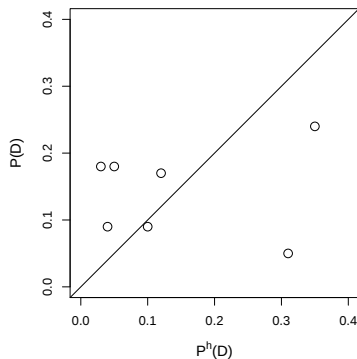
$$\begin{array}{ccc} P(\mathbf{S}) & = & P(\mathbf{S}|D)P(D) \\ 2^K \times 1 & & 2^K \times J \quad J \times 1 \end{array}$$
$$\implies Y = X\beta \implies \beta = (X'X)^{-1}X'y$$

Document category proportions (quantity of interest) Word stem profile proportions (estimate in unlabeled set by tabulation) Word stem profiles, by category (estimate in *labeled* set by tabulation) Alternative symbols (to emphasize the linear equation) Solve for quantity of interest (with no error term)

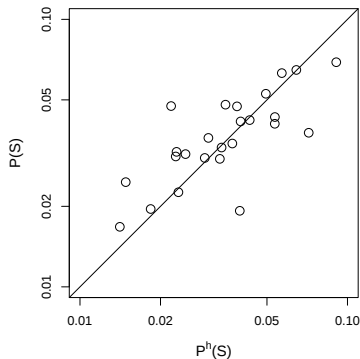
- Technical estimation issues:
 - 2^K is enormous, far larger than any existing computer
 - $P(\mathbf{S})$ and $P(\mathbf{S}|D)$ will be too sparse
 - Elements of $P(D)$ must be between 0 and 1 and sum to 1
- Solutions

A Nonrandom Hand-coded Sample

**Differences in Document
Category Frequencies**

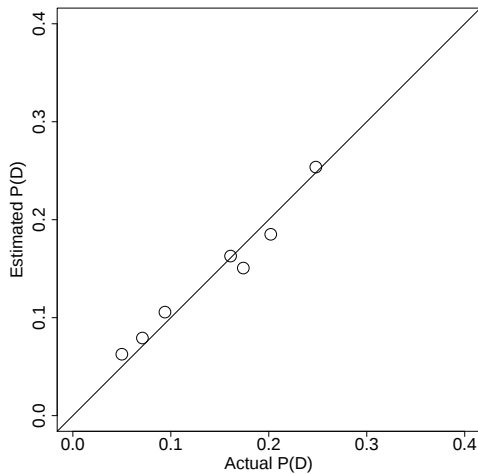


**Differences in Word
Profile Frequencies**

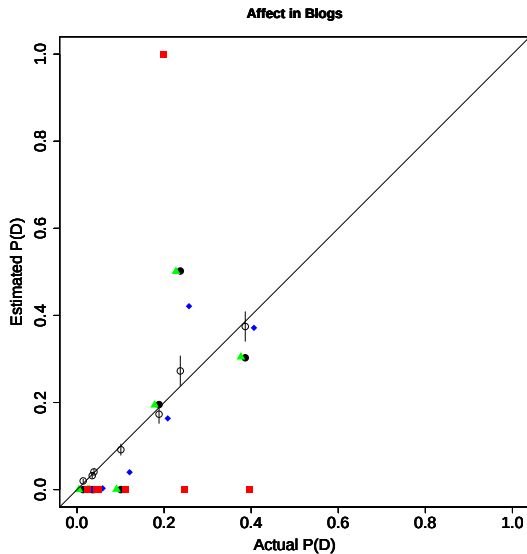


All existing methods would fail with these data.

Accurate Estimates

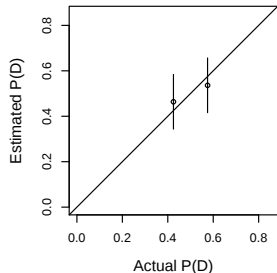


Out-of-sample Comparison: 60 Seconds vs. 8.7 Days

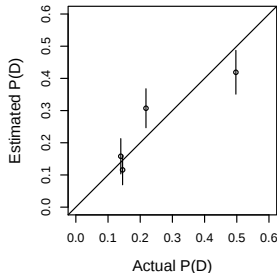


Out of Sample Validation: Other Examples

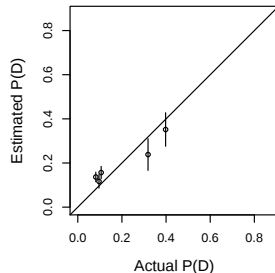
Congressional Speeches



Immigration Editorials



Enron Emails



- The Problem

- Policymakers need the **cause-specific mortality rate** to set research goals, budgetary priorities, and ameliorative policies
- High quality death registration: only 23/192 countries

- Existing Approaches

- Verbal Autopsy: Ask relatives or caregivers 50-100 symptom questions
- Ask physicians to determine cause of death (low intercoder reliability)
- Apply expert algorithms (high reliability, low validity)
- Find deaths with medically certified causes from a local hospital, trace caregivers to their homes, ask the same symptom questions, and statistically classify deaths in population (model-dependent, low accuracy)

An Alternative Approach

- Document ~~Category~~, Cause of ~~Death~~,

$$D_i = \begin{cases} 1 & \text{if bladder cancer} \\ 2 & \text{if cardiovascular disease} \\ 3 & \text{if transportation accident} \\ \vdots & \vdots \\ J & \text{if infectious respiratory} \end{cases}$$

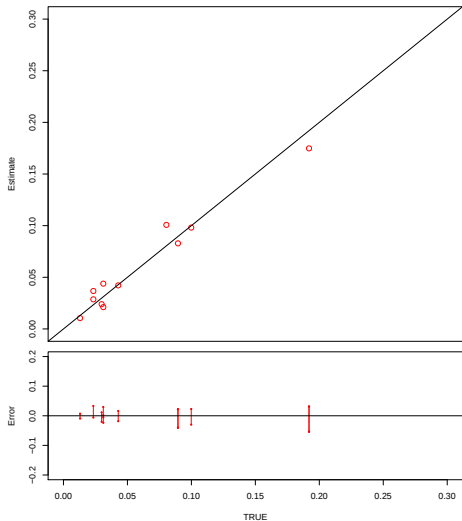
- Word ~~Stem~~ Profile, ~~Symptoms~~:

$$S_i = \begin{cases} S_{i1} = 1 & \text{if "breathing difficulties", 0 if not} \\ S_{i2} = 1 & \text{if "stomach ache", 0 if not} \\ \vdots & \vdots \\ S_{iK} = 1 & \text{if "diarrhea", 0 if not} \end{cases}$$

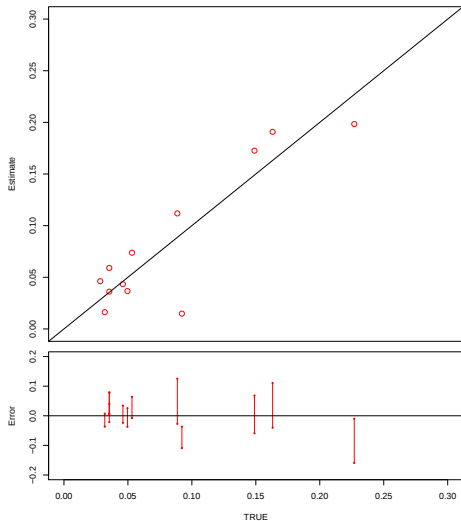
- Apply the ~~same~~ methods

Validation in Tanzania

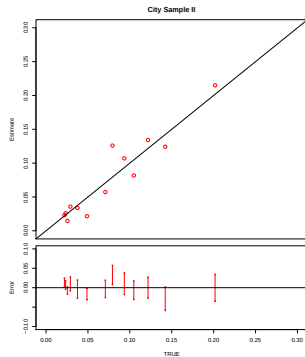
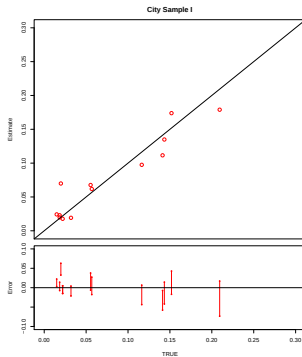
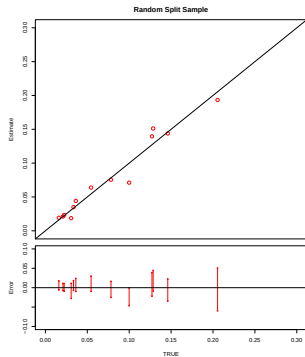
Random Split Sample



Community Sample



Validation in China



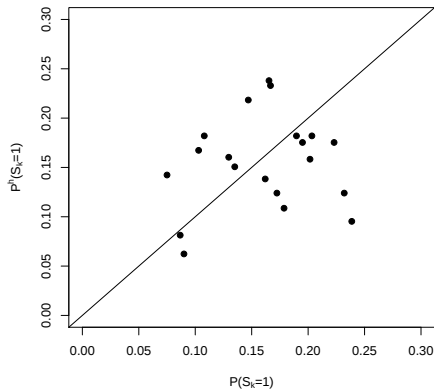
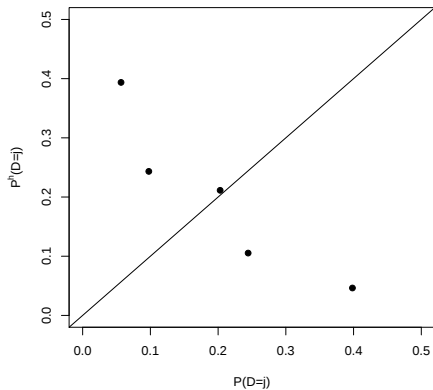
Implications for an Individual Classifier

- All existing classifiers assume: $P^h(S, D) = P(S, D)$
- For a different quantity we assume: $P^h(S|D) = P(S|D)$
- How to use this (less restrictive) assumption for classification (Bayes Theorem):

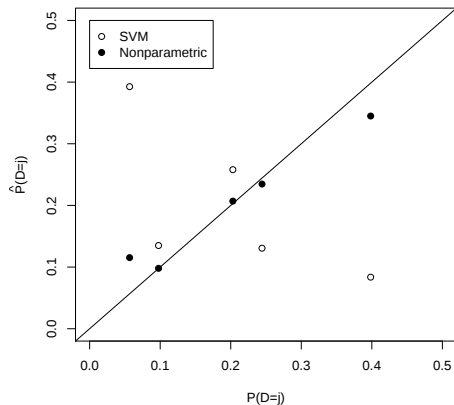
$$P(D_\ell | \mathbf{S}_\ell = \mathbf{s}_\ell) = \frac{P(\mathbf{S}_\ell = \mathbf{s}_\ell | D_\ell = j)P(D_\ell = j)}{P(\mathbf{S}_\ell = \mathbf{s}_\ell)}$$

The goal: individual classification Output from our estimator (described above)
Nonparametric estimate from labeled set (an assumption)
Nonparametric estimate from unlabeled set (no assumption)

Classification with Less Restrictive Assumptions



Classification with Less Restrictive Assumptions



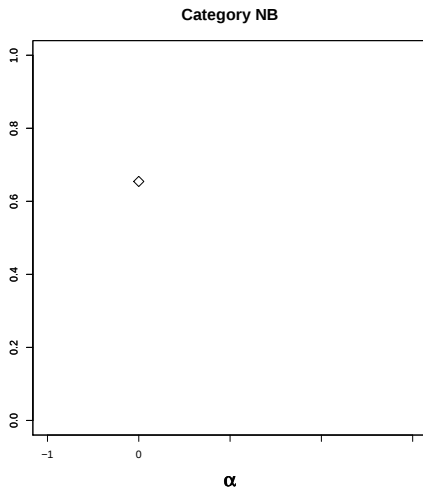
Percent correctly classified:

- SVM (best existing classifier): 40.5%
- Our nonparametric approach: 59.8%

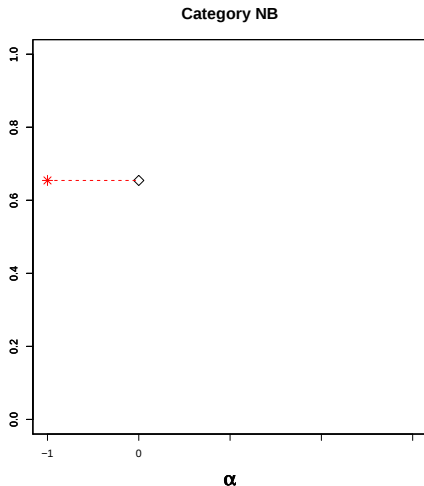
Misclassification Matrix for Blog Posts

	-2	-1	0	1	2	NA	NB	$P(D_1)$
-2	.70	.10	.01	.01	.00	.02	.16	.28
-1	.33	.25	.04	.02	.01	.01	.35	.08
0	.13	.17	.13	.11	.05	.02	.40	.02
1	.07	.06	.08	.20	.25	.01	.34	.03
2	.03	.03	.03	.22	.43	.01	.25	.03
NA	.04	.01	.00	.00	.00	.81	.14	.12
NB	.10	.07	.02	.02	.02	.04	.75	.45

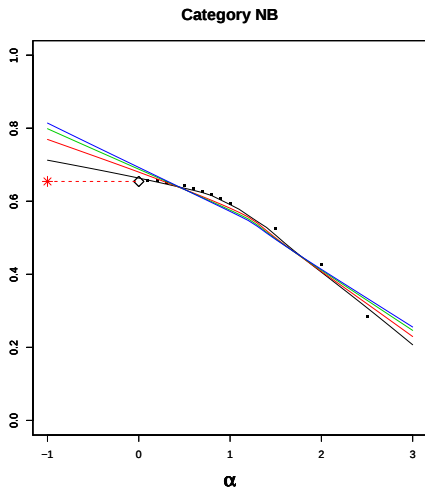
SIMEX Analysis of “Not a Blog” Category



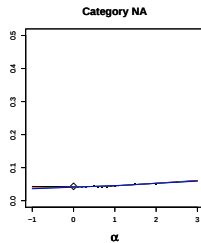
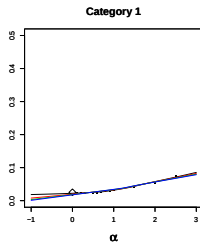
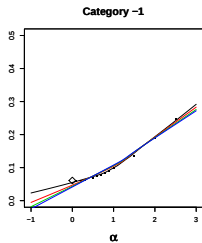
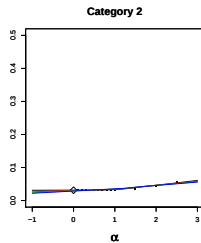
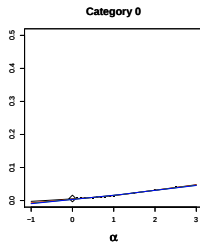
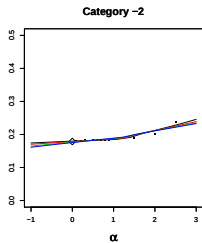
SIMEX Analysis of “Not a Blog” Category



SIMEX Analysis of “Not a Blog” Category



SIMEX Analysis of Other Categories



For more information

<http://GKing.Harvard.edu>