Most of the following exercises are extracted from the books “Pattern Recognition and Machine Learning” by Bishop and “Bayesian Reasoning and Machine Learning” by Barber.

Problem 1

Suppose we have some data \((x_i, y_i), i = 1, \ldots, m\) where \(x_i, y_i \in \mathbb{R}\) and we want to find a regression function \(y = f(x)\). We use a fully Bayesian model:

\[
y = \sum_{a=1}^{p} w_a x^a + \xi,
\]

where the inputs \(x \sim P_0\) are iid and generated according to a prior \(P_0\) and \(\xi \sim \mathcal{N}(0, \sigma^2)\) iid. The \(w_a \in \mathbb{R}\) are regression parameters. We take for \(w_a\) the prior \(\sim e^{-\alpha w_a^2}\) where \(\alpha\). The parameters \(\alpha\) and \(\sigma^2\) are supposed to be known. So our model for the data generating process is

\[
\mathcal{D}(y \mid x, w) \mathcal{D}(x) = (2\pi \sigma^2)^{-1/2} e^{-\frac{1}{2\sigma^2}(y-\sum_{a=1}^{p} w_a x^a)^2} P_0(x)
\]

1) Write down the joint distribution for \((y_1, \ldots, y_m, x_1, \ldots, x_m, w_1, \ldots, w_p)\).

2) Draw a Belief Network (directed acyclic graph) corresponding to this probabilistic model.

3) Remark 1: Show that the maximum likelihood principle (take \(\alpha = 0\) or equivalently no prior on \(w_a\)’s) is equivalent to empirical risk minimisation in the hypothesis class of functions \(H \ni f(x) = \sum_{a=1}^{p} w_a x^a\).

4) Remark 2: Consider the MAP principle for estimating \(w_a\)’s and show that it is equivalent to an empirical risk minimization with additional penalty term proportional to \(\alpha \sum_{a=1}^{p} w_a^2\) (this is called ridge regression).

5) Remark 3: The ML or MAP estimates of \(w_a\)’s are to viewed in general as summarized versions of a more detailed object, namely the complete posterior distribution \(P(w_1, \ldots, w_p \mid (x_i, y_i)_{i=1}^{m})\). Show that the optimal regression function in a fully Bayesian approach is

\[
f(x) = \sum_{a=1}^{p} \mathbb{E}_{w \mid data}[w_a] x^a
\]

where \(\mathbb{E}_{w \mid data}\) is the expectation with respect to the posterior distribution \(P(w_1, \ldots, w_p \mid (x_i, y_i)_{i=1}^{m})\).
Problem 2

For each case below, is $a \perp \perp b$ true? And is $a \perp \perp b|c$ true? If yes, prove your answer.

(1)  (2)  (3)

Problem 3

Consider a generic belief network (also called Bayesian network).
Let $MB(i) = \{\text{pa}(i), \text{child}(i), \text{co-parent}(i)\}$ be the Markov blanket of $x_i$. Show that

$p(x_i | \{x_j \}_{j \neq i}) = p(x_i | \{x_v \}_{v \in MB(i)})$.

Problem 4 (Bishop, p.371 & 419, Exercise 8.7)

The linear-Gaussian models for the above graph consists of three random variables $x_1, x_2, x_3$. The model has the structure equations

$$x_i = \sum_{j \in \text{pa}(i)} w_{ij} x_j + b_i + \sqrt{v_i} \epsilon_i, \quad i = 1, 2, 3$$

where $\text{pa}(i)$ is the set of parent nodes of node $i$ (\text{pa}(1) = \emptyset, \text{pa}(2) = \{1\}, \text{pa}(3) = \{2\})$.

Show that the mean and covariance of the joint distribution for the above graph are given by (hint: use a recursive calculation)

$$\mu = (b_1, b_2 + w_{21} b_1, b_3 + w_{32} b_2 + w_{32} w_{21} b_1)^\top$$

$$\Sigma = \begin{pmatrix}
 v_1 & w_{21} v_1 & w_{32} w_{21} v_1 \\
 w_{21} v_1 & v_2 + w_{21}^2 v_1 & w_{32} (v_2 + w_{21}^2 v_1) \\
 w_{32} w_{21} v_1 & w_{32} (v_2 + w_{21}^2 v_1) & v_3 + w_{32} (v_2 + w_{21}^2 v_1)
\end{pmatrix}$$
Problem 5 (Barber, p.75, Exercise 4.4)

The restricted Boltzmann machine (RBM) is a constrained Boltzmann machine on a bipartite graph, consisting of a layer of visible variables $v = (v_1, \cdots, v_V)^\top$ and hidden variables $h = (h_1, \cdots, h_H)^\top$:

$$p(v, h) = \frac{1}{Z(W, a, b)} \exp \left( v^\top W h + a^\top v + b^\top h \right)$$

All variables are binary taking value 0 or 1. Here $W$ is an $V \times H$ matrix of weight $W_{ji}$.

1) Show that the distribution of hidden units conditional on the visible unit is factorized as

$$p(h | v) = \prod_i p(h_i | v), \quad \text{with } p(h_i = 1 | v) = \sigma \left( b_i + \sum_j W_{ji} v_j \right)$$

where $\sigma(x) = e^x / (1 + e^x)$.

2) By symmetry arguments, write down the form of the conditional $p(v | h)$.

3) Is $p(h) = \prod_i p(h_i)$?

4) Can the partition function $Z(W, a, b)$ be computed efficiently for the RBM?

Problem 6 (Barber, p.77, Exercise 4.14)

Consider a pairwise binary Markov network defined on variables $x_i \in \{0, 1\}, i = 1, \ldots, N$, with $p(x) = \frac{1}{Z} \prod_{ij \in E} \phi_{ij}(x_i, x_j)$ where $E$ is a given edge set and the factors $\phi_{ij}$ are arbitrary (here edges are non necessarily maximal cliques). Explain how to translate such a Markov network into a Boltzmann machine.

Problem 7

Let $G = (V, E)$ an undirected graph whose vertices $V = \{1, \ldots, n\}$ are associated to random variables, and edges are given by the set of pairs $E$. For simplicity the random variables are assumed to be discrete. Denote $C$ the set of maximal cliques of $G$ and consider a probability distribution $p(x)$ which factorizes as

$$p(x) = \frac{1}{Z} \prod_{C \in C} \psi_C(x_C),$$

where $Z = \sum_x \prod_{C \in C} \psi_C(x_C)$ and $\forall C \in C, \forall x_C : \psi_C(x_C) > 0$. Remember that – for a Markov Random Field – the Markov blanket $\partial S$ of a subset $S \subseteq V$ is the set of all vertices that are directly connected to a vertex in $S$ and are not in $S$. Show that the following conditional independence properties is satisfied:

$$\forall S \subseteq V : p(x_S | x_{V \setminus S}) = p(x_S | x_{\partial S}).$$
Problem 8 (Barber, p.99, Exercise 5.4)

Consider the hidden Markov model (HMM)

\[ p(v, h) = p(h_1)p(v_1|h_1) \prod_{t=2}^{T} p(v_t|h_t)p(h_t|h_{t-1}) \]

in which \( \text{dom}(h_t) = \{1, \ldots, H\} \) and \( \text{dom}(v_t) = \{1, \ldots, V\} \) for all \( t = 1, \ldots, T \).

1) Draw a belief network representation of the above distribution.

2) Show that the belief network for \( p(h_1, \ldots, h_T) \) is a simple linear chain. Draw the belief network corresponding to \( p(v_1, \ldots, v_T) \) (this is called a fully connected cascade belief network).

3) Draw a factor graph representation of the above distribution.

4) Use the factor graph to derive a Sum-Product algorithm to compute marginals \( p(h_t|v_1, \ldots, v_T) \).

   Explain the sequence order of messages passed on your factor graph.

5) Explain how to compute \( p(h_t, h_{t+1}|v_1, \ldots, v_T) \).

Problem 9 (Barber, p.98, Exercise 5.1)

Given a pairwise tree Markov network of the form

\[ p(x) = \frac{1}{Z} \prod_{i \sim j} \phi(x_i, x_j), \]

explain how to efficiently compute the normalization factor (also called the partition function) \( Z \) as a function of the potentials \( \phi \).

Problem 10 (Bishop, p.397 & 421, Exercise 8.16 & 8.17)

The joint distribution for the above graph takes the form

\[ p(x) = \frac{1}{Z} \phi_{1,2}(x_1, x_2)\phi_{2,3}(x_2, x_3) \cdots \phi_{N-1,N}(x_{N-1}, x_N). \]

The marginal probability \( p(x_n) = \sum_{i \in \{1, \ldots, N\} \setminus n} p(x) \) can be written as

\[ p(x_n) = \frac{1}{Z_n} \mu_\alpha(x_n) \mu_\beta(x_n) \quad \text{with} \quad Z_n = \sum_{x_n} \mu_\alpha(x_n) \mu_\beta(x_n) \]
where $\mu_\alpha(x_n)$ is the message passing forward from node $n - 1$ to node $n$, and $\mu_\beta(x_n)$ is the message passing backward from node $n + 1$ to node $n$. The computation of $\mu_\alpha(x_n)$ and $\mu_\beta(x_n)$ can be done recursively by the following message passing equations:

\[
\begin{align*}
\mu_\alpha(x_2) &= \sum_{x_1} \phi_{1,2}(x_1, x_2) \\
\mu_\beta(x_{N-1}) &= \sum_{x_N} \phi_{N-1,N}(x_{N-1}, x_N) \\
\mu_\alpha(x_n) &= \sum_{x_{n-1}} \phi_{n-1,n}(x_{n-1}, x_n) \mu_\alpha(x_{n-1}) \\
\mu_\beta(x_n) &= \sum_{x_{n+1}} \phi_{n,n+1}(x_n, x_{n+1}) \mu_\beta(x_{n+1})
\end{align*}
\]

1) Discuss how to modify the above message passing algorithm in order to compute $p(x_n | x_N)$ efficiently.

2) Suppose $N = 5$, and nodes $x_3, x_5$ are observed. Show that if the message passing algorithm is applied to the evaluation of $p(x_2 | x_3, x_5)$, the result will be independent of the value of $x_5$. 

5