Contents

Indian Institute of Technology Kanpur
CS777 Topics in Learning Theory

Authors: Rushab Munot(14405), Ameya Loya
Instructor: Purushottam Kar
Date: March 5, 2018

Investigatory Project

Generating Sense Vectors Probabilistically

1 Introduction

We present a way to view each word as probability distribution rather than a deterministic fixed vector. In particular, each word is a mixture of distributions corresponding to its senses. This method is generalization of traditional method that views a word as a deterministic fixed vector which can be obtained by finding the mean of the distribution.

2 Model by Athiwaratkun and Wilson

Athiwaratkun *et al.* have represented each word w as a mixture of its senses where each sense is a Gaussian. More precisely, if w is a word and it has K senses, then its distribution would be given by

$$P(x) = \sum_{k=1}^{k=K} p_{wi}\mathcal{N}(x \mid \mu_{wi}, \Sigma_{wi})$$

where x is the instance of w, p_{wi} are the mixture proportions of the word w, μ_{wi} and Σ_{wi} are the mean and covariance of the ith component. For e.g., one component of the word "hard" may refer to something that is rigid or tough whereas the other component may refer to something that requires a lot of effort.

3 Problems with the model

Let the number of words in our vocabulary be V. Let the average number of senses of each word be K. If we represent each instance of a word by a d dimensional vector then the total number of parameters to be estimated will be of the order $\mathcal{O}\left(VKd^2\right)$. In a general setting, $V \approx 10^5$, $K \approx 5$ and $d \approx 50$ this means that the number of parameters to be estimated is of the order 10^9. Clearly, the model is not scalable. To solve this issue, Athiwaratkun *et. al.* have fixed the number of senses for each word to 2. Now, although the number of parameters to be estimated have reduced, we have payed the price of considering only a fixed number of senses for each word. In general, however, we would want the number of senses to not be fixed and grow or scale as a new sense of a word appears.

4 The idea of a concept pool

To allow for indefinite number of senses for each word, we introduce the notion of a concept pool. We assume there is a fixed number m, of concept pools shared across all the words. Each sense of a word can be thought of as coming from these m concept pools. More specifically, we have assumed each concept pool to be a Gaussian. There are such m Gaussians shared across all the words and each sense of a word is thought of to be a convex combination of these concept pools. Finally, similar to Athiwaratkun *et. al.* we have considered each word to be a mixture of its senses.

5 Our Model

Now we present our model. Each word is drawn from a distribution that is the mixture of its senses. Each sense can be considered to be a mixture over concept pools. More explicitly,

$$w_i \sim \sum_{k=1}^{K_i^+} \pi_{ik} w_{ik}(\theta_{ik}) \tag{1}$$

where w_i is the ith word, π_{ik}, $w_{ik}(\theta)$ and θ_{ik} are the mixing proportion, the distribution and the parameters for the kth sense of the ith word respectively. We have assumed that there are ${K_i}^+$ senses for each word. Each concept pool is assumed to be a Gaussian with the mean and covariance matrix of the mth pool being ν_m and T_m respectively. Since each sense is a mixture over concept pools, we have

$$w_{ik}(\theta) \equiv \sum_{m=1}^{M} \psi_{ikm} \mathcal{N}(\, \nu_m \,,\, T_m \,) \tag{2}$$

where ψ_{ikm} denotes the mixing proportion for mth concept pool for the kth sense of the ith word. Note that we have

$$\theta_{ik} = (\nu_1, \, \ldots, \, \nu_m, T_1, \, \ldots \, , T_m, \psi_{ikm}, \, \ldots \, , \psi_{ikm})$$

We assume a Dirichlet prior over π_i

$$\pi_i \sim Dir\Big(\frac{\alpha}{K}, \ldots, \frac{\alpha}{K} \Big) \quad \text{for all } i \in \{1, 2, \ldots, V\}$$

We assume a Gaussian prior over ν_m conditioned over T_m

$$\nu_m \,|\, T_m \sim \mathcal{N}\big(\xi \,,\, (\rho T_m)^{-1} \big) \quad \text{for all } m \in \{1, 2, \ldots, M\}$$

We take a Wishart prior over T_m

$$T_m \sim \mathcal{W}\big(\lambda, (\lambda W)^{-1} \big) \quad \text{for all } i \in \{1, 2, \ldots, M\}$$

Next we assume a Dirichlet prior over ψ_{ik}

$$\psi_{ik} \sim Dir\Big(\frac{\beta}{M}, \ldots, \frac{\beta}{M} \Big) \quad \text{for all } i \in \{1, 2, \ldots, V\}, \quad k \in \{1, 2, \ldots, {K_i}^+\}$$

It is worth noting at this stage that our hyperparameters are ξ, ρ, W, λ, α, K, M

5.1 Liklihood

Let x_{ik} = instance of the sense k of word i. We wish to find the likelihood of x_{ik} conditioned on the parameters ν, T and ψ_{ik} i.e $P(x_{ik} \mid \nu, T, \psi_{ik})$. By definition of our model we have,

$$P(x_{ik} \mid \nu, T, \psi_{ik}) = \sum_{m=1}^{m=M} \psi_{ikm} \mathcal{N}(x_{ik} \mid \nu_m, T_m) \tag{3}$$

Now given X, which consists of n observations, say, $\{x^1, \ldots, x^n\}$ such that the ith word is an instance of word w_i, and its sense is s_i, we have(assuming iid)

$$P(X \mid \nu, T, s) = \prod_{i=1}^{i=n} \sum_{m=1}^{m=M} \psi_{w_i s_i m} \mathcal{N}(x^i \mid \nu_m, T_m) \tag{4}$$

5.2 Marginal Posterior

Given a set of observations X, we have the following marginal posteriors :

- $P(\nu_m \mid X, T_m, \psi_{ik}) \propto P(X \mid \nu_m, T_m, \psi_{ik}) P(\nu_m \mid T_m)$
- $P(T_m \mid X, \nu_m, \psi_{ik}) \propto P(X \mid \nu_m, T_m, \psi_{ik}) P(T_m)$
- $P(\psi_{ik} \mid X, \nu_m, T_m) \propto P(X \mid \nu_m, T_m, \psi_{ik}) Dir(\psi_{ik})$

Note that owing to non conjugacy, none of the posteriors can found. However, we can evaluate the numerator for each of the posteriors which will facilitate the sampling process.

5.3 Updating cluster id for instance w_i

We wish to find the probability that the instance w_i belongs to cluster j i.e it is an instance of the jth sense, given the senses of previous instances. We have,

$$P(w_i \mid c_i = j, c_{-i}) \tag{5}$$

$$= \int \left[P(w_i \mid \psi, \nu, T) P(\psi) d\psi \prod_{l=1}^{M} P(\nu_l \mid T_l) P(T_l) d\nu_l dT_l \right] \tag{6}$$

$$= \sum_{m=1}^{M} \left[\int \psi_m \mathcal{N}(\, w_i \mid \nu_m \,,\, T_m \,) P(\psi) d\psi P(\nu_m \mid T_m) P(T_m) d\nu_m dT_m \right] \tag{7}$$

$$= \sum_{m=1}^{M} \left[\int \psi_m P(\psi) d\psi \int \underbrace{\mathcal{N}(\, w_i \mid \nu_m \,,\, T_m \,) P(\nu_m \mid T_m) P(T_m)}_{\text{Conjugate (Normal-Wishart)}} d\nu_m dT_m \right] \tag{8}$$

Thus we have log probability,

$$\log P(w_i \mid c_i = j, c_{-i}) =$$
$$\frac{1}{B(\beta)} \cdot \frac{M^2}{M+\beta} + \frac{D}{2}\frac{\rho + n_j}{\rho + n_j + 1} - \frac{D}{2}\log \pi +$$
$$\log \Gamma\left(\frac{\lambda + n_j + 1}{2}\right) - \log \Gamma\left(\frac{\lambda + n_j + 1 - D}{2}\right) + \frac{\lambda + n_j}{2}\log |W^*| -$$
$$\frac{\lambda + n_j + 1}{2}\log \left| W^* + \frac{\rho + n_j}{\rho + n_j + 1}(w_i - \xi^*)(w_i - \xi^*)^T \right|$$
$$\xi^* = \frac{(\rho\xi + \sum_{i':c_{i'}=j} w_{i'})}{\rho + n_j}$$
$$W^* = +\rho\xi\xi^T + \sum_{i':c_{i'}=j} w_{i'} w_{i'} w_{i'}^T - (\rho + n_j)\xi^*\xi^{*T}$$

Note it is here that we are able to non parameterize the number of senses. We find the probability for each sense that the new instance belongs to to that sense. The cluster to which the given instance belongs is the one for which $P(w_i \mid c_i = j, c_{-j})$ is maximum. If the probability that the instance doesn't belong to an existing cluster is maximum, we introduce a new sense for the word of which w_i is an instance.

6 Sampling Procedure

- Update ν_m, T_m
- Update hyperparameters conditional on ν_m, T_m
- Update indicators c_i
- update α and ψ using Adaptive Rejection Sampling (log-concave conditional posteriors)

7 Conclusion and Future work

We have successfully built upon the work of Athiwaratkun *et. al.* by non parameterizing the number of senses by assuming that a basic concept pool exists. The number of parameters to be estimated in our model is $\mathrm{O}(VKm)\, which in a general setting turns out to be approximately 10^7$. Thus, we have been able to reduce the number of parameters that have to be estimated by a factor of 10^2 which makes our model a feasible one. The next step is to train and test our model and see whether it is scalable or not.

8 Referrences

- Dilan Gorur and Carl Edward Rasmussen, February 2010, Dirichlet Process Gaussian Mixture Models: Choice of the Base Distribution
- Ben Athiwaratkun, Andrew Gordon Wilson, April 2017, Multimodal Word Distributions
- Radford M. Neal, June 2000, Markov Chain Sampling Methods for Dirichlet Process Mixture Models