Latent Dirichlet Allocation

David Mareček

October 26, 2022
Many of the slides in this presentation were taken from the presentations of Carl Edward Rasmussen (University of Cambridge)
With the Expectation-Maximization algorithm we have essentially estimated $\theta$ and $\beta$ by maximum likelihood.

\[
\begin{align*}
z_d & \sim \text{Cat}(\theta) \\
w_{nd} | z_d & \sim \text{Cat}(\beta_{z_d})
\end{align*}
\]
Bayesian Mixture of Categoricals Model

An alternative, Bayesian treatment infers these parameters starting from priors, e.g.:

- $\tilde{\theta} \sim Dir(\tilde{\alpha})$ is a symmetric Dirichlet over category probabilities,
- $\tilde{\beta}_k \sim Dir(\tilde{\gamma})$ are symmetric Dirichlets over vocabulary probabilities.

What is different?

- We no longer want to compute a point estimate of $\tilde{\theta}$ and $\tilde{\beta}$.
- We are now interested in computing posterior distributions.
Limitations of the mixture of categoricals model

A generative view of the mixture of categoricals model:

1. Draw a distribution $\vec{\theta}$ over $K$ topics from a $Dir(\vec{\alpha})$.
2. For each topic $k$, draw a distribution $\vec{\beta}_k$ over words from a $Dir(\vec{\gamma})$.
3. For each document $d$, draw a topic $z_d$ from a $Cat(\vec{\theta})$.
4. For each document $d$, draw $N_d$ words $w_{nd}$ from a $Cat(\vec{\beta}_{z_d})$.

Limitations:

- All words in each document are drawn from one specific topic distribution.
- This works if each document is exclusively about one topic, but if some documents span more than one topic, then "blurred" topics must be learnt.
Seeking Life’s Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here, two genome researchers with radically different approaches presented complementary views of the basic genes needed for life.

One research team, using computer analyses to compare known genomes, concluded that today’s bacteria can be sustained with just 250 genes—and that the earliest life forms required mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn’t be enough.

Although the numbers don’t match precisely, those predictions


"are not all that far apart," especially in comparison to the 75,000 genes in the human genome. George M. Church of Harvard University, in Boston, who arrived at the 800 number, but coming up with a consensus answer may be more than just a genetic numbers game. Particularly as more and more genomes are sequenced and compared, "It may be a way of organizing any newly sequenced genomes," explains Ardy G. T. M. van Boeijen, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

Striping down, computer analysis yields an estimate of the minimum modern and ancient genomes.
In reality, we only observe the documents.

The other structure are hidden variables.
Our goal is to infer the hidden variables.

This means computing their distribution conditioned on the documents

\[ p(\text{topics, proportions, assignments} | \text{documents}) \]
Nodes are random variables; edges indicate dependence.

Shaded nodes indicate observed variables.
Mixture of Categoricals vs. LDA

A generative view of LDA:

1. For each document $d$ draw a distribution $\vec{\theta}_d$ over topics from a $Dir(\vec{\alpha})$.
2. For each topic $k$ draw a distribution $\vec{\beta}_k$ over words from a $Dir(\vec{\gamma})$.
3. Draw a topic $z_{nd}$ for the $n$-th word in document $d$ from a $Cat(\vec{\theta}_d)$.
4. Draw word $w_{nd}$ from a $Cat(\vec{\beta}_{z_{nd}})$.

Differences with the mixture of categoricals model:

- In LDA, every word in a document can be drawn from a different topic,
- and every document has its own distribution over topics $\vec{\theta}_d$. 
• Initialize $z_{nd}$ randomly for all words in all documents
• Choose random word and sample a new category based on all other words in all other documents.
• The distribution across categories is the predictive distribution of the posterior Dirichlet distribution (integration across all possible $\vec{\theta}$ and $\vec{\beta}$).
• Perform these small changes in many iterations across the data. The algorithm will converge to a good solutions.