Phenotyping via Bayesian Nonparametric Tensor Factorization

***Abstract*—We do phenotyping.**

I. Introduction

II. Model

A. CANDECOMP/PARAFAC (CP) Tensor Factorization

B. Dirichlet Process Tensor Factorization

TABLE I. Data Parameters

Notation	Meaning
λ	The weight vector of phenotypes, here $\|\lambda\| = 1$
λ_k	The weight of k-th phenotype; $\lambda_k \in [0, 1]$
N_p	Number of patients
N_m	Number of medications
N_d	Number of diagnosis
T	Record Tensor. $T \in \{0, 1\}^{N_p \times N_m \times N_d}$

TABLE II. Model Parameters

Notation	Meaning
α	The hyper parameter of Beta Distribution
β_i	The stick-breaking process random variable. $\beta_i \sim Beta(1, \alpha)$
γ_p	The hyper parameter of prior Dirichlet Distribution of patient
γ_m	The hyper parameter of prior Dirichlet Distribution of medication
γ_d	The hyper parameter of prior Dirichlet Distribution of diagnosis
$\theta^{(p,k)}$	The multinomial distribution over patients given the phenotype k
$\theta^{(m,k)}$	The multinomial distribution over medications given the phenotype k
$\theta^{(d,k)}$	The multinomial distribution over diagnosis given the phenotype k

We place a stick-breaking process prior on the λ. The parametere β_i follows the Beta Distribution $\beta_i \sim Beta(1, \alpha), i = 1, 2, \ldots$. And:

$$\lambda_k = \beta_k \prod_{i=1}^{k-1} (1 - \beta_i), k = 1, 2, \ldots$$

A $k-$dimensional Dirichlet random variable θ can take values in the $(k-1)$-simplex where $\sum_{i=1}^{k} \theta_i = 1$, and the probability density is

$$p(\theta|\gamma) = \frac{\mathcal{T}(\sum_{i=1}^{k} \gamma_i)}{\prod_{i=1}^{k} \mathcal{T}(\gamma_i)} \theta_1^{\gamma_1 - 1} \ldots \theta_k^{\gamma_k - 1}$$

It is natural to place the conjugate prior (Dirichlet Distribution) over the multinomial distributions parameterized by $\theta^{(p)}, \theta^{(m)}$ and $\theta^{(d)}$:

$$\begin{aligned} \theta^{(p,k)} &\sim Dir(\gamma_p), k = 1, 2, \ldots \\ \theta^{(m,k)} &\sim Dir(\gamma_m), k = 1, 2, \ldots \\ \theta^{(d,k)} &\sim Dir(\gamma_d), k = 1, 2, \ldots \end{aligned}$$

The patients, medications and diagnosis all follow the multinomial distribution parameterised by θ, specifically:

$$\begin{aligned} \text{patient}_i &\sim Multi(\theta^{(p)}), i = 1, 2, \ldots, N_p \\ \text{medication}_i &\sim Multi(\theta^{(m)}), i = 1, 2, \ldots, N_m \\ \text{diagnosis}_i &\sim Multi(\theta^{(d)}), i = 1, 2, \ldots, N_d \end{aligned}$$

Given the parameters, the likelihood of given record triplet (i, j, k) (patient i had diagnostic record k, and has taken medicine j) can be formulated as:

$$p(T_{i,j,k}|\alpha, \gamma_p, \gamma_m, \gamma_d) = \sum_{r=1}^{\infty} \lambda_r p(i|r, \gamma_p) p(j|r, \gamma_m) p(k|r, \gamma_d)$$

III. Experiments

IV. Conclusion

Acknowledgment

References

[1] H. Kopka and P. W. Daly, *A Guide to LaTeX*, 3rd ed. Harlow, England: Addison-Wesley, 1999.