Machine Learning Group
Faculty IV – Electrical Engineering and Computer Science
Technische Universität Berlin

Exercise Sheet 11

Exercise 1: Mixture Density Networks (30 + 30 + 40 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

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

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. We would like to compute the gradient of the error E^q w.r.t. the mixture parameters

(a) *Show* that $\dfrac{\partial E^q}{\partial \alpha_i} = -\dfrac{\pi_i}{\alpha_i}$

(b) Show that $\dfrac{\partial E^q}{\partial \mu_{ik}} = \pi_i \Big(\dfrac{\mu_{ik} - t_k}{\sigma_i^2}\Big)$

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

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

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 $\partial E^q / \partial z_i^\alpha$.

Exercise 1: Mixture Density Networks (30 + 30 + 40 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

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

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. We would like to compute the gradient of the error E^q w.r.t. the mixture parameters

(a) *Show that* $\frac{\partial E^q}{\partial \alpha_i} = -\frac{\pi_i}{\alpha_i}$

$$\frac{\partial E^q}{\partial \alpha_i} = -\frac{1}{\sum_{i=1}^{m} \alpha_i(x^q)\,\phi_i(t^q|x^q)} \cdot \phi_i(t^q|x^q)$$

$$= -\frac{1}{\sum_{i=1}^{m}\alpha_i\phi_i} \cdot \phi_i$$

$$= -\frac{1}{\alpha_i} \cdot \frac{\alpha_i\phi_i}{\sum_{i=1}^{m}\alpha_i\phi_i}$$

$$= -\frac{\pi_i}{\alpha_i}$$

(b) Show that $\frac{\partial E^q}{\partial \mu_{ik}} = \pi_i\left(\frac{\mu_{ik} - t_k}{\sigma_i^2}\right)$

$$\frac{\partial E^q}{\partial \mu_{ik}} = \frac{\partial E^q}{\partial \phi_i} \cdot \frac{\partial \phi_i}{\partial \mu_{ik}}$$

$$= -\frac{\alpha_i}{\sum_{i=1}^{m}\alpha_i\phi_i} \cdot \frac{\partial \phi_i}{\partial \mu_{ik}}$$

$$= -\frac{\alpha_i}{\sum_{i=1}^{m}\alpha_i\phi_i} \cdot \frac{1}{(2\pi)^{\frac{c}{2}}\sigma_i(x)^c} \cdot \exp\left(-\frac{\|t_k - \mu_{ik}\|^2}{2\sigma_i^2}\right) \cdot \left(+\frac{2}{2\sigma_i^2}(t_k - \mu_{ik})\right)$$

$$= -\frac{\alpha_i}{\sum_{i=1}^{m}\alpha_i\phi_i} \cdot \phi_i \cdot \left(\frac{1}{\sigma_i^2}(t_k - \mu_{ik})\right)$$

$$= \pi_i \cdot \left(\frac{\mu_{ik} - t_k}{\sigma_i^2}\right)$$

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

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

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 $\partial E^q / \partial z_i^\alpha$.

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

in particular, if $i = k$, $\frac{\partial \alpha_k}{\partial z_i^\alpha} = \frac{\partial \alpha_i}{\partial z_i^\alpha} = \frac{1}{(\sum_{j=1}^{m} \exp(z_j^\alpha))^2} \cdot \left(\exp(z_i^\alpha) \cdot \sum_{j=1}^{m} \exp(z_j^\alpha) - \exp(z_i^\alpha)^2 \right)$

$$= \alpha_i - \alpha_i^2 = \alpha_k - \alpha_k^2$$

if $i \neq k$, $\frac{\partial \alpha_k}{\partial z_i^\alpha} = \frac{1}{(\sum_{j=1}^{m} \exp(z_j^\alpha))^2} \cdot \left(- \exp(z_k^\alpha) \cdot \exp(z_i^\alpha) \right)$

$$= - \alpha_k \cdot \alpha_i$$

We simplify this by $\delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$

$$\frac{\partial \alpha_k}{\partial z_i^\alpha} = \alpha_k \cdot (\delta_{ik} - \alpha_i)$$

$$\frac{\partial E^q}{\partial z_i^\alpha} = \sum_{k=1}^{m} \frac{\partial E^q}{\partial \alpha_k} \cdot \frac{\partial \alpha_k}{\partial z_i^\alpha}$$

$$= \sum_{k=1}^{m} \left(- \frac{\pi_k}{\cancel{\alpha_k}} \cdot \cancel{\alpha_k} \cdot (\delta_{ik} - \alpha_i) \right)$$

$$= \sum_{k=1}^{m} \left(\pi_k \alpha_i - \pi_k \delta_{ik} \right)$$

$$= \sum_{k=1}^{m} \left(\frac{\alpha_k \phi_k}{\sum_{j=1}^{m} \alpha_j \phi_j} \cdot \alpha_i \right) - \pi_i$$

$$= \alpha_i \cdot \sum_{k=1}^{m} \frac{\alpha_k \phi_k}{\sum_{j=1}^{m} \alpha_j \phi_j} - \pi_i$$

$$= \alpha_i - \pi_i$$