Exercise 1: Mixture Density Networks (20 + 10 P)

In this exercise, we prove some of the results from the paper Mixture Density Networks by Bishop (1994). The mixture density network is given by

$$p(\boldsymbol{t}|\boldsymbol{x}) = \sum_{i=1}^{m} \alpha_i(\boldsymbol{x})\phi_i(\boldsymbol{t}|\boldsymbol{x})$$

with the mixture elements

$$\phi_i(\boldsymbol{t}|\boldsymbol{x}) = \frac{1}{(2\pi)^{c/2}\sigma_i(\boldsymbol{x})^c} \exp\Big(-\frac{\|\boldsymbol{t}-\boldsymbol{\mu}_i(\boldsymbol{x})\|^2}{2\sigma_i(\boldsymbol{x})^2}\Big).$$

The contribution to the error function of one data point q is given by

$$E^q = -\log\Big\{\sum_{i=1}^{m} \alpha_i(\boldsymbol{x}^q)\phi_i(\boldsymbol{t}^q|\boldsymbol{x}^q)\Big\}$$

$\frac{\partial \|x-y\|^2}{\partial x} = 2(x-y)$

$\frac{\partial \|x-y\|^2}{\partial y} = 2(y-x)$

We also define the posterior distribution

$$\pi_i(\boldsymbol{x}, \boldsymbol{t}) = \frac{\alpha_i\phi_i}{\sum_{j=1}^{m}\alpha_j\phi_j}$$

which is obtained using the Bayes theorem.

(a) *Compute* the gradient of the error E^q w.r.t. the mixture parameters, i.e. show that

(i) $\frac{\partial E^q}{\partial \alpha_i} = -\frac{\pi_i}{\alpha_i}$

$\frac{\partial E^q}{\partial \alpha_i} = -\frac{1}{\sum_j \alpha_j\phi_j} \cdot \phi_i \cdot \frac{\alpha_i}{\alpha_i} \cdot \pi_i = -\frac{\pi_i}{\alpha_i}$

(ii) $\frac{\partial E^q}{\partial \mu_{ik}} = \pi_i\Big(\frac{\mu_{ik}-t_k}{\sigma_i^2}\Big)$

$\frac{\partial E}{\partial \mu_{ik}} = \frac{\alpha_i}{\sum_j \alpha_j\phi_j} \cdot \phi_i \cdot \Big(\frac{\mu_{ik}-t_k}{\sigma_i^2}\Big) = \pi_i\Big(\frac{\mu_{ik}-t_k}{\sigma_i^2}\Big)$

(b) We now assume that the neural network produces the mixture coefficients as:

$$\alpha_i = \frac{\exp(z_i^\alpha)}{\sum_{j=1}^{M}\exp(z_j^\alpha)}$$

$\left(\frac{f}{g}\right)' = \frac{f'g - fg'}{g^2}$

where z^α denotes the outputs of the neural network (after the last linear layer) associated to these mixture coefficients. *Compute* using the chain rule for derivatives (i.e. by reusing some of the results in the first part of this exercise) the derivative $\partial E^q/\partial z_i^\alpha$.

$$\frac{\partial E^q}{\partial z_i^\alpha} = \sum_j \frac{\partial E^q}{\partial \alpha_j}\cdot\frac{\partial \alpha_j}{\partial z_i}$$

$$\left[\frac{\partial \alpha_j}{\partial z_i} = \frac{\delta_{ij}\exp(z_i)\cdot(\sum_j \exp(z_j)) - \exp(z_j)\exp(z_i)}{(\sum_j \exp(z_j))^2} = \delta_{ij}\alpha_i - \alpha_i\alpha_j\right]$$

$$= \sum_j -\frac{\pi_j}{\alpha_j}(\delta_{ij}\alpha_i - \alpha_i\alpha_j) = \sum_j\Big(-\frac{\pi_j}{\alpha_j}\delta_{ij}\alpha_i + \frac{\pi_j}{\alpha_j}\alpha_i\alpha_j\Big) = -\pi_i + \Big(\sum_j \pi_j\Big)\alpha_i = -\pi_i + \alpha_i$$

Exercise 2: Conditional RBM (20 + 10 P)

The conditional restricted Boltzmann machine is a system of binary variables comprising inputs $\boldsymbol{x} \in \{0,1\}^d$, outputs $\boldsymbol{y} \in \{0,1\}^c$, and hidden units $\boldsymbol{h} \in \{0,1\}^K$. It associates to each configuration of these binary variables the energy:

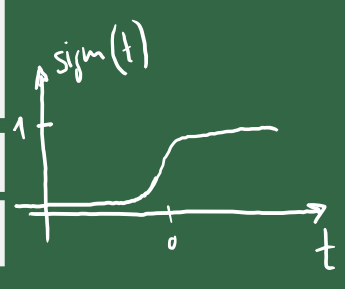
$$E(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{h}) = -\boldsymbol{x}^\top W\boldsymbol{h} - \boldsymbol{y}^\top U\boldsymbol{h} = \sum_k -\boldsymbol{x}^\top W_{:,k}h_k - \boldsymbol{y}^\top U_{:,k}h_k$$

$E(x, y, h_k)$

and the probability associated to each configuration is then given as:

$$p(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{h}) = \frac{1}{Z}\exp(-E(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{h}))$$

where Z is a normalization constant that makes probabilities sum to one.



(a) Let $\text{sigm}(t) = \exp(t)/(1+\exp(t))$ be the sigmoid function. *Show* that

(i) $p(h_k = 1 \,|\, \boldsymbol{x}, \boldsymbol{y}) = \text{sigm}\big(\boldsymbol{x}^\top W_{:,k} + \boldsymbol{y}^\top U_{:,k}\big)$

$$p(h_k=1 \mid x,y) = \frac{p(h_k=1, x, y)}{p(x,y)} \quad \text{| product rule}$$

$$= \frac{\sum_{h_{-k}} p(h_k=1, x, y, h_{-k})}{\sum_{q\in\{0,1\}}\sum_{h_{-k}} p(h_k=q, x, y, h_{-k})} \quad \text{| marginalization}$$

$\sum_{h_{-k}} : \sum_{q_1\in\{0,1\}}\sum_{q_2\in\{0,1\}}\sum_{q_3}\cdots$ (all hidden units except h_k)

$$= \frac{\sum_{h_{-k}} \exp\big(x^\top W_{:,k}1 + y^\top U_{:,k}1 - E(x,y,h_{-k})\big)/Z}{\sum_{q\in\{0,1\}}\sum_{h_{-k}} \exp\big(x^\top W_{:,k}q + y^\top U_{:,k}q - E(x,y,h_{-k})\big)/Z}$$

$$= \frac{\exp(x^\top W_{:,k}1 + y^\top U_{:,k}1)\cdot\sum_{h_{-k}}\exp(-E(x,y,h_{-k}))}{\sum_{q\in\{0,1\}}\exp(x^\top W_{:,k}q + y^\top U_{:,k}q)\cdot\sum_{h_{-k}}\exp(-E(x,y,h_{-k}))}$$

$$= \frac{\exp(x^\top W_{:,k}1 + y^\top U_{:,k}1)}{1+\exp(x^\top W_{:,k}1 + y^\top U_{:,k}1)} = \text{sigm}\big(x^\top W_{:,k} + y^\top U_{:,k}\big)$$

(ii) $p(y_j = 1 \,|\, \boldsymbol{h}, \boldsymbol{x}) = \text{sigm}\big(U_{j,:}^\top \boldsymbol{h}\big)$

(b) *Show* that

$$p(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{Z}\exp(-F(\boldsymbol{x}, \boldsymbol{y}))$$

where

$$F(\boldsymbol{x}, \boldsymbol{y}) = -\sum_{k=1}^{K}\log\big(1+\exp\big(\boldsymbol{x}^\top W_{:,k} + \boldsymbol{y}^\top U_{:,k}\big)\big)$$

is the free energy and where Z is again a normalization constant.

$$Z\cdot p(x,y) = \sum_h p(x,y,h)\cdot Z$$

$$= \sum_h \exp(-E(x,y,h)) = \sum_h \exp\Big(-\sum_k E(x,y,h_k)\Big)$$

$$= \sum_h \pi_k \exp(-E(x,y,h_k)) \qquad \sum_i\sum_j \pi_j a_{ij} = \pi_j\sum_i a_{ij}$$

$$= \pi_k \sum_{q_k\in\{0,1\}} \exp(-E(x,y,h_k=q_k))$$

$$= \pi_k(1 + \exp(-E(x,y,h_k=1))$$

$$= \exp\Big(\log\big(\pi_k 1 + \exp(-E(x,y,h_k=1))\big)\Big)$$

$$= \exp\Big(\sum_k \log\big(1+\exp(-E(x,y,h_k=1))\big)\Big)$$

$$= \exp(-F(x,y))$$