# Problem Set 4

May 24, 2017

## 1 Clustering: Mixture of Multinomials

### 1.1 MLE for multinomial

Derive the maximum-likelihood estimator for the parameter $\boldsymbol{\mu} = (\mu_i)_{i=1}^d$ of a multinomial distribution:

$$P(\boldsymbol{x}|\boldsymbol{\mu}) = \frac{n!}{\prod_i x_i!} \prod_i \mu_i^{x_i}, \quad i = 1, \cdots, d \tag{1}$$

where $x_i \in \mathbb{N}$, $\sum_i x_i = n$ and $0 < \mu_i < 1$, $\sum_i \mu_i = 1$.

### 1.2 EM for mixture of multinomials

Consider the following mixture-of-multinomials model to analyze a corpus of documents that are represented in the bag-of-words model.

Specifically, assume we have a corpus of $D$ documents and a vocabulary of $W$ words from which every word in the corpus is token. We are interested in counting how many times each word appears in each document, regardless of their positions and orderings. We denote by $T \in \mathbb{N}^{D \times W}$ the word occurrence matrix where the $w$th word appears $T_{dw}$ times in the $d$th document.

According to the mixture-of-multinomials model, each document is generated i.i.d. as follows. We first choose for each document $d$ a *latent* "topic" $c_d$ (analogous to choosing for each data point a component $z_n$ in the mixture-of-Gaussians) with

$$P(c_d = k) = \pi_k, \, k = 1, 2, \cdots, K; \tag{2}$$

And then given this "topic" $\boldsymbol{\mu}_k = (\mu_{1k}, \ldots, \mu_{Wk})$ which now simply represents a categorical distribution over the entire vocabulary, we generate the word bag of the document from the corresponding multinomial distribution[1]

$$P(d|c_d = k) = \frac{n_d!}{\prod_w T_{dw}!} \prod_w \mu_{wk}^{T_{dw}}, \text{ where } n_d = \sum_w T_{dw}. \tag{3}$$

[1] Make sure you understand the difference between a categorical distribution and a multinomial distribution. You may think about a Bernoulli distribution and a binomial distribution for reference.

Hence in summary

$$P(d) = \sum_{k=1}^{K} P(d|c_d = k)P(c_d = k) = \frac{n_d!}{\prod_w T_{dw}!} \sum_{k=1}^{K} \pi_k \prod_w \mu_{wk}^{T_{dw}}. \tag{4}$$

Given the corpus $T$, please design and implement an EM algorithm to learn the parameters $\{\boldsymbol{\pi}, \boldsymbol{\mu}\}$ of this mixture model and test it on the NIPS dataset (http://ml.cs.tsinghua.edu.cn/~jianfei/static/nips.tar.gz)

Set the number of topics $K$ to be $5, 10, 20, 30$ respectively and show the most-frequent words in each topic for each case. Observe the result and try to find the "best" $K$ value for this dataset and explain why.