Text Mining
Statistical Modeling of Textual Data
Lecture 3

Mattias Villani

Division of Statistics
Dept. of Computer and Information Science
Linköping University

Overview Lecture 3

- Demo of **document classification** in the **tm package** in R.
- **Topic models (LDA)**
- Demo of **topicmodels package** in R

Topic models

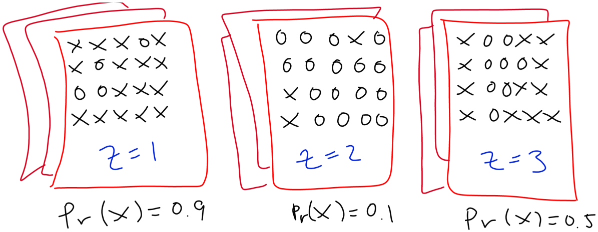
- Models for **unsupervised learning** [No need for labelled data!], but more recently also for **supervised learning**.
- **Probabilistic generative** models.
- **Very popular** model in applications and research. >8000 Google scholar citations in 11 years.
- The basic topic models are extensions of the bag-of-words (unigram) model.
- **Unigram** model: each word is assumed to be drawn from the same word (term) distribution.

$$\hat{P}(w) = \frac{\#w}{N}$$

- **Many extensions** in recent years: nGrams, supervised, nonparametric, relational topics, correlated topics, dynamically time-varying topics.

Mixture of unigrams

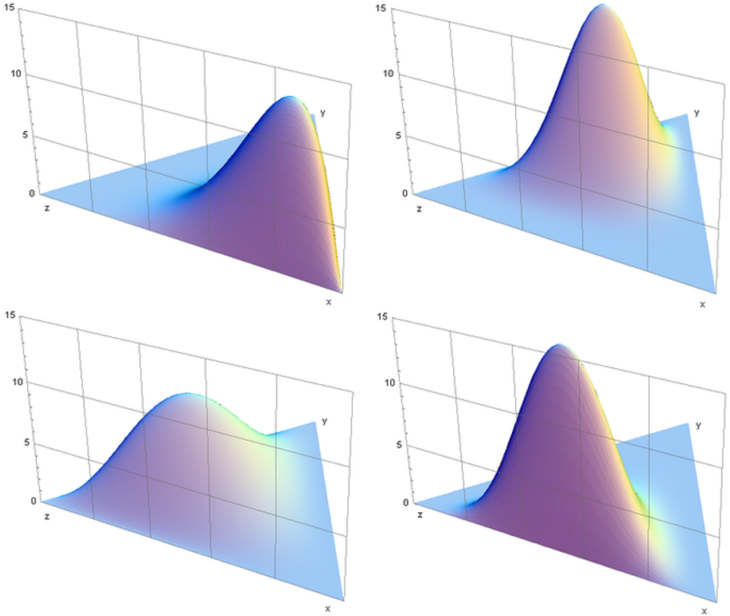
- **Mixture of unigrams**:
  1. Draw a *topic* z_d for the dth document from a topic distribution $\theta = (\theta_1, ..., \theta_K)$.
  2. Conditional on the drawn topic z_d draw words from a word distribution for that topic.



- Topic models are **mixed-membership models**: each document can belong to **several topics simultaneously**.

Multinomial and Dirichlet distributions

- **Multinomial distribution**: random discrete variable $X \in \{1, 2, ..., K\}$ that can assume exactly one of K (unordered) values.
  - $Pr(X = k) = \theta_k$
  - Parameters $\theta = (\theta_1, ..., \theta_K)$.
- **Dirichlet distribution**: random **vector** $X = (X_1, ..., X_K)$ satisfying the constraint $X_1 + X_2 + ... + X_K = 1$.
  - Unit simplex
  - Parameters: $\alpha = (\alpha_1, ..., \alpha_K)$
  - Uniform distribution: $\alpha = (1, 1, ..., 1)$
  - Small variance (informative) when the α's are large.
  - "Bathtub shape" when $\alpha_k < 1$ for all k.

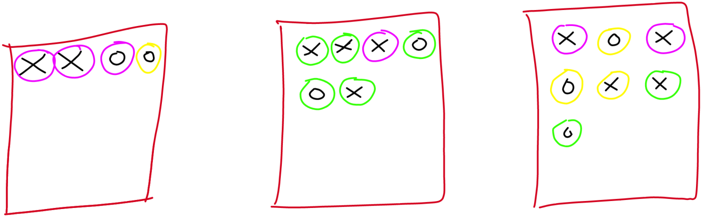

15
10
5
0
z
y
x
15
10
5
0
z
y
x
15
10
5
0
z
y
x
15
10
5
0
z
y
x

GENERATING A CORPUS FROM A TOPIC MODEL

- Assume that we have:
  - A fixed vocabulary V
  - D documents
  - N words in each document
  - K topics

1. **For each topic** $(k = 1, ..., K)$:
   A. Draw a distribution over the words $\beta_k \sim Dir(\eta, \eta, ..., \eta)$
2. **For each document** $(d = 1, ..., D)$:
   A. Draw a vector of topic proportions $\theta_d \sim Dir(\alpha_1, ..., \alpha_K)$
   B. **For each word** $(n = 1, ..., N)$:
      I. Draw a topic assignment $z_{d,n} \sim Multinomial(\theta_d)$
      II. Draw a word $w_{d,n} \sim Multinomial(\beta_{z_{d,n}})$

(Horrible picture of a) topic model



$\theta_1 = (0.9\ \ 0.1\ \ 0)$ $\theta_2 = (0.1\ \ 0.1\ \ 0.8)$ $\theta_3 = (0.3\ \ 0.4\ \ 0.3)$

$\beta_1 = (0.9\ \ 0.1)$

$\beta_2 = (0.1\ \ 0.9)$

$\beta_3 = (0.5\ \ 0.5)$

Example - simulation from two topics

Topic	Word distr.	probability	dna	gene	data	distribution
1	β_1	0.5	0.1	0.0	0.2	0.2
2	β_2	0.0	0.5	0.4	0.1	0.0

Doc 1	$\theta_1 = (0.2, 0.8)$		
	Word 1:	Topic=2	Word='gene'
	Word 2:	Topic=2	Word='gene'
	Word 3:	Topic=1	Word='data'

Doc 2	$\theta_2 = (0.9, 0.1)$		
	Word 1:	Topic=1	Word='probability'
	Word 2:	Topic=1	Word='data'
	Word 3:	Topic=1	Word='probability'

Doc 3	$\theta_2 = (0.5, 0.5)$		
	⋮	⋮	⋮

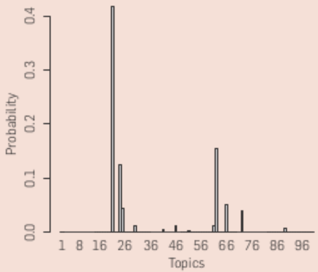
Learning/inference in topic models

- What do we know?
  - The words in the documents: $w_{1:D}$.
- What do we not know?
  - Topic proportions for each document: $\theta_{1:D}$
  - Topic assignments for each word in each document: $z_{1:D}$
  - Word distributions for each topic: $\beta_{1:K}$
- Do the Bayes dance: Posterior distribution

$$p(\theta_{1:D}, z_{1:D}, \beta_{1:K} | w_{1:D})$$

- The posterior is mathematically untractable. Solutions:
  - Gibbs sampling (MCMC) [Correct, but can be slow]
  - Variational Bayes [Crude approximation of the posterior *distribution*, but typically rather accurate about posterior mode (MAP)]
- The inferred $\theta_{1:D}$ can be used as features in supervised classification.

Figure 2. Real inference with LDA. We fit a 100-topic LDA model to 17,000 articles from the journal *Science*. At left are the inferred topic proportions for the example article in Figure 1. At right are the top 15 most frequent words from the most frequent topics found in this article.



"Genetics"	"Evolution"	"Disease"	"Computers"
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

From Blei (2012). Probabilistic topic models, Communication of the ACM.