CS 247 : Advanced Data Mining Learning **(Due: 11:59 pm 05/08/19)**

Homework Assignment #3

Instructor: Yizhou Sun *TA:* Zeyu Li

Homework Policy:

- Read the ***Homework Submission Guidance*** carefully before you start working on the assignment, and before you make a submission.
- This is an individual homework. Please **DO NOT** collaborate with others.
- Write your answers in a **pdf** (typeset by LaTeX) and submit it to Gradescope.

Problem 1: pLSA Initialization

In the iterative pLSA [2] training algorithm, word distribution vectors $\{\beta_k\}_{k=1}^K$ and topic distribution vectors $\{\theta_d\}_{d=1}^M$, where K denotes the number of topics and M denotes the number of documents, need to be initialized.

1. Is it a good initialization to set θ_d's and β_k's as uniform distribution, i.e., $\theta_{dk} = \frac{1}{K}$ for every d and k, and $\beta_{kw} = \frac{1}{N}$ for every k and w, where N is the total number of words in the dictionary? Why?
2. Can you give another example of bad initialization?

Problem 2: Multinomial Naïve Bayes with Dirichlet Prior

In LDA [1], Dirichlet priors can be added to topic distributions and word distributions. Similarly, Dirichlet priors can be added to word distribution vectors β_k's in multinomial naïve Bayes model, i.e., $\beta_k|\alpha \sim \mathrm{Dir}(\alpha)$, where α is the parameter vector associated with the Dirichlet distribution. We use $\mathbf{x}_d$ to denote the bag-of-words vector in document d and y_d to denote the latent topic for document d.

1. Please write down the joint distribution $p(\mathbf{x}_d, y_d, \beta|\alpha)$;
2. Please write down the inference procedure, i.e., find $y^* = \arg\max p(\mathbf{x}_d, y)$;
3. Please write down the posterior distribution for β, i.e, $p(\beta_k|D, \alpha)$, where $D = \{(\mathbf{x}_d, y_d)\}_{d=1}^M$ is the labeled document dataset with M documents. Compute the posterior mean for β, i.e., $\mathbb{E}(\beta_k|D, \alpha)$. If $\alpha = (1, 1, \cdots, 1)$, i.e., an all one vector with dimensionality N, where N denotes the number of words in the vocabulary, what is the posterior mean and what is the connection between it and add-1 smoothing?

(Hints: the integral of density function $p(\beta_k|\alpha)$ over β_k equals to 1.)

Problem 3: Word Embedding

Word2Vec [3] is trained based on local context window. Suppose we aggregate all the co-occurrence information between words based on local context window as done in GloVe [4], where X_{ij} is denoted as the counts that w_j has appeared in w_i's context.

1. Is X_{ij} symmetric, i.e., $\forall i, j$, $X_{ij} = X_{ji}$? Why?
2. What would be the form of original objective function for skip-gram using X_{ij}?
3. Suppose we fix negative samples before training for negative sampling, and obtain X_{ij}^+ and X_{ij}^-, which denote the number of times w_j appears in w_i's local contexts and the number of times w_j appears in w_i's negative samples. What would be the form of negative sampling based objective function for skip-gram using X_{ij}^+ and X_{ij}^-?

References

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

[2] T. Hofmann. Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 289–296. Morgan Kaufmann Publishers Inc., 1999.

[3] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[4] J. Pennington, R. Socher, and C. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.