10-708: Probabilistic Graphical Models 10-708, Spring 2018

12 : Introduction to Topic Modeling and Factor Analysis

Lecturer: Kayhan Batmanghelich *Scribes: Gregory Plumb, Aaron Rumack*

1 Topic Modeling

Topic modeling is a method that can organize documents into thematic categories, describe the categories and how they change over time, find relationships between categories, and explain the effect of authorship on content.

1.1 Latent Dirichlet Allocation: The Model

LDA is an admixture model, in which each document is classified as a weighted mixture of different topics.

The generative process is as follows:

for topic $k \in \{1, ..., K\}$ **do**
  $\phi \sim$ Dirichlet(β) {Draw distribution over words}
end for
for document $m \in \{1, ..., M\}$ **do**
  $\theta_m \sim$ Dirichlet(α) {Draw distribution over topics}
  for word $w \in \{1, ..., N_m\}$ **do**
    $z_{mn} \sim$ Mult$(1, \theta_m)$ {Draw topic assignment}
    $x_{mn} \sim \phi_{z_{mn}}$ {Draw word}
  end for
end for

The two Dirichlet distributions are parameterized by α and β. As α or β goes to 1, the respective distribution is more uniform, and as the parameter goes to 0, the distribution is more sparse.

As we can see, the generative model initially requires only a Dirichlet prior over the topics. Each topic is represented as a multinomial distribution over each word in the vocabulary. For example, the topic of hockey would have higher probabilities for words such as team, season, hockey, ice, and puck.

Each document is represented as a distribution over topics. Each word has a distribution of belonging to the set of topics, and the document is classified according to the probability of each word belonging to each topic.

1.2 Latent Dirichlet Allocation: The Context

Topic modeling has two goals: to assign each topic as few words as possible with high probability, and allocate words to as few topics as possible within a single document. However, allocating words to a few topics within a single document will increase the number of words associated with each topic; and decreasing the

number of words associated with each topic will increase the number of topics allocated to each document. To satisfy both of these goals, we need to find sets of tightly co-occuring words.

When learning the Dirichlet distribution, we cannot use a simple L-1 regularization of ϕ to induce sparsity, because every point on the Dirichlet distribution has an L-1 norm of 1. Instead, we minimize $\mathcal{L}(X, \phi) + \lambda \sum_i \log(\phi_i + \epsilon)$.

1.3 Case Study: Imaging and Genetic Data

In this study, each document was a patient's data, and each topic was a medical condition. Each patient had a 3-D image of the lungs, and a genetic screening. The features extracted from the data were supervoxel intensities from the lung image, and genetic loci of interest from the DNA sequencing.

Certain medical conditions are associated with certain abnormalities in the lung image and DNA sequence, and therefore, the researchers were able to use LDA to allocate a mixture of medical conditions to each patient based on the lung and DNA data. Even though there were two completely different types of data sources, LDA was still able to be a useful technique.

2 Factor Analysis

To begin, we need a couple of facts:

- If x_1 and x_2 are multivariate Gaussian, then $x1|x2 \sim \mathcal{N}(\mu_1 + \Sigma_{1,2}\Sigma_{2,2}^{-1}(x_2 - \mu_2), \Sigma_{1,1} - \Sigma_{1,2}\Sigma_{2,2}^{-1}\Sigma_{2,1})$
- The Woodbury matrix identity
- $tr[ABC] = tr[CAB] = tr[BCA]$, $\frac{\partial}{\partial A} tr[BA] = B^T$, $\frac{\partial}{\partial A} log(|A|) = A^{-1}$

Factor Analysis is an unsupervised regression from X to Y where $x \sim \mathcal{N}(0, I)$ and $y|x \sim \mathcal{N}(\mu + \Lambda x, \Psi)$. In other words, Y is generated linearly by a lower dimensional space X. Then $\begin{bmatrix} X \\ Y \end{bmatrix} \sim \mathcal{N}(\begin{bmatrix} 0 \\ \mu \end{bmatrix}, \begin{bmatrix} 0 & \Lambda^T \\ \Lambda & \Lambda\Lambda^T + \Psi \end{bmatrix})$

2.1 Inference

Using the conditional Gaussian and the matrix identity, we conclude that $X|Y \sim \mathcal{N}(\mu_{x|y}, V_{x|y})$, where $V_{x|y} = (I + \Lambda^T \Psi^{-1} \Lambda)^{-1})$ and $\mu_{x|y} = V_{x|y} \Lambda^T \Psi^{-1}(y - \mu)$ Note that:

- Only have to invert a matrix whose size depends on the dimension of X and not Y
- Posterior covariance doesn't depend on the data
- Poster mean is a linear function

2.2 Learning for Factor Analysis

Suppose now that we only observe the samples $\{y_n\}$. We need to estimate μ, Λ and Ψ. Estimating μ is trivial, but Λ and Ψ are coupled non-linearly in the log-likelihood. So they are estimated using an Expectation Maximization algorithm.

2.3 Model Invariance and Unidentifiability

Because the parameter Λ only appears as $\Lambda\Lambda^t$ in the final model, it is inherently invariant to any orthonormal transformation. As a result, the model is unidentifiable.