Dirichlet Variational Autoencoders

Jason (Jiaxin) Liu
University of Waterloo
jason@jxnl.co

ABSTRACT

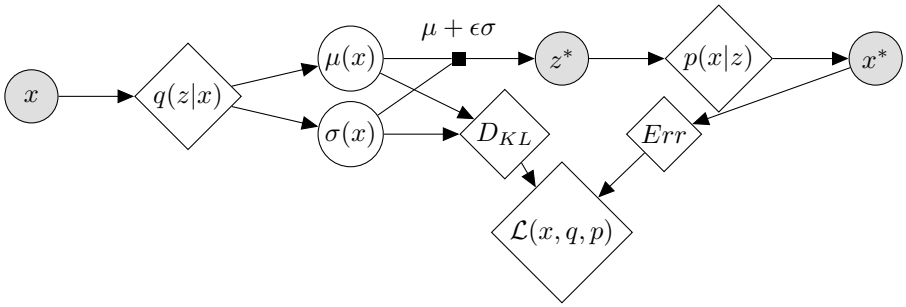
Variational Autoencoders (VAE) are extremely appealing models that allow for learning complicated distributions by taking advantage of recent progress in gradient descent algorithms and accelerated processing with GPUs. We modify the Stocastic Gradient Variational Bayes to perform posterior inference for latent spaces of multinomial distributions. We propose Dirichlet priors for a multinomial latent distribution which allows us to explore the data by interpreting the discrete probabilies as class attributes. We emperically show that our Dirichlet model learns superior representations that outperform simpler Logistic Normal models for nearest neighbours, and have similar performance in supervised settings with the Normal VAEs.

INTRODUCTION

Autoencoders use an encoder and decoder network to map between data points and latent variables (embeddings). Deep Autoencoders trained using Stochastic Gradient Variational Bayes (SGVB) (Kingma and Welling 2013) are attractive due to their ability to regularize the latent space of the neural network with probabilistic latent variables that are robust to changes to the embedding. This allows us to sample from the latent space to generate data from the posterior, infeasible in the unrestricted case.

In this report, we demonstrate how to use the Dirichlet distribution as a prior for the latent space with differentiable reparameterizations required by SGVB. By using automatic differentiation via TensorFlow (Abadi et al. 2016) we will demonstrate how these methods can be used as a preprocessing step for machine learning, model interpretation, and data visualization.

FIG. 1. Graphical Representation of Normal-VAE



Variational Autoencoder

Before we extend the Gauss-VAE, we will go over some details about VAEs. Instead of encoders and decoders, which map to single values, we can consider the decoder as a generator network with produces $g_\theta(x|z)$ and the encoder as the inference network which estimates $q_\phi(z|x)$ an approximation of the true posterior. Rather than using costly MCMC algorithms for estimation, the author (Kingma and Welling 2013) proposes a variational lower bound on the log-likelihood which allows for differentiation.

$$\mathcal{L}(\theta, \phi; x) = -D_{KL}(q(z|x)||p(z|x)) + E_{q(z|x)}\left[\log p(x|z)\right. \tag{1}$$

Since KL-Divergences can be solved analytically, $D_{KL}(q(z|x)||p(z|x))$ acts a regularizer for the inference network. Similar to how L_1 and L_2 regularization for linear models can be interpreted as having Laplacian or Normal priors over the model parameters. In conventional settings, the inference network produces a Normal distribution:

$$q(z|x) = \mathcal{N}(z|\mu(x), \sigma(x)^2 I) \tag{2}$$

While the D_{KL} term regularizes inference network, the generator network is regularized by using a sample drawn from $z \sim q(z|x)$. A reparameteriation technique is used to sample from the distribution by reframing it as adding noise to the output of the neural network.

$$z^* = \mu(x) + \sigma(x)\epsilon, \epsilon \sim \mathcal{N}(0, 1) \tag{3}$$

This way we can generate from $g(x|z*)$ and produce a complete computational graph in order for automatical differention to compute gradients.

Motivation

In many unsupervised situations a Normal distribution for the latent variable does not make much sense. Often times, these methods are used when the postiror is intractable because we deal with data that is complex, for example most benchmarks for unsupervixed learning use images of faces, plants, or animals. In this situation while the generative properies of the VAE are quite impressive. The underlying embedding is not well interpreted. However, when we do look for models with high levels of interpretibility we think of models more structured than a multivariate Normal.

Dirichlet Priors

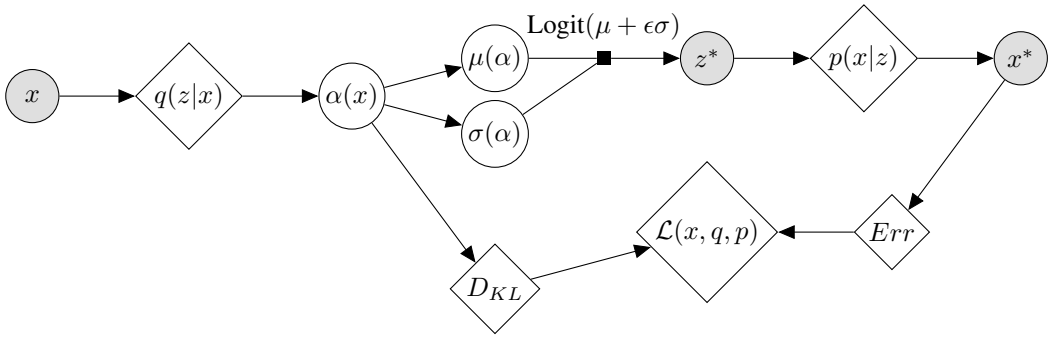
Dirichlet distributions are often used as priors over proportional data. Latent Dirichlet Allocation, for example, suggests that there are n topics and assigns a $P(\theta_i)$ of a data point being on a certain topic. Many models like soft-clustering or matrix factorization seek to find underlying categories and assigns proportions to each category. The Dirichlet Prior for our latent space is used in a similar manner to learn non-linear mappings to a mixture of classes. The difficulty lies in that we cannot reparameterize the Dirichlet for sampling the same way that we can with a Normal distribution. herefor we must find an alternative way to reparametrize the sample.

METHODOLOGY

Logitstic-Normal Approximation

An alternative for the Dirichlet distribution would be a Logistic-Normal generated by applying the logistic function $\text{Logit}(x) = \frac{\exp(x)}{\sum \exp(x)}$ to samples from a Normal.

Dir-VAE with Logit-Normal Sampling



Min KL-Divergence Approximation

One method of approximating a dirichlet is to find the Logistic-Normal with μ, σ as a function of α by minimizing the KL-divergence. The solution can be found in (Aitchison and Shen 1980).

D_{KL} Optimal Reparameterization

$$\min D_{KL}(p,q) = \min \int p(x|\mu,\Sigma) \log\left(\frac{p(x|\mu,\Sigma)}{q(x|\alpha)}\right) dx$$

Solution:

$$\mu_i = \psi(\alpha_i) - \psi(\alpha_k)$$
$$\sigma i = \sqrt{\psi'(\alpha_i) - \psi'(\alpha_k)}$$

Laplace Approximation

Minimizing KL-Divergence is not the only method of reparameteriztion. An alternative approach from (Hennig et al. 2012) constructs a Laplace approximation to the Dirichlet distribution.

Laplace Approximated Reparameterization

$$\mu_i = \log \alpha_i - \frac{1}{K}\sum \log \alpha$$
$$\sigma_i = \frac{1}{\alpha_i}\left(1 - \frac{2}{K}\right) + \frac{1}{K^2}\sum \frac{1}{\alpha}$$

Regularization

By applying any of reparameterizations in order to sample from the distribution, we only need to minimize the KL-Divergence between two Dirichlet with parameters $P(x|\alpha), Q(x|\beta)$, which can be written analytically:

$$\begin{aligned} D_{KL}(p_\alpha||q_\beta) = &\log\Gamma(\alpha_0) - \log\Gamma(\beta_0) \\ &- \sum^K \log\Gamma(\alpha_k) + \sum^K \log\Gamma(\beta_k) \\ &+ \sum^K (\alpha_k - \beta_k)(\psi(\alpha_k) - \psi(\alpha_0)) \end{aligned}$$

Since a symmetric Dirichlet is used as the regularizer, knowing that $\beta_i = 1/k, \beta_0 = 1$ allows us to define the minimization objective as a function of α.

$$\mathrm{R}(\alpha) = \log\Gamma(\alpha_0) - \sum^K \log\Gamma(\alpha_k) + \sum^K \alpha_k \psi(\alpha_k) - \frac{1}{K}\sum^K [\psi(\alpha_k) - \psi(\alpha_0)]$$

However, since we often use the symmetric Dirichlet as an uninformative prior for the latent space, any reparameterization for μ, σ when $\alpha_i = \frac{1}{k}$ results in an approximation $LN(0, \sigma I)$ which comes out as a unit logistic Normal since the σ term does not affect the minimization.

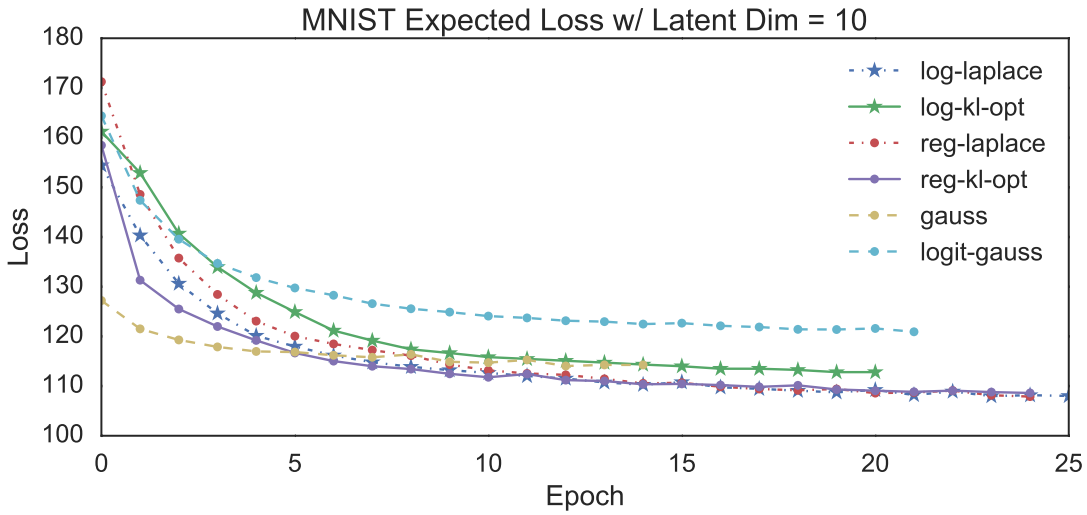


FIG. 2. Variational Autoencoder Model Performance

IMPLEMENTATION

There are many ways of implementing the multinomial code for each data point. Since neural networks consist of a sequence of non linear activations which allows us some flexibility.

The models were written in Python using Keras, a package that allows for simple specification and optimization using automatic differentiation. We implemented the Variational Autoencoder in both Normal, Logistic Normal, and Dirichlet architectures and found that while Dirichlet models were more complex in the number of operations, for the MNIST dataset, 60000 data points with 784 features showed no difference between implementations.

Activations and Reparameterizations

One choice of activation would be the scaled softmax where $\alpha_i = c\frac{\exp(h_i)}{\sum^K exp(h_k)}$, where h is the output of the hidden layer and c is a learned scaling parameter. Another is to use a centered activation like the $\tanh(x)$ to model the log of alphas. Where $\alpha_i = \exp(c\tanh(h_i))$. After using log or regular activations to generate alphas we then have the D_{KL} or the Laplace reparameterization to a logistic-Normal.

The figure above shows that various methods perform similar with laplace approximations performing the best.

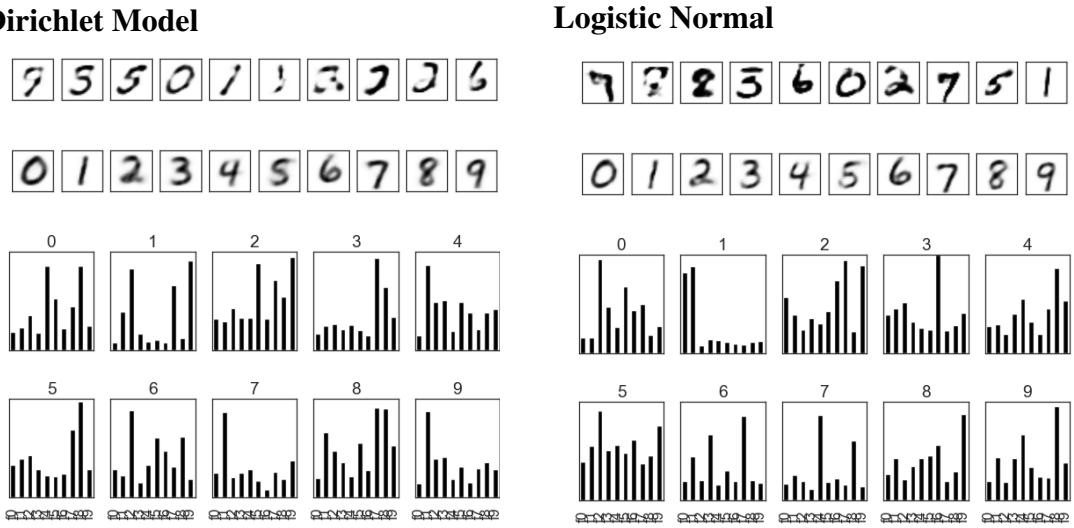


FIG. 3. Top: Reconstruction from single component, Middle: Average Digit Reconstruction, Bottom: Embedding for average digit

FIG. 4. Classification Error (%) w/ Supervised Methods

	K Nearest Neighbours			Linear Models	
	k=2	k=5	k=10	LR	LDA
Dirichlet Laplace	$9_{\pm 1.9}$	$7_{\pm 2.2}$	$7_{\pm 2.0}$	$12_{\pm 3.2}$	$14_{\pm 3.4}$
Dirichlet KL	$12_{\pm 1.8}$	$9_{\pm 2.0}$	$9_{\pm 2.3}$	$15_{\pm 3.1}$	$15_{\pm 3.3}$
Logistic Normal	$11_{\pm 1.4}$	$9_{\pm 1.5}$	$8_{\pm 1.7}$	$18_{\pm 1.9}$	$10_{\pm 2.4}$
Normal	$7_{\pm 1.8}$	$5_{\pm 2.0}$	$5_{\pm 1.6}$	$13_{\pm 2.9}$	$14_{\pm 2.7}$

RESULTS

When inspecting (Fig 3) the image reconstructions for average digits in both the Dirichlet and Logistic models come up with meaningful results. The embeddings for (4,7,9) are similar along with (2,3,5,8) as they have similar characteristics. In the appendix, there are examples of transitions from each model to qualitatively judge the quality of the latent space reconstructions.

In a supervised setting, the performance difference between the Dirichlet vs the logistic models become more apparent. To evaluated the latent space, we use K-nearest neighbors, logistic regression (LR), and multiclass linear discriminant analysis (LDA) as simple models to benchmark how well the latent space captures the data. As for the metric in K-Nearest Neighbors, we use euclidian distance for the Normal VAE and we use Hellinger distance for the rest since its is more data specific and is used to compare probabilities.

As we see from (Fig 4) The Dirichlet model (with Laplace Approximations) consistently outperforms the Logistic Normal model for K-Nearest neighbor and is competitive with the Normal VAE for supervised tasks allowing us to obtain comparable prediction accuracy with more interpretability.

DISCUSSION

In this report, we demonstrated that enforcing the latent space of a variational autoencoder to be Dirichlet vs Normal distributed allows us to generated embeddings that are both more interpretable while still being competitive in prediction and information retrieval. Rather than using a logistic Normal, which empirically performed worse, we are able to us the Laplace approximation in order to use a Dirichlet distribution in the model while taking advantage of the reparameterization trick in order to apply automatic differentiation to the samples. The Dirichlet Laplace VAE performs better than the Logistic model without incurring any extra computational costs and produces a smoother latent space (see appendix).

Future Work

This work was inspired by recent papers published in the International Conference on Learning Representations. (Nalisnick and Smyth 2016) uses a Stick-Breaking prior rather than a Dirichlet Process since they could not construct a differentable reparameterization, instead, they use the Kumaraswamy distribution as an approximation which was shown to outperform the Normal VAE. (Jang et al. 2016) proposes another reparametrization called the Gumbel-Softmax which allows for differentiable sampling from a categorical distribution. Future work would go in the direction of sparity and more model inference from the embedding.

REFERENCES

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y., and Zheng, X. (2016). "Tensorflow: A system for large-scale machine learning." *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation*, OSDI'16, Berkeley, CA, USA, USENIX Association, 265–283.

Aitchison, J. and Shen, S. M. (1980). "Logistic-normal distributions: Some properties and uses." *Biometrika*, 67(2), 261–272.

Hennig, P., Stern, D. H., Herbrich, R., and Graepel, T. (2012). "Kernel topic models.." *AISTATS*, 511–519.

Jang, E., Gu, S., and Poole, B. (2016). "Categorical reparameterization with gumbel-softmax." *Proceedings of the 5rd International Conference on Learning Representations (ICLR)*.

Kingma, D. P. and Welling, M. (2013). "Auto-encoding variational bayes." *Proceedings of the 2nd International Conference on Learning Representations (ICLR)*, number 2014.

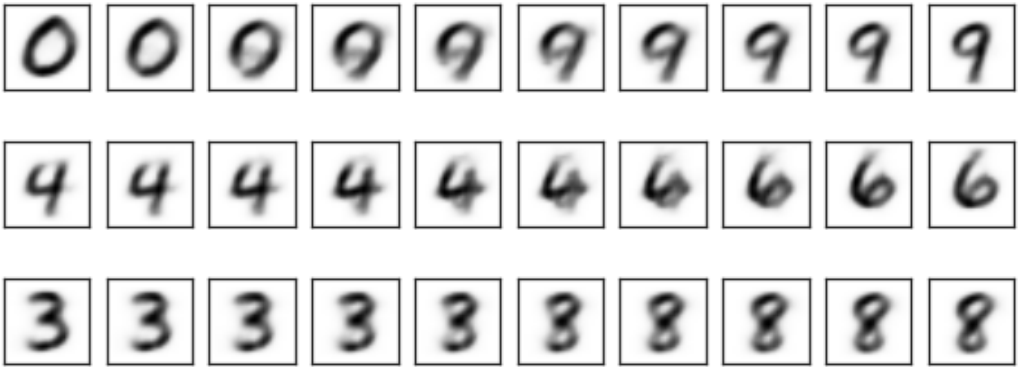
Nalisnick, E. and Smyth, P. (2016). "Stick-breaking variational autoencoders." *Proceedings of the 5rd International Conference on Learning Representations (ICLR)*.

APPENDIX I. CODE

All code can be found in the following location: `https://github.com/jxnl/dirichlet-vae`

APPENDIX II. TRANSITIONS

Dirichlet Laplace



Logistic Normal

Normal