Exercises for the course

Machine Learning 2

Summer semester 2024

Fachgebiet Maschinelles Lernen
Institut für Softwaretechnik und theoretische Informatik
Fakultät IV, Technische Universität Berlin
Prof. Dr. Klaus-Robert Müller
Email: klaus-robert.mueller@tu-berlin.de

Exercise Sheet 13

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

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

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) $\dfrac{\partial E^q}{\partial \alpha_i} = -\dfrac{\pi_i}{\alpha_i}$

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

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

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

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 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:

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

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

(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.

Exercise 3: Programming (40 P)

Download the programming files on ISIS and follow the instructions.