TEDnet Documentation

1 Model Specification

1.1 Mixture Density Outputs

$$x_t \in \mathbb{R}^{1\times k} \tag{1}$$

$$\pi_t \in \mathbb{R} \tag{2}$$

$$\mu_t \in \mathbb{R}^{1\times k} \tag{3}$$

$$\Sigma_t \in \mathbb{R}^{k\times k} \tag{4}$$

$$y_t = \left(\{\pi_t^j, \mu_t^j, \Sigma_t^j\}_{j=1}^M\right) \tag{5}$$

From Inputs to Mixture Gaussian Components

$$\hat{y}_t = \left(\{\hat{\pi}_t^j, \hat{\mu}_t^j, \hat{\Sigma}_t^j\}_{j=1}^M\right) = b_y + \sum_{n=1}^{N} W_{h^n} h_t^n \tag{6}$$

$$\pi_t^j = \frac{\exp\left(\hat{\pi}_t^j\right)}{\sum_{j'=1}^M \exp\left(\hat{\pi}_t^{j'}\right)} \implies \pi_t^j \in (0,1),\ \sum_{j=1}^M \pi_t^j = 1 \tag{7}$$

$$\mu_t^j = \hat{\mu}_t^j \tag{8}$$

There are two options for the number of input values necessary for describing the covariance matrix values. We have one option where the covariance is treated normally, and since the covariance matrix is symmetric we only need to account for k(k-1)/2 value (either upper or lower triangular value). The second option is to treat the covariance matrix as a strickly diagonal matrix and therefore we only account for k values.

Option 1: Full Covariance Matrix

$$\hat{\Sigma}_t \in \mathbb{R}^{1\times(k(k-1)/2)} \tag{9}$$

$$\Sigma_{t,(l,m)} = \begin{cases} \hat{\Sigma}_{t,(1,\ l\cdot(l-1)/2+m)}, & \text{if } m \le l \\ \hat{\Sigma}_{t,(1,\ m\cdot(m-1)/2+l)}, & \text{if } m > l \end{cases} \tag{10}$$

Option 2: Diagonal Covariance Matrix

$$\hat{\Sigma}_t \in \mathbb{R}^{1\times k} \tag{11}$$

$$\Sigma_{t,(l,m)} = \begin{cases} \hat{\Sigma}_{t,(1,\ l)}, & \text{if } l = m \\ 0, & \text{otherwise} \end{cases} \tag{12}$$

Probability and Loss

The probability density $\Pr(x_{t+1} \mid y_t)$ of the next input x_{t+1} given the output y_t is defined as follows:

$$\Pr(x_{t+1} \mid y_t) = \sum_{j=1}^{M} \pi_t^j \ \mathcal{N}(x_{t+1} \mid \mu_t^j, \Sigma_t^j) \tag{13}$$

where

$$\mathcal{N}(x \mid \mu, \Sigma) = \frac{1}{\sqrt{2\pi\,|\Sigma|}} \ \exp\big[-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\big] \tag{14}$$

Using eq 13 we can determine sequence loss as follows:

$$\mathcal{L}(x) = \sum_{t=1}^{T} -\log\Big(\sum_{j=1}^{M} \pi_t^j \ \mathcal{N}(x_{t+1} \mid \mu_t^j, \Sigma_t^j)\Big) \tag{15}$$

Deriving Gradient of Loss[1]

Using responsibilities γ_t^j will make derivations clearer:

$$\hat{\gamma}_t^j = \pi_t^j \mathcal{N}(x_{t+1} \mid \mu_t^j, \Sigma_t^j) \tag{16}$$

$$\gamma_t^j = \frac{\hat{\gamma}_t^j}{\sum_{j'}^{M} \hat{\gamma}_t^{j'}} \tag{17}$$

$$\frac{\delta \mathcal{L}(x)}{\delta \pi_t^j} = -\gamma_t^j \frac{1}{\pi_t^j} \tag{18}$$

$$\frac{\delta \mathcal{L}(x)}{\delta \mu_t^j} = -\gamma_t^j \big([\Sigma_t^j]^{-1}(x_{t+1} - \mu_t^j)\big) \tag{19}$$

$$\frac{\delta \mathcal{L}(x)}{\delta \Sigma_t^j} = -\gamma_t^j \big(\frac{1}{2}\big[-[\Sigma_t^j]^{-1} + [\Sigma_t^j]^{-1}(x_{t+1} - \mu_t^j)(x_{t+1} - \mu_t^j)^T [\Sigma_t^j]^{-1}\big]\big) \tag{20}$$

References

1. Petersen, K., Pedersen, M.: The Matrix Cookbook. Eq. 390–396, 44–45 (2008)