8. Unsupervised Techniques II: flipped

DS-GA 1015, Text as Data
Arthur Spirling

March 30, 2021

Housekeeping

Housekeeping

1 HW2 coming in today, March 30, 2021, at 11pm.

Housekeeping

1. HW2 coming in today, March 30, 2021, at 11pm.
2. Will post some general advice on the final paper.

Political Speech: US Senate

Political Speech: US Senate

Beauchamp, 2010 (Text-Based Scaling of Legislatures: A Comparison of Methods with Applications to the US Senate and UK House of Commons)

Political Speech: US Senate

Beauchamp, 2010 (Text-Based Scaling of Legislatures: A Comparison of Methods with Applications to the US Senate and UK House of Commons)

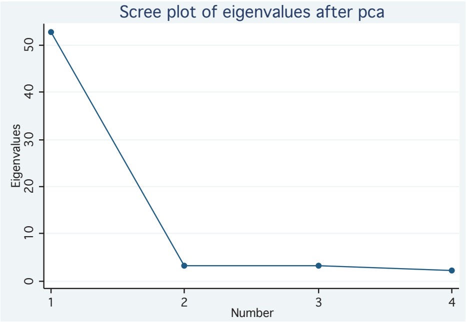
Considers PCA of (preprocessed) 1000-top-vectors for US Senators.

Political Speech: US Senate

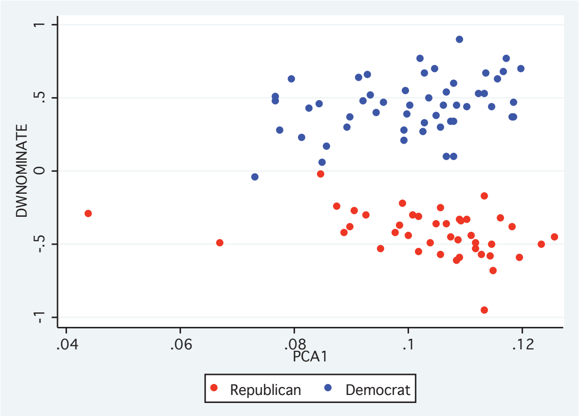
Beauchamp, 2010 (Text-Based Scaling of Legislatures: A Comparison of Methods with Applications to the US Senate and UK House of Commons)

Considers PCA of (preprocessed) 1000-top-vectors for US Senators.

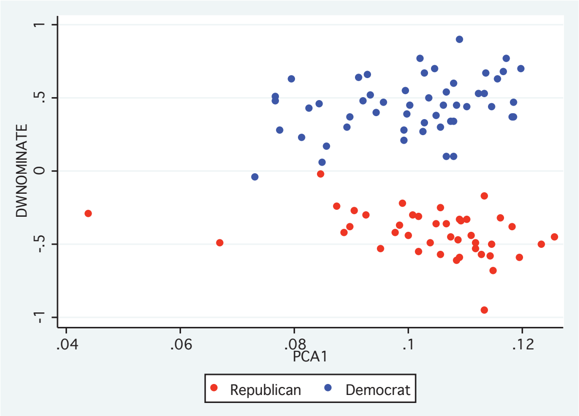
Fits several components, of which 1PC model looks very good. . .



Exercise

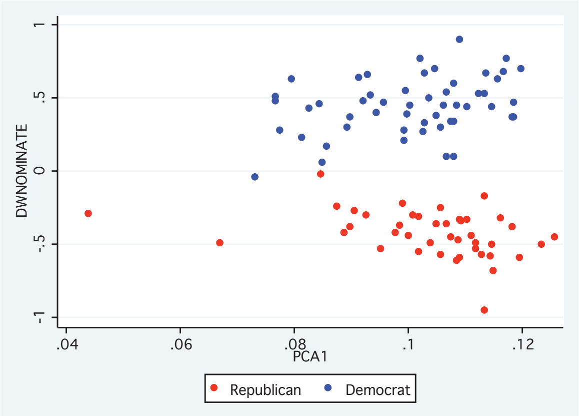


Exercise



Strangely, in Beauchamp's work, PC1 uncorrelated with first dimension of roll calls scores.

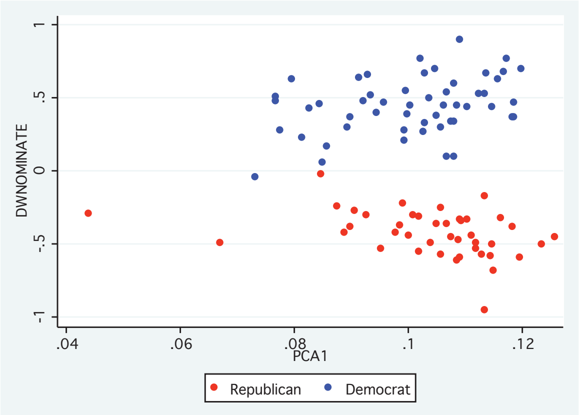
Exercise



Strangely, in Beauchamp's work, PC1 uncorrelated with first dimension of roll calls scores.

why?

Exercise



Strangely, in Beauchamp's work, PC1 uncorrelated with first dimension of roll calls scores.

why?

Clustering

Clustering

Clustering:

Clustering

Clustering: look for ‘groups’ in data explicitly.

Clustering

Clustering: look for 'groups' in data explicitly.

Partition methods are most common:

Clustering

Clustering: look for ‘groups’ in data explicitly.

Partition methods are most common:

→ Include K-means, for which one pre-specifies cluster number,

Clustering

Clustering: look for 'groups' in data explicitly.

Partition methods are most common:

→ Include K-means, for which one pre-specifies cluster number, and algorithm puts observations into clusters which minimize within cluster sum of squares

Clustering

Clustering: look for 'groups' in data explicitly.

Partition methods are most common:

→ Include K-means, for which one pre-specifies cluster number, and algorithm puts observations into clusters which minimize within cluster sum of squares

Clustering

Clustering: look for 'groups' in data explicitly.

Partition methods are most common:

→ Include K-means, for which one pre-specifies cluster number, and algorithm puts observations into clusters which minimize within cluster sum of squares

so pick s such that $\sum_{i=1}^{k} \sum_{\mathbf{x_j} \in S_i} ||\mathbf{x_j} - \boldsymbol{\mu_i}||^2$ is minimized where $\boldsymbol{\mu_i}$ is (vector) mean of points in cluster S_i.

Clustering

Clustering: look for 'groups' in data explicitly.

Partition methods are most common:

→ Include K-means, for which one pre-specifies cluster number, and algorithm puts observations into clusters which minimize within cluster sum of squares

so pick s such that $\sum_{i=1}^{k} \sum_{\mathbf{x_j} \in S_i} ||\mathbf{x_j} - \boldsymbol{\mu_i}||^2$ is minimized where $\boldsymbol{\mu_i}$ is (vector) mean of points in cluster S_i.

→ observations (documents) within clusters should be as similar as possible,

Clustering

Clustering: look for 'groups' in data explicitly.

Partition methods are most common:

→ Include K-means, for which one pre-specifies cluster number, and algorithm puts observations into clusters which minimize within cluster sum of squares

so pick s such that $\sum_{i=1}^{k} \sum_{\mathbf{x_j} \in S_i} ||\mathbf{x_j} - \boldsymbol{\mu_i}||^2$ is minimized where $\boldsymbol{\mu_i}$ is (vector) mean of points in cluster S_i.

→ observations (documents) within clusters should be as similar as possible, observations (documents) in different clusters should be as different as possible.

Exercise

Exercise

Suppose you were modeling schoolchildren in terms of intellectual ability. Would you model their data (correct/incorrect) on the test they took using PCA or clustering? Why?

Exercise

Suppose you were modeling schoolchildren in terms of intellectual ability. Would you model their data (correct/incorrect) on the test they took using PCA or clustering? Why?

Suppose you were modeling students organizing friendship groups based on who they know/see on a regular basis. Would you model this data using PCA or clustering? Why?

Item Response

Introduction

Introduction

- interest is in measuring some underlying, latent (unobservable) trait like ‘ability’ or ‘intelligence’

Introduction

- interest is in measuring some underlying, latent (unobservable) trait like ‘ability’ or ‘intelligence’
- generally seek to place students on arbitrary scale, with midpoint zero, measurement unit of one, bounds $(-\infty, \infty)$

Introduction

- interest is in measuring some underlying, latent (unobservable) trait like ‘ability’ or ‘intelligence’
- generally seek to place students on arbitrary scale, with midpoint zero, measurement unit of one, bounds $(-\infty, \infty)$
- students respond to items and get them ‘wrong’ or ‘correct’ $\{0, 1\}$

Introduction

- interest is in measuring some underlying, latent (unobservable) trait like ‘ability’ or ‘intelligence’
- generally seek to place students on arbitrary scale, with midpoint zero, measurement unit of one, bounds $(-\infty, \infty)$
- students respond to items and get them ‘wrong’ or ‘correct’ $\{0, 1\}$
- classical test theory says that score is (raw) sum of correct answers

Introduction

- interest is in measuring some underlying, latent (unobservable) trait like ‘ability’ or ‘intelligence’
- generally seek to place students on arbitrary scale, with midpoint zero, measurement unit of one, bounds $(-\infty, \infty)$
- students respond to items and get them ‘wrong’ or ‘correct’ $\{0, 1\}$
- classical test theory says that score is (raw) sum of correct answers
- item response theory pays attention to whether students got particular items correct,

Introduction

- interest is in measuring some underlying, latent (unobservable) trait like ‘ability’ or ‘intelligence’
- generally seek to place students on arbitrary scale, with midpoint zero, measurement unit of one, bounds $(-\infty, \infty)$
- students respond to items and get them ‘wrong’ or ‘correct’ $\{0, 1\}$
- classical test theory says that score is (raw) sum of correct answers
- item response theory pays attention to whether students got particular items correct, with particular characteristics

Introduction

- interest is in measuring some underlying, latent (unobservable) trait like ‘ability’ or ‘intelligence’
- generally seek to place students on arbitrary scale, with midpoint zero, measurement unit of one, bounds $(-\infty, \infty)$
- students respond to items and get them ‘wrong’ or ‘correct’ $\{0, 1\}$
- classical test theory says that score is (raw) sum of correct answers
- item response theory pays attention to whether students got particular items correct, with particular characteristics
- underlying ability of a student is θ,

Introduction

- interest is in measuring some underlying, latent (unobservable) trait like ‘ability’ or ‘intelligence’
- generally seek to place students on arbitrary scale, with midpoint zero, measurement unit of one, bounds $(-\infty, \infty)$
- students respond to items and get them ‘wrong’ or ‘correct’ $\{0, 1\}$
- classical test theory says that score is (raw) sum of correct answers
- item response theory pays attention to whether students got particular items correct, with particular characteristics
- underlying ability of a student is θ, and the probability he gets particular item correct is simply $p(\theta)$:

Introduction

- interest is in measuring some underlying, latent (unobservable) trait like ‘ability’ or ‘intelligence’
- generally seek to place students on arbitrary scale, with midpoint zero, measurement unit of one, bounds $(-\infty, \infty)$
- students respond to items and get them ‘wrong’ or ‘correct’ $\{0, 1\}$
- classical test theory says that score is (raw) sum of correct answers
- item response theory pays attention to whether students got particular items correct, with particular characteristics
- underlying ability of a student is θ, and the probability he gets particular item correct is simply $p(\theta)$: i.e. a function only of his ability.

Item Characteristic Curve

Item Characteristic Curve

for a given item. . .

Item Characteristic Curve

for a given item...

- the probability that student gets it right is ‘high’ when θ is high

Item Characteristic Curve

for a given item...

- the probability that student gets it right is 'high' when θ is high
- and low when θ is low

Item Characteristic Curve

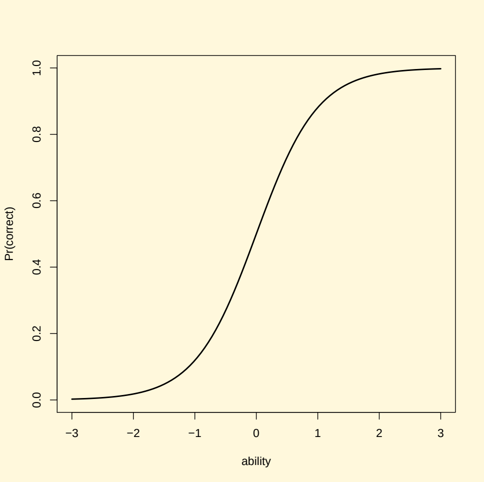
for a given item...

- the probability that student gets it right is 'high' when θ is high
- and low when θ is low
- in particular, we assume something like

Item Characteristic Curve

for a given item. . .

- the probability that student gets it right is ‘high’ when θ is high
- and low when θ is low
- in particular, we assume something like



Properties of IC curve I

Properties of IC curve I

difficulty of item

Properties of IC curve I

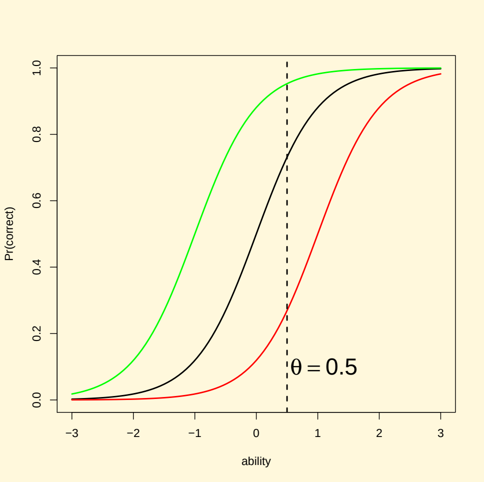
difficulty of item

- tells us about location

Properties of IC curve I

difficulty of item

- tells us about location
- easy items are ones where even those with low θ can get correct
- hard items are ones where only those with high θ can get correct
- can think about a particular individual with e.g. $\theta = 0.5$
- item difficulty can be given as θ for which $\Pr(\text{correct}) = 0.5$



Properties of IC curve I

Properties of IC curve I

discrimination of item (steepness)

Properties of IC curve I

discrimination of item (steepness)

- tells us about differentiation between students whose abilities lie above and below item location (e.g. zero)

Properties of IC curve I

discrimination of item (steepness)

- tells us about differentiation between students whose abilities lie above and below item location (e.g. zero)
- highly discriminating items are ones where those with high θ are much more likely to be correct than those with low θ

Properties of IC curve I

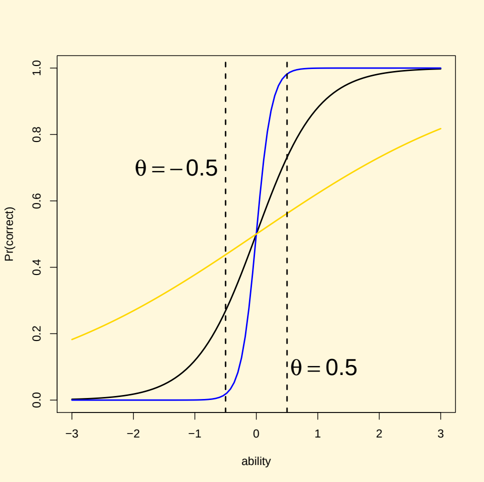
discrimination of item (steepness)

- tells us about differentiation between students whose abilities lie above and below item location (e.g. zero)
- highly discriminating items are ones where those with high θ are much more likely to be correct than those with low θ
- less discriminating items are ones where those with high θ are only just more likely to be correct than those with low θ

Properties of IC curve I

discrimination of item (steepness)

- tells us about differentiation between students whose abilities lie above and below item location (e.g. zero)
- highly discriminating items are ones where those with high θ are much more likely to be correct than those with low θ
- less discriminating items are ones where those with high θ are only just more likely to be correct than those with low θ
- contrast e.g. $\theta = 0.5$ student to $\theta = -0.5$ student



Modeling the IC Curve

- suppose we are interested in probability of a correct response for a given item

Modeling the IC Curve

- suppose we are interested in probability of a correct response for a given item
- logit is very popular,

Modeling the IC Curve

- suppose we are interested in probability of a correct response for a given item
- logit is very popular, taking the form (2PL)

Modeling the IC Curve

- suppose we are interested in probability of a correct response for a given item
- logit is very popular, taking the form (2PL)

$$p(\theta) = \frac{1}{1 + e^{-a(\theta - b)}} = F(a\theta - ab)$$

Modeling the IC Curve

- suppose we are interested in probability of a correct response for a given item
- logit is very popular, taking the form (2PL)

$$p(\theta) = \frac{1}{1 + e^{-a(\theta - b)}} = F(a\theta - ab)$$

where

- b is difficulty parameter,

Modeling the IC Curve

- suppose we are interested in probability of a correct response for a given item
- logit is very popular, taking the form (2PL)

$$p(\theta) = \frac{1}{1 + e^{-a(\theta - b)}} = F(a\theta - ab)$$

where

- b is difficulty parameter, and simply θ value for which $\Pr(\mathrm{correct}) = 0.5$

Modeling the IC Curve

- suppose we are interested in probability of a correct response for a given item
- logit is very popular, taking the form (2PL)

$$p(\theta) = \frac{1}{1 + e^{-a(\theta - b)}} = F(a\theta - ab)$$

where

- b is difficulty parameter, and simply θ value for which $\Pr(\text{correct}) = 0.5$
- a is discrimination parameter

Modeling the IC Curve

- suppose we are interested in probability of a correct response for a given item
- logit is very popular, taking the form (2PL)

$$p(\theta) = \frac{1}{1 + e^{-a(\theta - b)}} = F(a\theta - ab)$$

where

- b is difficulty parameter, and simply θ value for which $\Pr(\text{correct}) = 0.5$
- a is discrimination parameter: in practice, slope is not constant (this gets complicated),

Modeling the IC Curve

- suppose we are interested in probability of a correct response for a given item
- logit is very popular, taking the form (2PL)

$$p(\theta) = \frac{1}{1 + e^{-a(\theta - b)}} = F(a\theta - ab)$$

where

- b is difficulty parameter, and simply θ value for which $\Pr(\text{correct}) = 0.5$
- a is discrimination parameter: in practice, slope is not constant (this gets complicated), but think of this as the curve slope at $\theta = b$

Modeling the IC Curve

- suppose we are interested in probability of a correct response for a given item
- logit is very popular, taking the form (2PL)

$$p(\theta) = \frac{1}{1 + e^{-a(\theta - b)}} = F(a\theta - ab)$$

where

- b is difficulty parameter, and simply θ value for which $\Pr(\text{correct}) = 0.5$
- a is discrimination parameter: in practice, slope is not constant (this gets complicated), but think of this as the curve slope at $\theta = b$
- θ is ability

Modeling the IC Curve

- suppose we are interested in probability of a correct response for a given item
- logit is very popular, taking the form (2PL)

$$p(\theta) = \frac{1}{1 + e^{-a(\theta - b)}} = F(a\theta - ab)$$

where

- b is difficulty parameter, and simply θ value for which $\Pr(\mathrm{correct}) = 0.5$
- a is discrimination parameter: in practice, slope is not constant (this gets complicated), but think of this as the curve slope at $\theta = b$
- θ is ability

Kim, Ratkovic & Londregan (2018)

Attempt to combine roll call vote model (the 2PL), with word choice model.

Kim, Ratkovic & Londregan (2018)

Attempt to combine roll call vote model (the 2PL), with word choice model. Introduce sparse factor analysis.

Kim, Ratkovic & Londregan (2018)

Attempt to combine roll call vote model (the 2PL), with word choice model. Introduce sparse factor analysis.

Assume legislators choose votes and words;

Kim, Ratkovic & Londregan (2018)

Attempt to combine roll call vote model (the 2PL), with word choice model. Introduce sparse factor analysis.

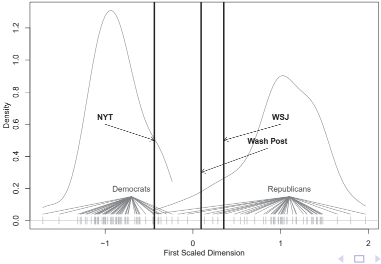
Assume legislators choose votes and words; allow these two factors to influence ideal points with different weights that will be estimated from data (basically, whatever discriminates between individuals best)

Kim, Ratkovic & Londregan (2018)

Attempt to combine roll call vote model (the 2PL), with word choice model. Introduce sparse factor analysis.

Assume legislators choose votes and words; allow these two factors to influence ideal points with different weights that will be estimated from data (basically, whatever discriminates between individuals best)

Incorporation of words allows one to place e.g. newspapers in same space as legislators:



Latent Semantic Analysis

Process

Process

Not supervised,

Process

Not supervised, in sense that there is no *a priori* dictionary, or grammar, or morphology etc

Process

Not supervised, in sense that there is no *a priori* dictionary, or grammar, or morphology etc

Begin with (unstemmed, unstopped) TDM for some set of texts:

Process

Not supervised, in sense that there is no *a priori* dictionary, or grammar, or morphology etc

Begin with (unstemmed, unstopped) TDM for some set of texts: terms are rows, texts are columns.

Process

Not supervised, in sense that there is no *a priori* dictionary, or grammar, or morphology etc

Begin with (unstemmed, unstopped) TDM for some set of texts: terms are rows, texts are columns. Cell entries are transformed via product of. . .

Process

Not supervised, in sense that there is no *a priori* dictionary, or grammar, or morphology etc

Begin with (unstemmed, unstopped) TDM for some set of texts: terms are rows, texts are columns. Cell entries are transformed via product of...

local weight function:

Process

Not supervised, in sense that there is no *a priori* dictionary, or grammar, or morphology etc

Begin with (unstemmed, unstopped) TDM for some set of texts: terms are rows, texts are columns. Cell entries are transformed via product of...

local weight function: e.g. $\log(tf_{ij} + 1)$

Process

Not supervised, in sense that there is no *a priori* dictionary, or grammar, or morphology etc

Begin with (unstemmed, unstopped) TDM for some set of texts: terms are rows, texts are columns. Cell entries are transformed via product of. . .

local weight function: e.g. $\log(tf_{ij} + 1)$
(infinite variety)

Process

Not supervised, in sense that there is no *a priori* dictionary, or grammar, or morphology etc

Begin with (unstemmed, unstopped) TDM for some set of texts: terms are rows, texts are columns. Cell entries are transformed via product of. . .

- local weight function: e.g. $\log(tf_{ij} + 1)$ (infinite variety)
- global weight function: e.g.

Process

Not supervised, in sense that there is no *a priori* dictionary, or grammar, or morphology etc

Begin with (unstemmed, unstopped) TDM for some set of texts: terms are rows, texts are columns. Cell entries are transformed via product of. . .

local weight function: e.g. $\log(tf_{ij} + 1)$
(infinite variety)
global weight function: e.g. $1 + \sum_j \frac{p_{ij} \log p_{ij}}{\log n}$

Process

Not supervised, in sense that there is no *a priori* dictionary, or grammar, or morphology etc

Begin with (unstemmed, unstopped) TDM for some set of texts: terms are rows, texts are columns. Cell entries are transformed via product of. . .

local weight function: e.g. $\log(tf_{ij} + 1)$
(infinite variety)
global weight function: e.g. $1 + \sum_j \frac{p_{ij} \log p_{ij}}{\log n}$ (infinite variety:

Process

Not supervised, in sense that there is no *a priori* dictionary, or grammar, or morphology etc

Begin with (unstemmed, unstopped) TDM for some set of texts: terms are rows, texts are columns. Cell entries are transformed via product of...

local weight function: e.g. $\log(tf_{ij} + 1)$
(infinite variety)
global weight function: e.g. $1 + \sum_j \frac{p_{ij} \log p_{ij}}{\log n}$ (infinite variety: this one is '1 + entropy' in some lits)

Process

Not supervised, in sense that there is no *a priori* dictionary, or grammar, or morphology etc

Begin with (unstemmed, unstopped) TDM for some set of texts: terms are rows, texts are columns. Cell entries are transformed via product of...

- local weight function: e.g. $\log(tf_{ij} + 1)$ (infinite variety)
- global weight function: e.g. $1 + \sum_j \frac{p_{ij} \log p_{ij}}{\log n}$ (infinite variety: this one is '1 + entropy' in some lits)

NB take $0 \times \log(0)$ to be zero

Process

Not supervised, in sense that there is no *a priori* dictionary, or grammar, or morphology etc

Begin with (unstemmed, unstopped) TDM for some set of texts: terms are rows, texts are columns. Cell entries are transformed via product of. . .

local weight function: e.g. $\log(tf_{ij} + 1)$
(infinite variety)

global weight function: e.g. $1 + \sum_j \frac{p_{ij} \log p_{ij}}{\log n}$ (infinite variety: this one is '1 + entropy' in some lits)

NB take $0 \times \log(0)$ to be zero
where

Process

Not supervised, in sense that there is no *a priori* dictionary, or grammar, or morphology etc

Begin with (unstemmed, unstopped) TDM for some set of texts: terms are rows, texts are columns. Cell entries are transformed via product of...

local weight function: e.g. $\log(tf_{ij} + 1)$
(infinite variety)
global weight function: e.g. $1 + \sum_j \frac{p_{ij} \log p_{ij}}{\log n}$ (infinite variety: this one is '$1 +$ entropy' in some lits)

NB take $0 \times \log(0)$ to be zero
where n is simply total number of documents;

Process

Not supervised, in sense that there is no *a priori* dictionary, or grammar, or morphology etc

Begin with (unstemmed, unstopped) TDM for some set of texts: terms are rows, texts are columns. Cell entries are transformed via product of...

local weight function: e.g. $\log(tf_{ij} + 1)$
(infinite variety)
global weight function: e.g. $1 + \sum_j \frac{p_{ij} \log p_{ij}}{\log n}$ (infinite variety: this one is '$1 +$ entropy' in some lits)

NB take $0 \times \log(0)$ to be zero
where n is simply total number of documents; $p_{ij} = \frac{tf_{ij}}{gf_i}$.

Process

Not supervised, in sense that there is no *a priori* dictionary, or grammar, or morphology etc

Begin with (unstemmed, unstopped) TDM for some set of texts: terms are rows, texts are columns. Cell entries are transformed via product of. . .

> local weight function: e.g. $\log(tf_{ij} + 1)$ (infinite variety)
>
> global weight function: e.g. $1 + \sum_j \frac{p_{ij} \log p_{ij}}{\log n}$ (infinite variety: this one is '1 + entropy' in some lits)

NB take $0 \times \log(0)$ to be zero
where n is simply total number of documents; $p_{ij} = \frac{tf_{ij}}{gf_i}$. Note that gf_i is simply total number of times term i appears in corpus (i.e. over all docs).

Process

Not supervised, in sense that there is no *a priori* dictionary, or grammar, or morphology etc

Begin with (unstemmed, unstopped) TDM for some set of texts: terms are rows, texts are columns. Cell entries are transformed via product of. . .

local weight function: e.g. $\log(tf_{ij} + 1)$ (infinite variety)

global weight function: e.g. $1 + \sum_j \frac{p_{ij} \log p_{ij}}{\log n}$ (infinite variety: this one is '$1 +$ entropy' in some lits)

NB take $0 \times \log(0)$ to be zero
where n is simply total number of documents; $p_{ij} = \frac{tf_{ij}}{gf_i}$. Note that gf_i is simply total number of times term i appears in corpus (i.e. over all docs).

So. . .

So. . .

e.g. a row of some (3 text) corpus TDM is:

So. . .

e.g. a row of some (3 text) corpus TDM is:

term	doc1	doc2	doc3
dog	1	3	2

So. . .

e.g. a row of some (3 text) corpus TDM is:

term	doc1	doc2	doc3
dog	1	3	2

- applying LWF gives:

So. . .

e.g. a row of some (3 text) corpus TDM is:

term	doc1	doc2	doc3
dog	1	3	2

- applying LWF gives: `c(0.69, 1.39, 1.10)`

So. . .

e.g. a row of some (3 text) corpus TDM is:

term	doc1	doc2	doc3
dog	1	3	2

- applying LWF gives: `c(0.69, 1.39, 1.10)`
- applying GWF gives:

So...

e.g. a row of some (3 text) corpus TDM is:

term	doc1	doc2	doc3
dog	1	3	2

- applying LWF gives: `c(0.69, 1.39, 1.10)`
- applying GWF gives: $p_{i,1} = \frac{1}{6}$ and thus $1/6 \log(1/6)$; $p_{i,2} = \frac{3}{6}$ and thus $3/6 \log(3/6)$; $p_{i,3} = \frac{2}{6}$ and thus $2/6 \log(2/6)$.

So. . .

e.g. a row of some (3 text) corpus TDM is:

term	doc1	doc2	doc3
dog	1	3	2

- applying LWF gives: `c(0.69, 1.39, 1.10)`
- applying GWF gives: $p_{i,1} = \frac{1}{6}$ and thus $1/6 \log(1/6)$; $p_{i,2} = \frac{3}{6}$ and thus $3/6 \log(3/6)$; $p_{i,3} = \frac{2}{6}$ and thus $2/6 \log(2/6)$.
- we then have $1 + \left(\frac{1/6 \log(1/6) + 3/6 \log(3/6) + 2/6 \log(2/6)}{\log(3)} \right) = 0.079$

So...

e.g. a row of some (3 text) corpus TDM is:

term	doc1	doc2	doc3
dog	1	3	2

- applying LWF gives: `c(0.69, 1.39, 1.10)`
- applying GWF gives: $p_{i,1} = \frac{1}{6}$ and thus $1/6 \log(1/6)$; $p_{i,2} = \frac{3}{6}$ and thus $3/6 \log(3/6)$; $p_{i,3} = \frac{2}{6}$ and thus $2/6 \log(2/6)$.
- we then have $1 + \left(\frac{1/6 \log(1/6) + 3/6 \log(3/6) + 2/6 \log(2/6)}{\log(3)}\right) = 0.079$

... which we multiply by the LWF to give:

So. . .

e.g. a row of some (3 text) corpus TDM is:

term	doc1	doc2	doc3
dog	1	3	2

- applying LWF gives: `c(0.69, 1.39, 1.10)`
- applying GWF gives: $p_{i,1} = \frac{1}{6}$ and thus $1/6\log(1/6)$; $p_{i,2} = \frac{3}{6}$ and thus $3/6\log(3/6)$; $p_{i,3} = \frac{2}{6}$ and thus $2/6\log(2/6)$.
- we then have $1 + \left(\frac{1/6\log(1/6)+3/6\log(3/6)+2/6\log(2/6)}{\log(3)}\right) = 0.079$

. . . which we multiply by the LWF to give:

term	doc1	doc2	doc3
dog	0.055	0.110	0.087

An Example

An Example

An Example

79 (not many for LSA!)

An Example

79 (not many for LSA!) State of the Union Speeches,

An Example

79 (not many for LSA!) State of the Union Speeches, between 1934 and 2009:

An Example

79 (not many for LSA!) State of the Union Speeches, between 1934 and 2009: 16878 (stemmed, de-punctuated) terms

An Example

79 (not many for LSA!) State of the Union Speeches, between 1934 and 2009: 16878 (stemmed, de-punctuated) terms

Consider five dimensional representation

An Example

79 (not many for LSA!) State of the Union Speeches, between 1934 and 2009: 16878 (stemmed, de-punctuated) terms

Consider five dimensional representation (that is, reconstruct TDM as mix of five concepts

An Example

79 (not many for LSA!) State of the Union Speeches, between 1934 and 2009: 16878 (stemmed, de-punctuated) terms

Consider five dimensional representation (that is, reconstruct TDM as mix of five concepts—probably over-fitting)

An Example


79 (not many for LSA!) State of the Union Speeches, between 1934 and 2009: 16878 (stemmed, de-punctuated) terms

Consider five dimensional representation (that is, reconstruct TDM as mix of five concepts—probably over-fitting)

Q1 What are these documents about?

An Example

79 (not many for LSA!) State of the Union Speeches, between 1934 and 2009: 16878 (stemmed, de-punctuated) terms

Consider five dimensional representation (that is, reconstruct TDM as mix of five concepts—probably over-fitting)

Q1 What are these documents about? What do they have as their 'highest' (weighted) words?

An Example

79 (not many for LSA!) State of the Union Speeches, between 1934 and 2009: 16878 (stemmed, de-punctuated) terms

Consider five dimensional representation (that is, reconstruct TDM as mix of five concepts—probably over-fitting)

Q1 What are these documents about? What do they have as their 'highest' (weighted) words?

Q2 How are terms related? What words are closely associated conceptually?

Q1

Q1

	1942	1985	2002
original	war	freedom	america
	world	tax	security
	united	american	world
	people	time	american
	forces	growth	terror

Q1

	1942	1985	2002
original	war	freedom	america
	world	tax	security
	united	american	world
	people	time	american
	forces	growth	terror
transformed	1944	dollars	iraq
	japanese	tonight	iraqi
	war	we've	terrorists
	1942	million	qaida
	french	thats	terror
	germans	war	terrorist

Q2

Q2

Q2

words	original	transformed

Q2

words	original	transformed
communist, zarqawi	-0.08	-0.28

Q2

words	original	transformed
communist, zarqawi	-0.08	-0.28
america, freedom	0.06	0.41

Q2

words	original	transformed
communist, zarqawi	-0.08	-0.28
america, freedom	0.06	0.41
camp, david	0.56	0.57

Q2

words	original	transformed
communist, zarqawi	-0.08	-0.28
america, freedom	0.06	0.41
camp, david	0.56	0.57
inflation, unemployment	0.59	0.80

Q2

words	original	transformed
communist, zarqawi	-0.08	-0.28
america, freedom	0.06	0.41
camp, david	0.56	0.57
inflation, unemployment	0.59	0.80

Exercise

Exercise

words	original	transformed

Exercise

words	original	transformed
communist, zarqawi	-0.08	-0.28

Exercise

words	original	transformed
communist, zarqawi	-0.08	-0.28
america, freedom	0.06	0.41

Exercise

words	original	transformed
communist, zarqawi	-0.08	-0.28
america, freedom	0.06	0.41
camp, david	0.56	0.57

Exercise

words	original	transformed
communist, zarqawi	-0.08	-0.28
america, freedom	0.06	0.41
camp, david	0.56	0.57
inflation, unemployment	0.59	0.80

Exercise

words	original	transformed
communist, zarqawi	-0.08	-0.28
america, freedom	0.06	0.41
camp, david	0.56	0.57
inflation, unemployment	0.59	0.80

Exercise

words	original	transformed
communist, zarqawi	-0.08	-0.28
america, freedom	0.06	0.41
camp, david	0.56	0.57
inflation, unemployment	0.59	0.80

How do you interpret these transformed correlations? What do they suggest about the relevant concepts?

Wordfish

Parameters

Parameters

$$\boxed{\lambda_{ijt} = \exp\left(\alpha_{it} + \psi_j + \beta_j \times \omega_{it}\right)}$$

Parameters

$$\boxed{\lambda_{ijt} = \exp\left(\alpha_{it} + \psi_j + \beta_j \times \omega_{it}\right)}$$

α_{it} fixed effect(s) for party i in time t: why do we need this?

Parameters

$$\boxed{\lambda_{ijt} = \exp\left(\alpha_{it} + \psi_j + \beta_j \times \omega_{it}\right)}$$

α_{it} fixed effect(s) for party i in time t: why do we need this?

ψ_j word fixed effect: why do we need this?

Parameters

$$\boxed{\lambda_{ijt} = \exp\left(\alpha_{it} + \psi_j + \beta_j \times \omega_{it}\right)}$$

α_{it} fixed effect(s) for party i in time t: why do we need this?

ψ_j word fixed effect: why do we need this?

β_j word specific weight: what is this for?

Parameters

$$\boxed{\lambda_{ijt} = \exp\left(\alpha_{it} + \psi_j + \beta_j \times \omega_{it}\right)}$$

α_{it} fixed effect(s) for party i in time t: why do we need this?

ψ_j word fixed effect: why do we need this?

β_j word specific weight: what is this for?

ω_{it} : what is this for?

Recap

Recap

What role does the Poisson distribution play here?

Recap

What role does the Poisson distribution play here?

What is identification in this case? Why does it mean we need to do?

Recap

What role does the Poisson distribution play here?

What is identification in this case? Why does it mean we need to do?

What is the expectation maximization algorithm for in this case?

Recap

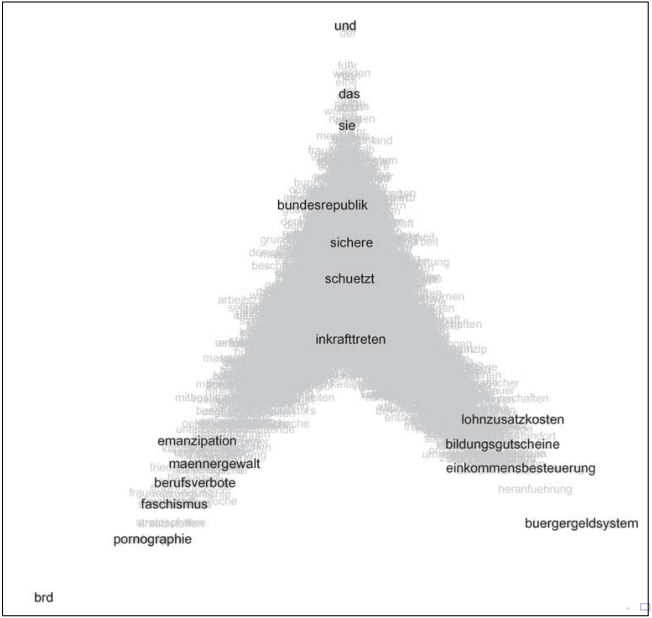
What role does the Poisson distribution play here?

What is identification in this case? Why does it mean we need to do?

What is the expectation maximization algorithm for in this case?
What other (Bayesian) ways could we use to proceed here?

‘Eiffel Tower’ plot

‘Eiffel Tower’ plot



this plot shape in common: why?
What is x and y?

Semi-supervised Techniques

Semi-Supervised Techniques

Semi-Supervised Techniques

May be prohibitively costly to provide enough labeled data for a supervised learning problem.

Semi-Supervised Techniques

May be prohibitively costly to provide enough labeled data for a supervised learning problem.

Turns out that accuracy of supervised text classifier can be (markedly) improved by adding large pool of unlabeled documents to a small number of training documents.

Semi-Supervised Techniques

May be prohibitively costly to provide enough labeled data for a supervised learning problem.

Turns out that accuracy of supervised text classifier can be (markedly) improved by adding large pool of unlabeled documents to a small number of training documents.

Intuition: unlabeled set provides useful information about joint probability over words.

Semi-Supervised Techniques

May be prohibitively costly to provide enough labeled data for a supervised learning problem.

Turns out that accuracy of supervised text classifier can be (markedly) improved by adding large pool of unlabeled documents to a small number of training documents.

Intuition: unlabeled set provides useful information about joint probability over words.

e.g. in labeled data,

Semi-Supervised Techniques

May be prohibitively costly to provide enough labeled data for a supervised learning problem.

Turns out that accuracy of supervised text classifier can be (markedly) improved by adding large pool of unlabeled documents to a small number of training documents.

Intuition: unlabeled set provides useful information about joint probability over words.

e.g. in labeled data, the word "`military`" is associated with being a Republican speech.

Semi-Supervised Techniques

May be prohibitively costly to provide enough labeled data for a supervised learning problem.

Turns out that accuracy of supervised text classifier can be (markedly) improved by adding large pool of unlabeled documents to a small number of training documents.

Intuition: unlabeled set provides useful information about joint probability over words.

e.g. in labeled data, the word "`military`" is associated with being a Republican speech. We then use this fact to classify thousands of unlabeled speeches.

Semi-Supervised Techniques

May be prohibitively costly to provide enough labeled data for a supervised learning problem.

Turns out that accuracy of supervised text classifier can be (markedly) improved by adding large pool of unlabeled documents to a small number of training documents.

Intuition: unlabeled set provides useful information about joint probability over words.

e.g. in labeled data, the word "`military`" is associated with being a Republican speech. We then use this fact to classify thousands of unlabeled speeches.

but then we find that the word "`defence`" co-occurs with '`military`' in the unlabeled documents (which we just classified)

Semi-Supervised Techniques

May be prohibitively costly to provide enough labeled data for a supervised learning problem.

Turns out that accuracy of supervised text classifier can be (markedly) improved by adding large pool of unlabeled documents to a small number of training documents.

Intuition: unlabeled set provides useful information about joint probability over words.

- e.g. in labeled data, the word "`military`" is associated with being a Republican speech. We then use this fact to classify thousands of unlabeled speeches.
- but then we find that the word "`defence`" co-occurs with '`military`' in the unlabeled documents (which we just classified)
- → use this to build more accurate classifier.

Exercise

Exercise

Consider a human infant learning certain concepts.

Exercise

Consider a human infant learning certain concepts.

1. How does an (average) infant learn the correct way to hold a cup?

Exercise

Consider a human infant learning certain concepts.

1. How does an (average) infant learn the correct way to hold a cup? Supervised, unsupervised, semi-supervised?

Exercise

Consider a human infant learning certain concepts.

1. How does an (average) infant learn the correct way to hold a cup? Supervised, unsupervised, semi-supervised?
2. How does an (average) infant learn that a Sharpei is a dog, not a cat?

Exercise

Consider a human infant learning certain concepts.

1. How does an (average) infant learn the correct way to hold a cup? Supervised, unsupervised, semi-supervised?

2. How does an (average) infant learn that a Sharpei is a dog, not a cat? Supervised, unsupervised, semi-supervised?