A PROBABILISTIC SEMI-SUPERVISED APPROACH WITH TRIPLET MARKOV CHAINS

Anonymous

Anonymous

ABSTRACT

In this article, xxx

***Index Terms*—** Generative Models; Variational Inference; Semi-Supervised Learning; Triplet Markov Chains.

1. INTRODUCTION

This paper focuses on semi-supervised learning for bayesian classification in general generative models.

1.1. Sequential Bayesian classification

We denote as $\mathbf{x}_T = (x_0, \ldots, x_T)$ a sequence of observed random variables (r.v.) and $\mathbf{z}_T = (z_0, \ldots, z_T)$ a sequence of latent r.v. We also introduce a sequence of labels $\mathbf{y}_T = (y_0, \ldots, y_T)$ associated to the previous sequence $\mathbf{x}_T$. We will assume that $x_t \in \mathbb{R}^{d_x}$, $z_t \in \mathbb{R}^{d_z}$ while the label y_t is discrete, so $y_t \in \Omega = \{\omega_1, \ldots, \omega_C\}$. As far as notations are concerned, we do not distinguish r.v. and their realizations. For example, $\mathbf{x}_T$ can represent a noisy grayscale image while $\mathbf{y}_T$ represents the original black and white image. For this application, one can assume that the latent variable z_t can govern the distribution of the noise.

When the labels associated to $\mathbf{x}_T$ are not observed, the objective associated to Bayesian classification consists in computing, for all t, the posterior distributions

$$p(y_t|\mathbf{x}_T) = \frac{\sum_{\mathbf{y}_{0:t-1}, \mathbf{y}_{t+1:T}} \int p(\mathbf{y}_T, \mathbf{x}_T, \mathbf{z}_T) \mathrm{d}\mathbf{z}_T}{\sum_{\mathbf{y}_T} \int p(\mathbf{y}_T, \mathbf{x}_T, \mathbf{z}_T) \mathrm{d}\mathbf{z}_T}. \quad (1)$$

Consequently, we first need to define a parameterized model $p_\theta(\mathbf{x}_T, \mathbf{y}_T, \mathbf{z}_T)$ which aims at describing the r.v. involved in the problem and from which it is possible to estimate θ and next to compute (1) in a reasonable computational cost.

The estimation of θ (i.e. the learning step) can be realized from sequences where we have at our disposal $(\mathbf{x}_T, \mathbf{y}_T)$ (supervised learning) or only $\mathbf{x}_T$ (unsupervised learning). This general problem is commonly used in many fields, such as speech recognition XX, natural language processing XX, and activity recognition XX, etc.

Anonymous.

1.2. Semi-Supervised learning

The problem we consider in this paper is a little bit different. In many real-world applications, it is expensive or impossible to obtain labels for the entire sequence due to various reasons such as the high cost of labeling, the lack of expertise, or the lack of time. So from now on, we assume that we have at our disposal sequences with partially observed labels and that i) we want to train generative models and ii) we need to estimate the missing labels associated to each sequences. In other words, if we decompose the sequence of labels $\mathbf{y}_T$ as

$$\mathbf{y}_T = \mathbf{y}_{T^\mathcal{L}} \cup \mathbf{y}_{T^\mathcal{U}},$$

where $\mathbf{y}_{T^\mathcal{L}} = \{y_t\}_{t\in\mathcal{L}}$ (resp. $\mathbf{y}_{T^\mathcal{U}} = \{y_t\}_{t\in\mathcal{U}}$) denotes the observed (resp. the unobserved) labels, we now look for estimating θ from $(\mathbf{x}_T, \mathbf{y}_{T^\mathcal{L}})$, and next compute, for all $t \in \mathcal{U}$,

$$p(y_t|\mathbf{x}_T, \mathbf{y}_{T^\mathcal{L}}) = \frac{\sum_{y_s \mathrm{s.t.} s\in\mathcal{U}\setminus\{t\}} \int p(\mathbf{y}_T, \mathbf{x}_T, \mathbf{z}_T) \mathrm{d}\mathbf{z}_T}{\sum_{y_s \mathrm{s.t.} s\in\mathcal{U}} \int p(\mathbf{y}_T, \mathbf{x}_T, \mathbf{z}_T) \mathrm{d}\mathbf{z}_T}.$$

Sequential data is characterized by a temporal ordering of observations.

In particular, generative models are a popular approach for semi-supervised learning, as they allow the incorporation of prior knowledge about the data distribution. In the case of sequential data, probabilistic generative models can be used to model the underlying distribution of the data and generate new samples, make predictions outcomes, and estimate missing or unobserved data.

TMC [1, 2]

1.3. Scope of the paper

Since only a subset of the observations has corresponding labels, the sequence of labels can be expressed as $\mathbf{y}_T = \mathbf{y}_{T^\mathcal{L}} \cup \mathbf{y}_{T^\mathcal{U}}$, where $\mathbf{y}_{T^\mathcal{L}} = \{y_t\}_{t\in\mathcal{L}}$ and $\mathbf{y}_{T^\mathcal{U}} = \{y_t\}_{t\in\mathcal{U}}$ are the sets of observed labels and unobserved (hidden) labels, respectively; and $\mathcal{L}$ and $\mathcal{U}$ are the set of labeled and unlabeled time steps, respectively.

In this article, our interest is to present a general generative model for semi-supervised learning, based on latent r.v. which is defined by a joint distribution $p_\theta(\mathbf{x}_T, \mathbf{y}_T, \mathbf{z}_T)$ and provides learning from the observations $\mathbf{x}_T$ and the observed labels $\mathbf{y}_{T^\mathcal{L}}$, since the distribution reads

979-8-3503-2411-2/23/$31.00 ©2023 IEEE

$$p_\theta(\mathbf{x}_T, \mathbf{y}_{T^\mathcal{L}}) = \sum_{y_t \in \mathbf{y}_{T^\mathcal{U}}} \int p_\theta(\mathbf{x}_T, \mathbf{y}_T, \mathbf{z}_T) \mathrm{d}\mathbf{z}_T. \quad (2)$$

Our model is based on the TMC [1, 2] which relies on the assumption that the triplet $(z_t, y_t, x_t)_{t\geq 0}$ is a Markov chain with transition $p(z_t, y_t, x_t | z_{t-1}, y_{t-1}, x_{t-1})$.

2. PROPOSED METHOD

This section presents a review of variational Bayesian inference and its application to general Triplet Markov Chains for semi-supervised learning.

2.1. Background: Variational Inference

The aim is to estimate the parameter θ from a realization $\mathbf{x}_T$. A popular estimate is the Maximum-Likelihood (ML) estimate $\hat{\theta} = \arg\max_\theta p_\theta(\mathbf{x}_T)$ due to its statistical properties [3, 4]. However, a direct maximization of $p_\theta(\mathbf{x}_T)$ is not always possible, particularly in models with latent variables where the likelihood $p_\theta(\mathbf{x}_T) = \int p_\theta(\mathbf{x}_T, \mathbf{z}_T) \mathrm{d}\mathbf{z}_T$ may be not computable. In the variational inference framework, a variational lower bound called evidence lower bound (ELBO) on the log-likelihood is optimized in order to estimate the parameters θ. This variational lower bound relies on the introduction of a parameterized variational distribution $q_\phi(\mathbf{z}_T | \mathbf{x}_T)$, which is parameterized by a set of parameters ϕ, and it is given by:

$$\log(p_\theta(\mathbf{x}_T)) \geq \tilde{Q}(\theta, \phi),$$
$$\tilde{Q}(\theta, \phi) = -\int \log\left(\frac{q_\phi(\mathbf{z}_T|\mathbf{x}_T)}{p_\theta(\mathbf{x}_T, \mathbf{z}_T)}\right) q_\phi(\mathbf{z}_T|\mathbf{x}_T) \mathrm{d}\mathbf{z}_T. \quad (3)$$

Equality holds if $q_\phi(\mathbf{z}_T \mathbf{x}_T) = p_\theta(\mathbf{z}_T|\mathbf{x}_T)$. When the posterior distribution $p_\theta(\mathbf{z}_T|\mathbf{x}_T)$ is computable, the alternating maximization w.r.t. θ and q_ϕ of the ELBO, $\tilde{Q}(\theta, \phi)$, coincides with the EM algorithm [5].

Variational inference consists in maximizing $Q(\theta, \phi)$ with respect to (θ, ϕ) for a given class of distributions q_ϕ. The choice of the variational distribution $q_\phi(\mathbf{z}_T|\mathbf{x}_T)$ is important; $q_\phi(\mathbf{z}_T|\mathbf{x}_T)$ should be close to $p_\theta(\mathbf{z}_T|\mathbf{x}_T)$ but should also be chosen in a such way that the associated ELBO can be exactly computed or easily approximated while remaining differentiable w.r.t. (θ, ϕ). A simple way to approximate $\tilde{Q}(\theta, \phi)$ is by using the reparametrization trick [6] which consists in chosing a parametric distribution $q_\phi(\mathbf{z}_T|\mathbf{x}_T)$ such that a sample $\mathbf{z}_T^{(i)} \sim q(\mathbf{z}_T|\mathbf{x}_T)$ can be written as a differentiable function of ϕ.

It remains to choose a variational distribution $q_\phi(\mathbf{z}_T|\mathbf{x}_T)$. One popular strategy is to use distributions from a mean-field variational distribution which factorizes as $q_\phi(\mathbf{z}_T|\mathbf{x}_T) = q(z_0|\mathbf{x}_T) \prod_{t=1}^t q_\phi(z_t|z_{0:t-1}, \mathbf{x}_T)$.

2.2. Semi-supervised Variational Inference for TMC

The distribution $p_\theta(\mathbf{x}_T, \mathbf{y}_T, \mathbf{z}_T)$ reads

$$p_\theta(z_0, y_0, x_0) \prod_{t=1}^T p_\theta(z_t, y_t, x_t | z_{t-1}, y_{t-1}, x_{t-1}). \quad (4)$$

Since only a subset of labels is observed $\mathbf{y}_{T^\mathcal{L}}$, the set of unobserved labels $\mathbf{y}_{T^\mathcal{U}}$ is treated as latent variables and variational inference involves finding a lower bound on the marginal likelihood of the observed data $\mathbf{x}_T$ and $\mathbf{y}_{T^\mathcal{L}}$. The variational lower bound is given by:

$$\begin{aligned}\log(p_\theta(\mathbf{x}_T, \mathbf{y}_{T^\mathcal{L}})) \geq &- \int \sum_{\mathbf{y}_{T^\mathcal{U}}} q_\phi(\mathbf{z}_T, \mathbf{y}_{T^\mathcal{U}} | \mathbf{x}_T, \mathbf{y}_{T^\mathcal{L}}) \times \\ &\log\left(\frac{q_\phi(\mathbf{z}_T, \mathbf{y}_{T^\mathcal{U}}|\mathbf{x}_T)}{p_\theta(\mathbf{z}_T, \mathbf{y}_T, \mathbf{x}_T)}\right) \mathrm{d}\mathbf{z}_T \\ &= Q(\theta, \phi), \end{aligned} \quad (5)$$

where ϕ denotes the parameters of the variational distribution $q_\phi(\mathbf{z}_T, \mathbf{y}_{T^\mathcal{U}} | \mathbf{x}_T, \mathbf{y}_{T^\mathcal{L}})$.

Since our model has both discrete and continuous latent variables, the approximation of the ELBO in Eq. (5) becomes more complex. To that end, we can use the Gumbel-Softmax (G-S) trick [7, 8] and the reparametrization trick [6] to approximate $Q(\theta, \phi)$ simustaneously. For the continuous latent variables $\mathbf{z}_T$, the reparametrization trick introduced in section 2.1 is still valid. On the other hand, for the discrete latent variables $\mathbf{y}_{T^\mathcal{U}}$, the G-S trick enables to obtain a differentiable approximation to the discrete categorical distribution.

It remains to choose a factorization of the variational distribution $q_\phi(\mathbf{z}_T, \mathbf{y}_{T^\mathcal{U}} | \mathbf{x}_T, \mathbf{y}_{T^\mathcal{L}})$, which is a crucial step in variational inference. Different models can be obtained by choosing different factorizations of the variational distribution, which will be discussed in the next section 2.3. For example, we can first consider $q_\phi(\mathbf{z}_T, \mathbf{y}_{T^\mathcal{U}} | \mathbf{x}_T, \mathbf{y}_{T^\mathcal{L}}) = q_\phi(\mathbf{z}_T | \mathbf{x}_T, \mathbf{y}_T) q_\phi(\mathbf{y}_{T^\mathcal{U}} | \mathbf{x}_T, \mathbf{y}_{T^\mathcal{L}})$ and then consider a mean-field variational distribution.

I think we should add a paragraph here to explain what we do in the case of y is missing, we replace it by a sample .... — Katy

2.3. Particular cases of the TMC

The choice of the factorization of the transition distribution $p_\theta(z_t, y_t, x_t | z_{t-1}, y_{t-1}, x_{t-1})$ has an impact on the performance of the model for a specific task (classification, prediction, detection or generation). The goal of this section is to present particular cases of our proposed TMC model by choosing different factorizations of the transition and variational distribution. For the sake of clarity, let us now denote the triplet $v_t = (x_t, z_t, y_t)$.

2.3.1. VSL

Variational Sequential Labelers model is a particular case of the TMC model where the transition distribution is factorized as follows:

$$p_\theta(v_t|v_{t-1}) = p_\theta(y_t|z_t)p_\theta(z_t|x_{t-1}, y_{t-1})p_\theta(x_t|z_t). \quad (6)$$

In the case of SVRNN, the chosen variational distribution $q_\phi(\mathbf{z}_T, \mathbf{y}_{T^\mathcal{U}}|\mathbf{x}_T, \mathbf{y}_{T^\mathcal{L}}) = q_\phi(\mathbf{z}_T|\mathbf{x}_T, \mathbf{y}_{T^\mathcal{L}})q_\phi(\mathbf{y}_{T^\mathcal{U}}|\mathbf{x}_T, \mathbf{z}_T, \mathbf{y}_{T^\mathcal{L}})$ is given by

$$q_\phi(\mathbf{z}_T|\mathbf{x}_T, \mathbf{y}_{T^\mathcal{L}}) = \prod_{t=0}^{T} q_\phi(z_t|\mathbf{x}_T)$$

$$q_\phi(\mathbf{y}_{T^\mathcal{U}}|\mathbf{x}_T, \mathbf{z}_T, \mathbf{y}_{T^\mathcal{L}}) = \prod_{t\in\mathcal{U}} q_\phi(y_t|z_t)$$

In particular, VLS considers $q_\phi(y_t|z_t) = p_\theta(y_t|z_t)$.

2.3.2. SVRNN

A Semi-supervised Variational Recurrent Neural Network can be seen as an adapted version of the TMC model where the latent variable z_t is set as the pair (z'_t, h_t). The variable z'_t is a stochastic latent variable and h_t is deterministically given by $h_t = f_\theta(x_t, y_t, z'_t)$, where f_θ is a deterministic non-linear transition function. In ths case, the transition distribution is factorized as follows:

$$p_\theta(v_t|v_{t-1}) = p_\theta(y_t|v_{t-1})p_\theta(z_t|y_t, v_{t-1})p_\theta(x_t|y_t, z_t, v_{t-1}). \quad (7)$$

On the other hand, the chosen variational distribution $q_\phi(\mathbf{z}_T, \mathbf{y}_{T^\mathcal{U}}|\mathbf{x}_T, \mathbf{y}_{T^\mathcal{L}}) = q_\phi(\mathbf{z}_T|\mathbf{x}_T, \mathbf{y}_T)q_\phi(\mathbf{y}_{T^\mathcal{U}}|\mathbf{x}_T, \mathbf{y}_{T^\mathcal{L}})$ factorizes as follows:

$$q_\phi(\mathbf{z}_T|\mathbf{x}_T, \mathbf{y}_T) = q_\phi(z_0|\mathbf{x}_t, \mathbf{y}_t)\prod_{t=1}^{T} q_\phi(z_t|\mathbf{z}_{t-1}, \mathbf{x}_t, \mathbf{y}_t)$$

$$q_\phi(\mathbf{y}_{T^\mathcal{U}}|\mathbf{x}_T, \mathbf{y}_{T^\mathcal{L}}) = \prod_{t\in\mathcal{U}} q_\phi(y_t|\mathbf{y}_{t-1}, \mathbf{x}_t),$$

2.3.3. mTMC

Finally, we present the minimal TMC which is a particular case of the TMC model presented in [9] where the transition distribution is factorized as follows:

$$p_\theta(v_t|v_{t-1}) = p_\theta(y_t|y_{t-1})p_\theta(z_t|z_{t-1})p_\theta(x_t|t_t, z_t). \quad (8)$$

The variational distribution is chosen as in the SVRNN model.

3. SIMULATIONS

In this section, we present the results of the proposed models on semi-supervised binary image segmentation. Our goal is to recover the segmentation of a binary image ($\Omega = \{\omega_1, \omega_2\}$) from the noisy observations $\mathbf{x}_T$ when a partially segmentation $\mathbf{y}_{T^\mathcal{L}}$ is available.

3.1. Deep TMCs

The set of parameters (θ, ϕ) can be described by any differentiable flexible function $\psi(\cdot)$. In particular, we consider the case where the parameters are produced by a (deep) neural network.

Due to the different factorizations of the generating (resp. variational) distributions, we consider a general notation $p_\theta(x_t|\cdot)$, $p_\theta(z_t|\cdot)$ and $p_\theta(y_t|\cdot)$ (resp. $q_\phi(y_t|\cdot)$, $q_\phi(z_t|\cdot)$) in order to avoid presenting a specific dependence between variables. These dependencies are specified for each model and are presented in the previous Section 2.3. The general model is described by:

$$p_\theta(v_t|v_{t-1}) = p_\theta(x_t|\cdot)p_\theta(z_t|\cdot)p_\theta(y_t|\cdot)$$
$$p_\theta(x_t|\cdot) = \mathcal{N}(x_t; \mu_{px,t}, \mathrm{diag}(\sigma_{px,t})) \quad (9)$$
$$p_\theta(z_t|\cdot) = \mathcal{N}(x_t; \mu_{pz,t}, \mathrm{diag}(\sigma_{pz,t})) \quad (10)$$
$$p_\theta(y_t|\cdot) = \mathcal{B}er(y_t; \rho_{py,t}), \quad (11)$$

where $\mathrm{diag}(.)$ denotes the diagonal matrix deduced from the values of $\sigma_{.,t}$; and $[\mu_{.,t}, \sigma_{.,t}]$ and $\rho_{.,t}$ denote the parameters of the Gaussian and Bernoulli distributions, respectively. Finally, the variational distribution is given by

$$q_\phi(z_t|\cdot) = \mathcal{N}(z_t; \mu_{qz,t}, \mathrm{diag}(\sigma_{qz,t})) \quad (12)$$
$$q_\phi(y_t|\cdot) = \mathcal{B}er(y_t; \rho_{qy,t}), \quad (13)$$

The parameters $\theta = \{\mu_{pz,t}, \sigma_{pz,t}, \mu_{px,t}, \sigma_{px,t}, \rho_{py,t}\}$ and $\phi = \{\mu_{qz,t}, \sigma_{qz,t}, \rho_{qy,t}\}$ can be derived from neural networks $\psi_.(\cdot)$. For example, the parameters of the deep mTMC model are given by

$$[\mu_{px,t}, \sigma_{px,t}] = \psi_{px}(y_t, z_t)$$
$$[\mu_{qz,t}, \sigma_{qz,t}] = \psi_{qz}(\mathbf{z}_{t-1}, \mathbf{x}_t, \mathbf{y}_t),$$
$$[\mu_{pz,t}, \sigma_{pz,t}] = \psi_{pz}(z_{t-1}),$$
$$\rho_{py,t} = \psi_{px}(y_{t-1})$$
$$\rho_{qy,t} = \psi_{qy}(\mathbf{y}_{t-1}, \mathbf{x}_t).$$

Note that in the VLS model, ψ_{qy} is not necessary since the assumption $q_\phi(y_t|z_t) = p_\theta(y_t|z_t)$ is made.

Explicar la incorporacion de h_t en el modelo SVRNN y como fue configurada en los otros modelos. — Katy

3.2. Experiments settings

We consider the cattle-type and the camel-type images of the Binary Shape Database [10]. The images are transformed into a 1-D signal $\mathbf{x}_T$ with a Hilbert-Peano filling curve [11]. They are next blurred with non-linear noises to highlight the ability, of the models presented in Section 2.3, to learn such a signal corruption. In fact, the cattle-type image is blurred with a general stationary noise and the camel-type image is blurred with a stationary multiplicative noise. More details about generation are given in [9]. On the other hand, the pixels $y_t \in \mathbf{y}_{T^\mathcal{L}}$

are randomly chosen. In particular, we consider that 60% of pixels are labeled, the rest of pixels are considered as unobserved.

Each model was trained with stochastic gradient descent on the negative associated ELBO using the Adam optimizer [12]. Additionally, we match the total number of parameters of all models to be equal or close between them, so the number of hidden units is different for each model. The SVRNN (resp. mTMC and VLS) model has XX (resp. XX and XX) hidden units.

3.3. Results

3.4. Discussion

4. CONCLUSION

5. REFERENCES

[1] W. Pieczynski, "Chaines de Markov triplet," *Comptes Rendus de l'Academie des Sciences - Mathematiques*, vol. 335, pp. 275–278, 2002, in French.

[2] W. Pieczynski and F. Desbouvries, "On triplet Markov chains," in *International Symposium on Applied Stochastic Models and Data Analysis, (ASMDA)*, 2005.

[3] H. White, "Maximum likelihood estimation of misspecified models," *Econometrica*, vol. 50, no. 1, pp. 1–25, January 1982.

[4] R. Douc and E. Moulines, "Asymptotic properties of the maximum likelihood estimation in misspecified hidden Markov models," *Annals of Statistics*, vol. 40, no. 5, pp. 2697–2732, 2012.

[5] D. G. Tzikas, A. C. Likas, and N. P. Galatsanos, "The variational approximation for Bayesian inference," *IEEE Signal Processing Magazine*, vol. 25, no. 6, pp. 131–146, 2008.

[6] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *2nd International Conference on Learning Representations, ICLR*, 2014.

[7] Chris J Maddison, Andriy Mnih, and Yee Whye Teh, "The concrete distribution: A continuous relaxation of discrete random variables," *arXiv preprint arXiv:1611.00712*, 2016.

[8] Eric Jang, Shixiang Gu, and Ben Poole, "Categorical reparameterization with gumbel-softmax," *arXiv preprint arXiv:1611.01144*, 2016.

[9] H. Gangloff, K. Morales, and Y. Petetin, "Deep parameterizations of pairwise and triplet markov models for unsupervised classification of sequential data," *Computational Statistics & Data Analysis*, vol. 180, pp. 107663, 2023.

[10] LEMS-Computer Vision Group, "Binary shape," *https://vision.lems.brown.edu/content/available-software-and-databases*.

[11] H. Sagan, *Space-filling curves*, Springer, 2012.

[12] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *International conference on learning representations*, 12 2014.