Project 2: Latent Dirichlet Allocation

Kennard Ng and Luo Wen Han

November 8, 2019

1 Introduction

In this project, we find the underlying population distribution of an individual by his genotype. We model the this problem using the Latent Dirichlet Allocation (LDA) [1] graphical model. Given that inference and learning on the LDA is intractable, we use variational inference with the mean field assumption to perform approximate inference.

2 Background

2.1 Latent Dirichlet Distribution

The Latent Dirichlet Allocation (LDA) is a generative probabilistic model for a collection of discrete data proposed by Blei et al. [1]. In the original paper, Blei et al. use LDA to model text corpora, where they represent documents as a random mixture over latent topics where each topic is further characterized by their word distribution.

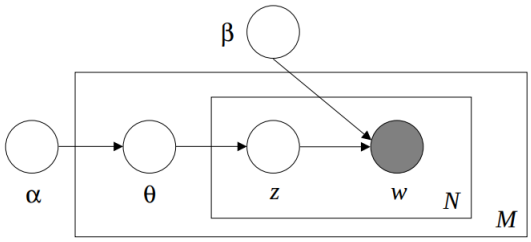

Figure 1: Graphical Model of LDA. Taken from [1]. The plates represent replicas. In our context, the inner plate represents the repeated choice of ancestor populations and genotypes of a person while the outer plate represents the population with M individuals.

Similarly, we model a population's genetics by representing each individual as a mixture of his ancestor populations, where each ancestor population can be characterized by its distribution over the genotypes i.e. genetic data. We assume the following generative process for the genetic makeup of an individual:

1. Assume that each individual has N genotypes.
2. Choose $\theta \sim \text{Dir}(\alpha)$.

3. For each of the N genotypes w_n:
   (a) Choose an ancestor population $z_n \sim \text{Multinomial}(\theta)$.
   (b) Choose a genotype w_n from $p(w_n|z_n, \beta)$, where $p(w_n|z_n, \beta)$ is a multinomial distribution conditioned on the ancestor population z_n and genotype prior β.

2.2 Variational Inference

During inference, we want to find the distributions over hidden variables θ, z given parameters α, β and observed variables w. We express this as the following posterior distribution:

$$p(\theta, z \mid w, \alpha, \beta) = \frac{p(\theta, z, w \mid \alpha, \beta)}{p(w \mid \alpha, \beta)}. \tag{1}$$

Unfortunately, we see that exact inference is intractable given that the normalizer is intractable where

$$p(w|\alpha, \beta) = \int_{-\infty}^{\infty} \sum_{n=1}^{N} p(\theta, z_n, w \mid \alpha, \beta) d\theta. \tag{2}$$

Hence, we make an approximate inference instead. We use the mean field assumption that assumes fully factorized distributions and we approximate the posterior distribution p with an approximate distribution:

$$q(\theta, z \mid \gamma, \phi) = q(\theta \mid \gamma) \prod_{n=1}^{N} q(z_n \mid \phi_n). \tag{3}$$

The graphical model of our approximate distribution is given in Figure 2.

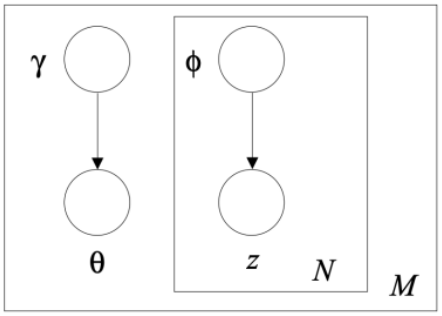

Figure 2: Approximate Distribution under the Mean Field Assumption.

References

[1] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.