VAE Topic Detection

MATTEO POSENATO, TOMASO ERSEGHE*, Università degli studi di Padova, Italy

tbd

Additional Key Words and Phrases: tbd

ACM Reference Format:
Matteo Posenato, Tomaso Erseghe. 2023. VAE Topic Detection. *ACM Trans. Knowl. Discov. Data.* 1, 1 (May 2023), 3 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

[...]

2 NEURAL TOPIC MODEL BASED ON VAE

2.1 Generative model - Generation network

We review the topic detection model of [1]. We consider a corpus of D documents using a vocabulary of W words. Each document is represented by a (variable length) vector $\boldsymbol{d}_i$, $i = 1, \ldots, D$ collecting the words occurrences in the document, so that $d_{i,n} \in \{1, \ldots, W\}$. We let the corpus be organised in T topics, and denote with $\boldsymbol{t}_i$ (vector of length T) the latent topic representation of document $\boldsymbol{d}_i$.

Our reference generative model starts from an hidden prior variable $\boldsymbol{z} \in \mathbb{R}^T$ normally distributed, i.e., with probability distribution function (PDF) $p(\boldsymbol{z}) = p_{\mathcal{N}}(\boldsymbol{z}; \boldsymbol{0}, \boldsymbol{I})$ where

$$p_{\mathcal{N}}(\boldsymbol{x}; \boldsymbol{m}, \Sigma) = \frac{1}{\sqrt{\det(2\pi\Sigma)}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{m})^T \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{m})} \tag{1}$$

is the multivariate normal PDF. The latent topic representation $\boldsymbol{t} \in \mathbb{R}^T$ is approximated by a multilayer perceptron (MLP), to build a differentiable map of the form (if I understood correctly, but the text also mentions two MLPs?!?)

$$\boldsymbol{t} = \boldsymbol{\vartheta}(\boldsymbol{z}) = \boldsymbol{W}_2 \tanh(\boldsymbol{W}_1 \boldsymbol{z} + \boldsymbol{b}_1) + \boldsymbol{b}_2 \ . \tag{2}$$

The word-occurrence-pattern vector is then generated via softmax construction from the latent topic representation, that is (if I understood correctly, but the text also mentions a Gaussian softmax?!? also if this is correct then $\boldsymbol{W}_2$ and $\boldsymbol{W}_3$ are redundant so something is wrong)

$$\log\left(p_{d|t}(\boldsymbol{d}|\boldsymbol{t})\right) = \sum_n \log(s_{d_n}) \ , \quad \boldsymbol{s} = \text{softmax}(\boldsymbol{W}_3 \boldsymbol{t} + \boldsymbol{b}_3)$$

*All authors contributed equally to this research.

Author's address: Matteo Posenato, Tomaso Erseghe, tomaso.erseghe@unipd.it, Università degli studi di Padova, Via VII Febbraio, 2, Padova, Italy, 35131.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2023 Association for Computing Machinery.
Manuscript submitted to ACM

where s_{d_n} denotes the d_nth entry of $\boldsymbol{s}$. Hence, we have

$$p_\theta(\boldsymbol{d}|z) = p_{d|t}(\boldsymbol{d}|\boldsymbol{\vartheta}(z)) \tag{3}$$

with parameters $\theta = \{\boldsymbol{b}_1, \boldsymbol{b}_2, \boldsymbol{b}_3, \boldsymbol{W}_1, \boldsymbol{W}_2, \boldsymbol{W}_3\}$.

2.2 Approximation of posterior probability - Inference network

The posterior probability $q_\phi(z|\boldsymbol{d})$ is approximated by the multivariate normal distribution

$$q_\phi(z|\boldsymbol{d}) = p_{\mathcal{N}}(z; \boldsymbol{\mu}_\phi(\boldsymbol{d}), \mathrm{diag}(\boldsymbol{\sigma}^2_\phi(\boldsymbol{d}))) \tag{4}$$

where $\boldsymbol{\mu}_\phi$ and $\boldsymbol{\sigma}^2_\phi$ are differential maps generated through two MLPs. Specifically, we have (this is my guess from [2])

$$\begin{aligned} \boldsymbol{\mu}_\phi(\boldsymbol{d}) &= \boldsymbol{W}_5\boldsymbol{h} + \boldsymbol{b}_5\,, \qquad \boldsymbol{h} = \tanh(\boldsymbol{W}_4\boldsymbol{d} + \boldsymbol{b}_4) \\ \log(\boldsymbol{\sigma}^2_\phi(\boldsymbol{d})) &= \boldsymbol{W}_6\boldsymbol{h} + \boldsymbol{b}_6 \end{aligned}$$

2.3 Target function

According to the variational auto encoder (VAE) approach of [2] we define a variational lower bound $f_{\theta,\phi}(\boldsymbol{d}) \le \log p_\theta(\boldsymbol{d})$ as

$$\begin{aligned} f_{\theta,\phi}(\boldsymbol{d}) &= \log p_\theta(\boldsymbol{d}) - D_{\mathrm{KL}}\left(q_\phi(z|\boldsymbol{d}) \middle\| p_\theta(z|\boldsymbol{d})\right) \\ &= \int dz\; q_\phi(z|\boldsymbol{d}) \log\left(\frac{p_\theta(z,\boldsymbol{d})}{q_\phi(z|\boldsymbol{d})}\right) \\ &= \underbrace{\int dz\; q_\phi(z|\boldsymbol{d}) \log\left(p_\theta(\boldsymbol{d}|z)\right)}_{f_1} - \underbrace{\int dz\; q_\phi(z|\boldsymbol{d}) \log\left(\frac{p(z)}{q_\phi(z|\boldsymbol{d})}\right)}_{f_2} \end{aligned} \tag{5}$$

with target function $f_{\theta,\phi}(\boldsymbol{d})$ to be maximized with respect to the parameters θ and ϕ. By exploiting (4), the target function can be rewritten in the form

$$\begin{aligned} f_1(\boldsymbol{d}) &= \int d\boldsymbol{u}\; p_{\mathcal{N}}(\boldsymbol{u}; \boldsymbol{0}, \boldsymbol{I}) \log\left(p_\theta\left(\boldsymbol{d}\middle|\boldsymbol{\mu}_\phi(\boldsymbol{d}) + \boldsymbol{\sigma}_\phi(\boldsymbol{d}) \circ \boldsymbol{u}\right)\right) \\ &\simeq \frac{1}{L}\sum_{\ell=1}^{L} \log\left(p_\theta\left(\boldsymbol{d}\middle|\boldsymbol{\mu}_\phi(\boldsymbol{d}) + \boldsymbol{\sigma}_\phi(\boldsymbol{d}) \circ \boldsymbol{u}_\ell\right)\right) \\ f_2(\boldsymbol{d}) &= -\tfrac{1}{2}\boldsymbol{1}^T\left(\boldsymbol{1} + \boldsymbol{\mu}^2_\phi(\boldsymbol{d}) + \boldsymbol{\sigma}^2_\phi(\boldsymbol{d}) + \log(\boldsymbol{\sigma}^2_\phi(\boldsymbol{d}))\right) \end{aligned} \tag{6}$$

where $\circ$ stands for element-wise product, and where $\boldsymbol{u}_\ell \in \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ are independent normal samples.

2.4 Topic detection

One important question. If I understood correctly $\boldsymbol{W}_3$ is the matrix that must be paired with the sentiment one, and this matrix identifies the topics. However, how can we assign documents to topics? After VAE we know $q_\phi(z|\boldsymbol{d})$ so we can guess the statistics on $\boldsymbol{t}$ via

$$p(\boldsymbol{t}|\boldsymbol{d}) = q_\phi(\boldsymbol{\vartheta}^{-1}(\boldsymbol{t})|\boldsymbol{d})$$

but does it make any sense?

REFERENCES

[1] Lin Gui, Jia Leng, Jiyun Zhou, Ruifeng Xu, and Yulan He. 2020. Multi task mutual learning for joint sentiment classification and topic detection. *IEEE Transactions on Knowledge and Data Engineering* 34, 4 (2020), 1915–1927.

[2] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

Received tbd