SVAE objective

July 26, 2017

1 Definition of the model

Define the following model:

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

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:

$$p_{z|\theta}(z_n) = \exp\Big(\langle \eta^0_z(\theta), t_z(z_n)\rangle - \log\big(Z_z(\eta^0_z(\theta))\big)\Big)$$

where:

$$\begin{aligned} \eta^0_z(\theta)^\top &= \big[\ln(\pi^0_1), \ldots, \ln(\pi^0_K)\big] \\ t_z(z_n)^\top &= \big[\delta(z=1), \ldots, \delta(z=K)\big] \\ \log\big(Z_z(\eta)\big) &= 0 \quad \forall\eta \end{aligned}$$

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

$$p_{x|z,\theta}(x) = \exp\Big(\langle \eta^0_x(\theta, z), t_x(x)\rangle - \log\big(Z_x(\eta^0_x(\theta, z))\big)\Big)$$

where: $t_x(x)^\top = \big[x, vec(xx^\top)^\top\big]$ and $\forall k \in [1, K]$, $\eta^0_x(\theta, z=k)^\top = \big[\Sigma^{0\,-1}_k \mu^0_k, vec(-\frac{1}{2}\Sigma^{0\,-1}_k)^\top\big]$.

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

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

and thus, we have:

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

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

$$
\eta_x^0(\theta, z=k)^\top = \big[\Sigma_k^{0-1}\mu_k^0, vec(-\frac{1}{2}\Sigma_k^{0-1})^\top, -\frac{1}{2}\ln\left(|2\pi\Sigma_k^0|\right) - \frac{1}{2}\mu_k^{0\top}\Sigma_k^{0-1}\mu_k^0\big]
$$

2 SVAE objective

2.1 SVAE objective Definition

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

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

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

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

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

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

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 $\widehat{\mathcal{L}}$ *w.r.t* η_x and η_z defining η_x^* and η_z^*:

$$\eta_x^*(\phi), \eta_z^*(\phi) = argmax_{\eta_x, \eta_z} \widehat{\mathcal{L}}(\eta_z, \eta_x, \phi)$$

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

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

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

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

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 respectively q_x^* and using the following approximation:

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

where $\hat{x} \sim q_x^*$

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:

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

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

$$\begin{aligned} KL(q_z^* || p_{z|\theta}) = &\Big\langle \eta_z^*(\phi) - \eta_z^0(\theta), \mathbb{E}_{q_z^*}[t_z(z)] \Big\rangle \\ &- \Big(\log\big(Z_z(\eta_z^*(\phi))\big) - \log\big(Z_z(\eta_z^0(\theta))\big)\Big) \end{aligned} \tag{4}$$

and

$$\begin{aligned} \mathbb{E}_{q_z^*}\Big[KL(q_x^* || p_{x|z,\theta})\Big] = &\Big\langle \eta_x^*(\phi) - \mathbb{E}_{q_z^*}[\eta_x^0(\theta, z)], \mathbb{E}_{q_x^*}[t_x(x)] \Big\rangle \\ &- \Big(\log\big(Z_x(\eta_x^*(\phi))\big) - \mathbb{E}_{q_z^*}[\log\big(Z_x(\eta_x^0(\theta, z))\big)]\Big) \end{aligned} \tag{5}$$

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

$$\begin{aligned} KL(q_z^* q_x^* || p_{z|\theta} p_{x|z,\theta}) = &\Big[\Big\langle \eta_x^*(\phi) - \mathbb{E}_{q_z^*}[\eta_x^0(\theta, z)], \mathbb{E}_{q_x^*}[t_x(x)] \Big\rangle - \log\big(Z_x(\eta_x^*(\phi))\big)\Big] \\ &+ \Big[\Big\langle \eta_z^*(\phi) - \eta_z^0(\theta), \mathbb{E}_{q_z^*}[t_z(z)] \Big\rangle + \mathbb{E}_{q_z^*}[\log\big(Z_x(\eta_x^0(\theta, z))\big)] \\ &- \Big(\log\big(Z_z(\eta_z^*(\phi))\big) - \log\big(Z_z(\eta_z^0(\theta))\big)\Big)\Big] \end{aligned} \tag{6}$$

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 $\propto$:

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

and thus:

$$\eta_x^*(\phi) = \mathbb{E}_{q_z^*}[\eta_x^0(\theta,z)] + r_\phi(y) \tag{7}$$

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 $\propto$:

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

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

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

and thus:

$$\eta_z^*(\phi) = \mathbb{E}_{q_x^*}\big[\eta_x^0(\theta)^\top \begin{bmatrix} t_x(x) \\ 1 \end{bmatrix}\big] + \eta_z^0(\theta) \tag{8}$$

We can inject the expressions (7) and (8) in (6) to get:

$$\begin{aligned} KL(q_z^* q_x^* || p_{z|\theta}p_{x|z,\theta}) = &\Big[\Big\langle r_\phi(y), \mathbb{E}_{q_x^*}[t_x(x)]\Big\rangle - \log\big(Z_x(\eta_x^*(\phi))\big)\Big] \\ &+ \Big[\Big\langle \mathbb{E}_{q_z^*}\big[\eta_x^0(\theta,z)\big], \mathbb{E}_{q_x^*}[t_x(x)]\Big\rangle \\ &- \Big(\log\big(Z_z(\eta_z^*(\phi))\big) - \log\big(Z_z(\eta_z^0(\theta))\big)\Big)\Big] \end{aligned} \tag{9}$$

We denote the first term of the *RHS* of (9) the *gaussian KL* and the second term of the *RHS* of (9) the *label KL*:

$$gaussian\_kl = \Big\langle r_\phi(y), \mathbb{E}_{q_x^*}[t_x(x)]\Big\rangle - \log\big(Z_x(\eta_x^*(\phi))\big) \tag{10}$$

$$label\_kl = \Big\langle \mathbb{E}_{q_z^*}\big[\eta_x^0(\theta,z)\big], \mathbb{E}_{q_x^*}[t_x(x)]\Big\rangle - \Big(\log\big(Z_z(\eta_z^*(\phi))\big) - \log\big(Z_z(\eta_z^0(\theta))\big)\Big) \tag{11}$$

We can then compute the local KL and perform the block ascent algorithm using (10) and (11).

4 PSEUDO CODE

Algorithm 1 Implementation of SVAE

function SVAE
 Initialization of the parameters
  Mean parameters $\theta = (\pi^0, \{\mu_k^0, \Sigma_k^0\}_{k \in [0,K]})$
  Recognition network: γ
  Generative network: ϕ

 Process mean parameters
  $\psi \leftarrow r_\phi(y)$ ▷ Get the node potential from the recognition network
  $\eta_z^0 \leftarrow natpar(\pi^0)$ ▷ Get natural parameters for Discrete distribution
  $\eta_x^0 \leftarrow natpar(\{\mu_k^0, \Sigma_k^0\}_k)$ ▷ Get natural parameters for Gaussian distribution

 $(\eta_x^*, \bar{t_x}, \eta_z^*, \bar{t_z}) \leftarrow$ **FIXEDmeanfield**$(\psi, \eta_z^0, \eta_x^0)$ ▷ Coordinate block ascent algorithm
 $\mathrm{KL}^{local} \leftarrow$ **LOCALmeanfield**$(\psi, \eta_z^0, \eta_x^0, \bar{t_z}, \bar{t_x})$ ▷ Compute local meanfield
 $\hat{x} \sim q_x^*$ ▷ Sample mean parameter of observations
 loglike $\leftarrow$ **Xentropy**$(y, f_\gamma(\hat{x}))$ ▷ Estimate loglikelihood term
return loglike$-\mathrm{KL}^{local}$