

Exercise Sheet 8

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

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

with the mixture elements

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

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

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

We also define the posterior distribution

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

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

$$\begin{aligned} \text{(i)} \quad & \frac{\partial E^q}{\partial \alpha_i} = -\frac{\pi_i}{\alpha_i} \\ \text{(ii)} \quad & \frac{\partial E^q}{\partial \mu_{ik}} = \pi_i \left(\frac{\mu_{ik} - t_k}{\sigma_i^2} \right) \end{aligned}$$

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

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

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 2: Conditional RBM (20 + 10 P)

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

$$E(\mathbf{x}, \mathbf{y}, \mathbf{h}) = -\mathbf{x}^\top \mathbf{W} \mathbf{h} - \mathbf{y}^\top \mathbf{U} \mathbf{h}$$

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

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

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) $p(h_k = 1 \mid \mathbf{x}, \mathbf{y}) = \text{sigm}(\mathbf{x}^\top W_{:,k} + \mathbf{y}^\top U_{:,k})$

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

(b) *Show* that

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

where

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

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

Exercise 3: Programming (40 P)

Download the programming files on ISIS and follow the instructions.