# Computational Intelligence Laboratory

## Tutorial session 5, part 2
## pLSA and LDA

Antonio Orvieto

ETH Zurich – `cil.inf.ethz.ch`

19 March 2020

- Reversed classroom is a very good opportunity to understand the content. Today went great.. so please consider participating more if you like!
- These slides are heavily based on the lecture slides.. but they are not meant to substitute them[1].

[1] Got the citation? Write it in the chat..

# Get excited! Text analysis is beautiful..

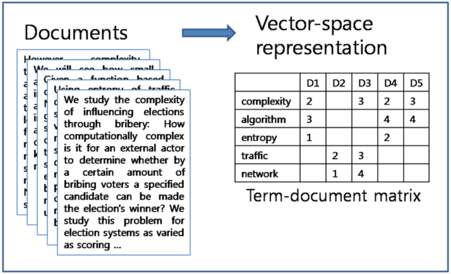
Our ability to understand and interact with the world is due to language..

A few books for your sweet quarantine:

- *Myth and meaning*, by Claude Levi-Strauss;
- *Plato's Pharmacy*, by Derrida.

# Topic Models

- given: corpus of text documents (e.g. web pages)
- goal: find (in an unsupervised way) low-dimensional document representation in **semantic space** of topics – **aboutness** of documents .
- assumption: Bag-of-word Representation
  - ignore order of words in sentences/document
  - reduce data to co-occurrence counts

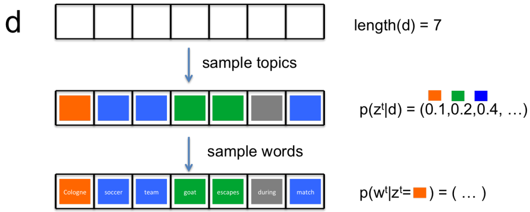



| | D1 | D2 | D3 | D4 | D5 |
|---|---|---|---|---|---|
| complexity | 2 | | 3 | 2 | 3 |
| algorithm | 3 | | | 4 | 4 |
| entropy | 1 | | | 2 | |
| traffic | | 2 | 3 | | |
| network | | 1 | 4 | | |

# pLSA (model)

Let's start by thinking about how we could see documents..

- each document = specific mix of topics (colors): $p(z|d)$
- each topic (color) = specific distribution of words: $p(w|z)$




Hence, we get the following model

$$p(w|d) = \sum_z p(w,z|d) = \sum_z p(w|d,z)p(z|d) \overset{*}{=} \sum_z p(w|z)p(z|d)$$

Conditional independence assumption (*)

# pLSA (cost function)

Let $x_{ij}$ be # occurrences of $w_j$ in document $d_i$ (i.e. our data).

We want our probabilistic model to explain the data — i.e. to maximize the log likelihood $\ell$!

$$\max \ell := \sum_{i,j} x_{ij} \log p(w_j | d_i)$$

$$= \sum_{i,j} x_{ij} \log \sum_{z=1}^{K} \underbrace{p(w_j | z)}_{=: v_{zj}} \underbrace{p(z | d_i)}_{=: u_{zi}},$$

where

- $u_{zi} \geq 0$ such that $\sum_z u_{zi} = 1$ $(\forall i)$
- $v_{zj} \geq 0$ such that $\sum_j v_{zj} = 1$ $(\forall z)$

goal: learn matrices $U$ and $V$ — i.e. the **model parameters**. How can we do that?

# Exercise 3 (i)

$$\max_{U,V} \ell(U,V) = \sum_{i,j} x_{ij} \log \sum_{z=1}^{K} v_{zj} u_{zi}$$

*Is this problem convex? Closed form solution?*

Consider two topics for one document and one word, then

$$-\ell(x) = -\log(u_1 v_1 + u_2 v_2), \quad x = (u_1, v_1, u_2, v_2)$$

The above function is not convex. Pick

$$x = (1,1,0,0), \quad y = (0,0,1,1)$$

$$-\ell(x/2 + y/2) = -\log(1/2) > 0 = (-\ell(x) - \ell(y))/2 \ \#$$

Note: this does not mean the problem is necessarily hard!
One can solve it with Projected Gradient Descent, and find a local minimizer. However, this is just slow!!

# pLSA (algorithm, 1)

Note: **we do not observe what is the topic** (color) for each word in each document.. otherwise likelihood maximization is trivial (see next slide)!

Assume we have this variable (even though its latent) and

- it is called $Q_{zij} \in \{0, 1\}$. It is 1 if $w_j$ in $d_i$ generated via $z$.
- $q_{zij} = \Pr(Q_{zij} = 1)$, $\sum_z q_{zij} = 1$, *variational parameters*.

Note that, if $U, V$ known, there is *some meaningful way* to find the $q_{zij}$:

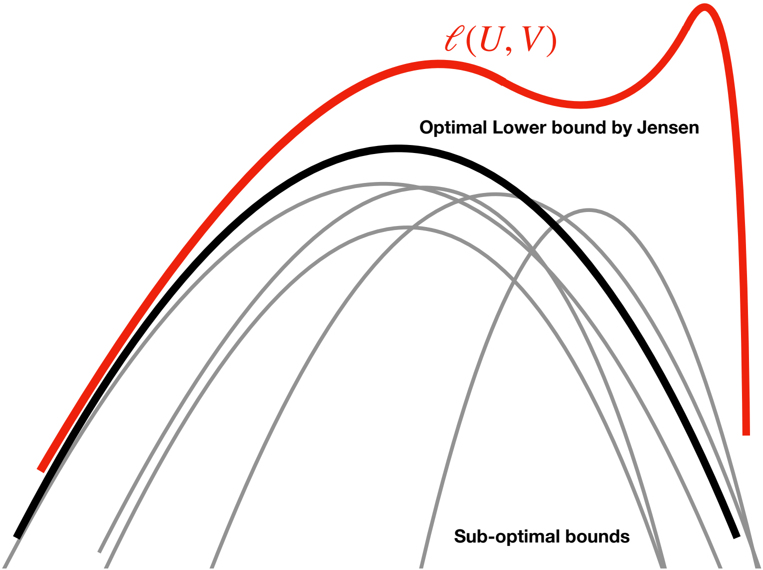
- Lower bound from Jensen's inequality

$$\ell(U, V) = \sum_{i,j} x_{ij} \log \sum_{z=1}^{K} q_{zij} \frac{u_{zi} v_{zj}}{q_{zij}}$$

$$\geq \sum_{i,j} x_{ij} \sum_{z=1}^{K} q_{zij} \left[\log u_{zi} + \log v_{zj} - \log q_{zij}\right].$$

- Solve for optimal $q$ (Expectation Step)

$$q_{zij} = \frac{u_{zi} v_{zj}}{\sum_{k=1}^{K} u_{ki} v_{kj}} = \frac{p(w_j|z)p(z|d_i)}{\sum_{k=1}^{K} p(w_j|k)p(k|d_i)}.$$


$\ell(U,V)$
Optimal Lower bound by Jensen
Sub-optimal bounds

# pLSA (algorithm, 2)

Solve for optimal parameters (Maximization Step)

$$u_{zi} = \frac{\sum_j x_{ij} q_{zij}}{\sum_j x_{ij}}, \qquad v_{zj} = \frac{\sum_i x_{ij} q_{zij}}{\sum_{i,l} x_{il} q_{zil}},$$

Alternate between the two!

- guaranteed convergence (cf. mixture models)
- **not** guaranteed to find global optimum

> *I thought that instead of the great number of precepts of which logic is composed, I would have enough with the four following ones, provided that I made a firm and unalterable resolution not to violate them even in a single instance. The first rule was never to accept anything as true unless I recognized it to be certainly and evidently such . The second was to* **divide each of the difficulties which I encountered into as many parts as possible, and as might be required for an easier solution***.*
>
> – Descartes

# Important remark

Why the first step is called *expectation*? Why are the $q_{zij}$ called *variational*?

- $q_{zij}$ is the posterior of $Q_{zij}$ given the current pair $(U, V)$ under the model. Since it is a binary variable, this posterior coincides with the expectation.
- At each step, $q_{zij}$ can be thought as an approximation of the true posterior. In that case, we can think of distance between distributions (hence calculus of *variations* on functionals such as the KL divergence).

If interested: Read from Bishop's book *Pattern recognition and machine learning*

- 9.4 EM algorithm in general;
- 10.1 Connection to variational inference.