

SVAE objective

July 27, 2017

1 Definition of the model

Define the following model:

$$\begin{aligned} z_n | \theta &\sim \text{discrete}(\pi^0) \\ x_n | z_n, \theta &\sim \mathcal{N}(\mu_{z_n}^0, \Sigma_{z_n}^0) \\ \text{where } \theta &= (\pi^0, \{\mu_k^0, \Sigma_k^0\}_{k \in [0, K]}) \\ \text{and } y_n | x_n, \gamma &\sim \text{Bernoulli}(f_\gamma(x_n), M) \end{aligned}$$

where $f_\gamma(x_n)$ is the output of a neural network parametrized by γ to the input x_n , K is the number of mixtures (10 for MNIST) and M is the dimension of the observable variables y_n (28×28 for MNIST). Now using the notation under the exponential family form, we can write:

$$p_{z|\theta}(z_n) = \exp \left(\langle \eta_z^0(\theta), t_z(z_n) \rangle - \log(Z_z(\eta_z^0(\theta))) \right)$$

where:

$$\begin{aligned} \eta_z^0(\theta)^\top &= [\ln(\pi_1^0), \dots, \ln(\pi_K^0)] \\ t_z(z_n)^\top &= [\delta(z=1), \dots, \delta(z=K)] \\ \log(Z_z(\eta)) &= 0 \quad \forall \eta \end{aligned}$$

For the mixture of gaussians part $p_{x|z, \theta}$, we can write:

$$p_{x|z, \theta}(x) = \exp \left(\langle \eta_x^0(\theta, z), t_x(x) \rangle - \log(Z_x(\eta_x^0(\theta, z))) \right)$$

where:

$$\begin{aligned} t_x(x) &= \begin{bmatrix} x \\ \text{vec}(xx^\top) \end{bmatrix} \\ \eta_x^0(\theta, z=k) &= \begin{bmatrix} \Sigma_k^{0-1} \mu_k^0 \\ \text{vec}(-\frac{1}{2} \Sigma_k^{0-1}) \end{bmatrix} \quad \forall k \in [1, K] \\ \log(Z_{x|z}(\eta_x^0(\theta, z))) &= \frac{1}{2} \ln(|2\pi \Sigma_z^0|) + \frac{1}{2} \mu_z^{0\top} \Sigma_z^{0-1} \mu_z^0 \end{aligned}$$

Using the conjugacy of the model, we can rewrite $p_{x|z,\theta}$ as follow:

$$\begin{aligned}
p_{x|z,\theta}(x) &= \frac{1}{\sqrt{|2\pi\Sigma_z^0|}} \exp\left(-\frac{1}{2}(x - \mu_z^0)^\top \Sigma_z^{0-1}(x - \mu_z^0)\right) \\
&= \prod_{k=1}^K \left[\frac{1}{\sqrt{|2\pi\Sigma_k^0|}} \exp\left(-\frac{1}{2}(x - \mu_k^0)^\top \Sigma_k^{0-1}(x - \mu_k^0)\right) \right]^{\delta(z=k)} \\
&= \exp\left(\sum_{k=1}^K \delta(z=k) \left[-\frac{1}{2} \ln(|2\pi\Sigma_k^0|) - \frac{1}{2}(x - \mu_k^0)^\top \Sigma_k^{0-1}(x - \mu_k^0) \right]\right) \\
&= \exp\left(\sum_{k=1}^K \delta(z=k) \left[-\frac{1}{2} \ln(|2\pi\Sigma_k^0|) - \frac{1}{2} \mu_k^{0\top} \Sigma_k^{0-1} \mu_k^0 + x^\top \Sigma_k^{0-1} \mu_k^0 - \frac{1}{2} x^\top \Sigma_k^{0-1} x \right]\right) \\
&= \exp\left(\sum_{k=1}^K \delta(z=k) \left[-\frac{1}{2} \ln(|2\pi\Sigma_k^0|) - \frac{1}{2} \mu_k^{0\top} \Sigma_k^{0-1} \mu_k^0 + \eta_x^0(\theta, z=k)^\top t_x(x) \right]\right) \\
&= \exp\left(\sum_{k=1}^K \delta(z=k) \left[\eta_x^0(\theta)[k]^\top \begin{bmatrix} t_x(x) \\ 1 \end{bmatrix} \right]\right)
\end{aligned}$$

and thus, we have:

$$p_{x|z,\theta}(x) = \exp\left(\left\langle t_z(z), \eta_x^0(\theta)^\top \begin{bmatrix} t_x(x) \\ 1 \end{bmatrix} \right\rangle\right) \quad (1)$$

where $\eta_x^0(\theta) = [\eta_x^0(\theta)[1], \dots, \eta_x^0(\theta)[K]]$ is a $(N + N^2 + 1) \times K$ matrix with its columns $\eta_x^0(\theta)[k] \in \mathbb{R}^{N+N^2+1}$, $\forall k \in [1, K]$, given by:

$$\eta_x^0(\theta)[k] = \begin{bmatrix} \Sigma_k^{0-1} \mu_k^0 \\ \text{vec}(-\frac{1}{2} \Sigma_k^{0-1}) \\ -\frac{1}{2} \ln(|2\pi\Sigma_k^0|) - \frac{1}{2} \mu_k^{0\top} \Sigma_k^{0-1} \mu_k^0 \end{bmatrix}$$

or, using the previous expressions,

$$\eta_x^0(\theta)[k] = \begin{bmatrix} \eta_x^0(\theta, z=k) \\ -\log(Z_{x|z=k}(\eta_x^0(\theta, z=k))) \end{bmatrix}$$

2 SVAE objective

2.1 SVAE objective Definition

Taking the respective variational factors in the corresponding exponential families, we define:

$$\begin{aligned}
q_z(z_n) &= \exp\left(\langle \eta_z, t_z(z_n) \rangle - \log(Z_z(\eta_z))\right) \\
q_x(x_n) &= \exp\left(\langle \eta_x, t_x(x_n) \rangle - \log(Z_x(\eta_x))\right)
\end{aligned}$$

We can now define the mean field objective of the problem \mathcal{L} :

$$\mathcal{L}(\eta_z, \eta_x) = \mathbb{E}_{q_z q_x} \left[\log \left(\frac{p_{z|\theta}(z_n) p_{x|z,\theta}(x_n) p_{y|x,\gamma}(y_n)}{q_z(z_n) q_x(x_n)} \right) \right]$$

In the *SVAE* algorithm, we express η_x and η_z as function of the remaining parameters using a surrogate objective $\hat{\mathcal{L}}$ and introducing a recognition network r_ϕ defined as follow:

$$\hat{\mathcal{L}}(\eta_z, \eta_x, \phi) = \mathbb{E}_{q_z q_x} \left[\log \left(\frac{p_{z|\theta}(z_n) p_{x|z, \theta}(x_n) \exp(\psi(x_n; \phi))}{q_z(z_n) q_x(x_n)} \right) \right]$$

where $\psi(x_n; \phi) = \langle r_\phi(y_n), t_x(x_n) \rangle$ and $r_\phi(y_n)$ a the output of a neural network parametrized by ϕ , the recognition network, to the input y_n .

We now partially optimize the surrogate objective $\hat{\mathcal{L}}$ w.r.t η_x and η_z defining η_x^* and η_z^* :

$$\eta_x^*(\phi), \eta_z^*(\phi) = \operatorname{argmax}_{\eta_x, \eta_z} \hat{\mathcal{L}}(\eta_z, \eta_x, \phi)$$

And finally, the SVAE objective, $\mathcal{L}_{\text{SVAE}}$ as:

$$\mathcal{L}_{\text{SVAE}}(\gamma, \theta, \phi) = \mathcal{L}(\eta_z^*(\phi), \eta_x^*(\phi))$$

Expanding the expression of $\mathcal{L}_{\text{SVAE}}$, we can write:

$$\mathcal{L}_{\text{SVAE}}(\gamma, \theta, \phi) = \mathbb{E}_{q_x^*} [\log (p_{y|x, \gamma}(y_n))] - KL(q_x^* q_z^* || p_{z|\theta} p_{x|z, \theta}) \quad (2)$$

2.2 SVAE objective derivation

The first term of the *R.H.S* of the equation can be computed using the reparametrization trick, sampling \hat{x} from q_x^* , $\hat{x} \sim q_x^*$, and using the following approximation:

$$\mathbb{E}_{q_x^*} [\log (p_{y|x, \gamma}(y_n))] \approx \log (p_{y|x, \gamma}(y_n | \hat{x}, \gamma)) \quad (3)$$

We will denote the last term of the *R.H.S* as the *local meanfield* term of the objective. Using the definition of the *KL*, we can expand the *local meanfield* as:

$$KL(q_z^* q_x^* || p_{z|\theta} p_{x|z, \theta}) = KL(q_z^* || p_{z|\theta}) + \mathbb{E}_{q_z^*} [KL(q_x^* || p_{x|z, \theta})]$$

Using the result for the *KL* in the exponential family case, we have:

$$\begin{aligned} KL(q_z^* || p_{z|\theta}) &= \left\langle \eta_z^*(\phi) - \eta_z^0(\theta), \mathbb{E}_{q_z^*} [t_z(z)] \right\rangle \\ &\quad - \left(\log (Z_z(\eta_z^*(\phi))) - \log (Z_z(\eta_z^0(\theta))) \right) \end{aligned}$$

and

$$\begin{aligned} \mathbb{E}_{q_z^*} [KL(q_x^* || p_{x|z, \theta})] &= \left\langle \eta_x^*(\phi) - \mathbb{E}_{q_z^*} [\eta_x^0(\theta, z)], \mathbb{E}_{q_x^*} [t_x(x)] \right\rangle \\ &\quad - \left(\log (Z_x(\eta_x^*(\phi))) - \mathbb{E}_{q_z^*} [\log (Z_x(\eta_x^0(\theta, z)))] \right) \end{aligned}$$

Thus, we can rewrite our *local meanfield* as:

$$\begin{aligned} KL(q_z^* q_x^* || p_{z|\theta} p_{x|z, \theta}) &= \left[\left\langle \eta_x^*(\phi) - \mathbb{E}_{q_z^*} [\eta_x^0(\theta, z)], \mathbb{E}_{q_x^*} [t_x(x)] \right\rangle - \log (Z_x(\eta_x^*(\phi))) \right] \\ &\quad + \left[\left\langle \eta_z^*(\phi) - \eta_z^0(\theta), \mathbb{E}_{q_z^*} [t_z(z)] \right\rangle + \mathbb{E}_{q_z^*} [\log (Z_x(\eta_x^0(\theta, z)))] \right] \\ &\quad - \left(\log (Z_z(\eta_z^*(\phi))) - \log (Z_z(\eta_z^0(\theta))) \right) \end{aligned} \quad (4)$$

3 Block coordinate ascent algorithm

So it remains to compute the partial optimum of our surrogate objective, $\eta_x^*(\phi)$ and $\eta_z^*(\phi)$ as functions of the other parameters. Using the classical result for meanfield, we have, omitting everything constant *w.r.t* x in the \propto :

$$\begin{aligned} q_x^*(x; \theta, \phi) &\propto \exp \left(\mathbb{E}_{q_z^*} [\log (p_{z|\theta}(z) p_{x|z,\theta}(x) \exp(\psi(x; \phi))) \right] \\ &\propto \exp \left(\mathbb{E}_{q_z^*} [\langle \eta_x^0(\theta, z), t_x(x) \rangle + \langle r_\phi(y), t_x(x) \rangle] \right) \\ &\propto \exp \left(\langle \mathbb{E}_{q_z^*} [\eta_x^0(\theta, z)] + r_\phi(y), t_x(x) \rangle \right) \end{aligned}$$

and thus:

$$\eta_x^*(\phi) = \mathbb{E}_{q_z^*} [\eta_x^0(\theta, z)] + r_\phi(y) \quad (5)$$

In the same way, we can express $\eta_z^*(\phi)$ as a function of θ and ϕ , omitting everything constant *w.r.t* x in the \propto :

$$q_z^*(z; \theta, \phi) \propto \exp \left(\mathbb{E}_{q_x^*} [\log (p_{z|\theta}(z) p_{x|z,\theta}(x) \exp(\psi(x; \phi))) \right]$$

Using equation (1), we have, keeping only what depends of z :

$$\begin{aligned} q_z^*(z; \theta, \phi) &\propto \exp \left(\mathbb{E}_{q_x^*} [\langle \eta_z^0(\theta), t_z(z) \rangle + \langle t_z(z), \eta_x^0(\theta)^\top \begin{bmatrix} t_x(x) \\ 1 \end{bmatrix} \rangle] \right) \\ &\propto \exp \left(\langle \mathbb{E}_{q_x^*} [\eta_x^0(\theta)^\top (t_x(x), x)] + \eta_z^0(\theta), t_z(z) \rangle \right) \end{aligned}$$

and thus:

$$\eta_z^*(\phi) = \mathbb{E}_{q_x^*} [\eta_z^0(\theta)^\top \begin{bmatrix} t_x(x) \\ 1 \end{bmatrix}] + \eta_z^0(\theta) \quad (6)$$

We can inject the expressions (3) and (6) in (4) to get:

$$\begin{aligned} KL(q_z^* q_x^* || p_{z|\theta} p_{x|z,\theta}) &= \left[\left\langle r_\phi(y), \mathbb{E}_{q_x^*} [t_x(x)] \right\rangle - \log (Z_x(\eta_x^*(\phi))) \right] \\ &\quad + \left[\left\langle \mathbb{E}_{q_z^*} [\eta_x^0(\theta, z)], \mathbb{E}_{q_x^*} [t_x(x)] \right\rangle \right. \\ &\quad \left. - \left(\log (Z_z(\eta_z^*(\phi))) - \log (Z_z(\eta_z^0(\theta))) \right) \right] \end{aligned} \quad (7)$$

We denote the first term of the *RHS* of (7) the *gaussian KL* and the second term of the *RHS* of (7) the *label KL*:

$$gaussian_kl = \left\langle r_\phi(y), \mathbb{E}_{q_x^*} [t_x(x)] \right\rangle - \log (Z_x(\eta_x^*(\phi)))$$

$$label_kl = \left\langle \mathbb{E}_{q_z^*} [\eta_x^0(\theta, z)], \mathbb{E}_{q_x^*} [t_x(x)] \right\rangle - \left(\log (Z_z(\eta_z^*(\phi))) - \log (Z_z(\eta_z^0(\theta))) \right)$$

So, replacing the terms of the *RHS* of (2) respectively (3) and (7), we can compute efficiently the SVAE objective.

4 PSEUDO CODE

Algorithm 1 Implementation of SVAE

```

1: function SVAE
2:   Initialization of the parameters
3:   Mean parameters  $\theta = (\pi^0, \{\mu_k^0, \Sigma_k^0\}_{k \in [0, K]})$ 
4:   Recognition network:  $\gamma$ 
5:   Generative network:  $\phi$ 
6:
7:   Process mean parameters
8:    $\psi \leftarrow r_\phi(y)$   $\triangleright$  Get the node potential from the recognition network
9:    $\eta_z^0 \leftarrow \text{natpar}(\pi^0)$   $\triangleright$  Get natural parameters for Discrete distribution
10:   $\eta_x^0 \leftarrow \text{natpar}(\{\mu_k^0, \Sigma_k^0\}_k)$   $\triangleright$  Get natural parameters for Gaussian distribution
11:
12:   $(\eta_x^*, \bar{t}_x, \eta_z^*, \bar{t}_z) \leftarrow \mathbf{FIXEDmeanfield}(\psi, \eta_z^0, \eta_x^0)$   $\triangleright$  Coordinate block ascent algorithm
13:   $\text{KL}^{local} \leftarrow \mathbf{LOCALmeanfield}(\psi, \eta_z^0, \eta_x^0, \bar{t}_z, \bar{t}_x)$   $\triangleright$  Compute local meanfield
14:   $\hat{x} \sim q_x^*$   $\triangleright$  Sample mean parameter of observations
15:  loglike  $\leftarrow \mathbf{Xentropy}(y, f_\gamma(\hat{x}))$   $\triangleright$  Estimate loglikelihood term
16: return loglike  $-\text{KL}^{local}$ 

```
