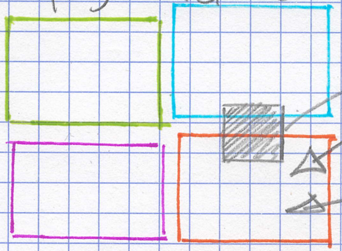
MIXED MEMBERSHIP MODELING VIA LATENT DIRICHLET ALLOCATION

Mixed membership models for documents

So far, clustered articles into groups



Doc labeled with a topic assignment

Soft assignment r_{ik} tells us this doc could be about world news or science.

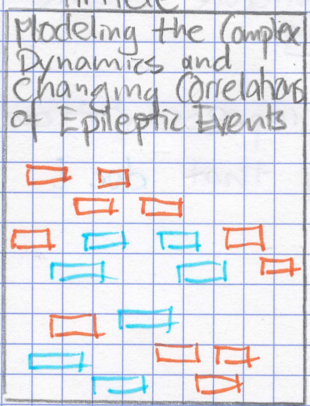
Is this article just about science?

Clustering goal: discover groups of related docs

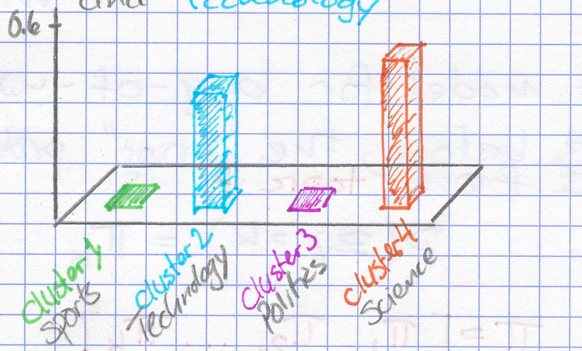
Soft assignments capture uncertainty

But, clustering model still specifies each doc belongs to 1 topic

Article



Really, it's about science and technology



"z_i" is both 2 and 4

Mixed membership models

Want to discover a set of memberships

(In contrast, cluster models aim at discovering a single membership)

An alternative document clustering model

So far we have considered...

Article

$x_i =$ [| | tf-idf vector | |]

Bag-of-word representation

Article

epilepsy
modeling
complex
clinical
Bayesian

$x_i = \{$modeling, complex, epilepsy, modeling, Bayesian, clinical, epilepsy, EEG, data, dynamic ...$\}$

multiset

= unordered set of words with duplication of unique elements mattering

A model for bag-of-words representation

As before, the "prior" probability that doc i is from topic k is:

$$p(z_i = k) = \pi_k$$

$$\pi = [\pi_1 \ \pi_2 \ \dots \ \pi_K]$$

represents corpus-wide topic prevalence

Assuming doc i is from topic k, words occur with probabilities:

SCIENCE

$$\left.\begin{array}{ll} \text{patients} & 0.05 \\ \text{clinical} & 0.01 \\ \text{epilepsy} & 0.002 \\ \text{seizures} & 0.0015 \\ \text{EEG} & 0.001 \\ \dots & \dots \end{array}\right\} \text{Words in vocab}$$

Topic-specific word probabilities

Distribution on words in vocab for each topic

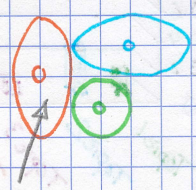
SCIENCE		TECH		SPORTS		
experiment	0.1	develop	0.18	player	0.15	...
test	0.08	computer	0.09	score	0.07	
discover	0.05	processor	0.032	team	0.06	
hypothesize	0.03	user	0.027	goal	0.03	
climate	0.01	internet	0.02	injury	0.01	
...	...			...	...	

Comparing and contrasting

Previously

Prior topic probabilities: $p(z_i = k) = \pi_k$

likelihood under each topic:



tf-idf vector

Compute likelihood of tf-idf vector under each Gaussian

Now

$p(z_i = k) = \pi_k$

SCIENCE | TECH | SPORTS | ...

{modeling, complex, epilepsy, modeling, Bayesian, clinical, epilepsy, EEG, data, dynamic ...}

compute likelihood of the collection of words in doc under each topic distribution

Components of latent Dirichlet allocation model

LDA is a mixed membership model

Topic vocab distribution

SCIENCE

TECH

SPORTS

Article

Modeling the Complex Dynamics and Changing Correlations of Epileptic Events

:

Clustering:

One topic indicator z_i per document i

All words come from (get scored under) same topic z_i

Distribution on prevalence of topics in corpus

$\pi = [\pi_1 \; \pi_2 \; \dots \; \pi_k]$

In LDA:

One topic indicator z_{iw} per word in doc i

Each word gets scored under its topic z_{iw}

Distribution on prevalence of topics in document

$\pi_i = [\pi_{i1}\ \pi_{i2}\ \dots\ \pi_{ik}]$

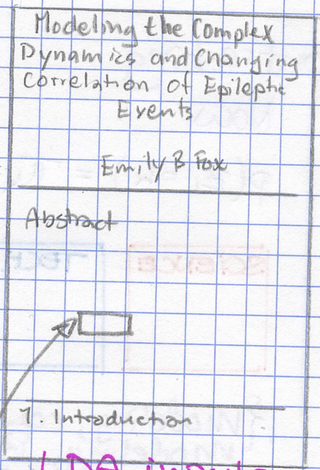
Goal of LDA inference

Topic vocab distributions:

TOPIC 1
Word 1 ?
Word 2 ?
Word 3 ?
Word 4 ?
Word 5 ?
... ...

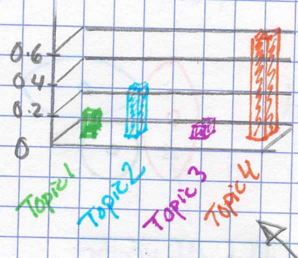
TOPIC 2
Word 1 ?
Word 2 ?

TOPIC 3
Word 1 ?



Document topic proportions

$\pi_i = [\pi_{i1}\ \pi_{i2}\ \dots\ \pi_{ik}]$



LDA inputs:
- Set of words per doc for each doc in corpus

LDA Outputs:
- Corpus-wide topic vocab distributions
- Topic assignments per word
- Topic proportions per doc

Interpreting LDA outputs

TOPIC 1

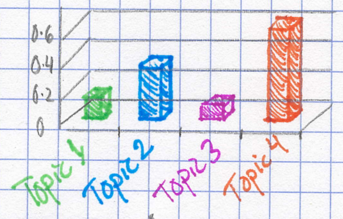
experiment	0.1
test	0.08
discover	0.05
hypothesize	0.03
climate	0.01

TOPIC 2

TOPIC 3

Examine coherence of learned topics
- What are top words per topic?
- Do they form meaningful groups?
- Use to post-facto label topics e.g. science, tech, sports

Interpreting LDA outputs (cont'd)



Doc-specific topic proportions can be used to:

- Relate documents
- Study user topic preferences
- Assign docs to multiple categories