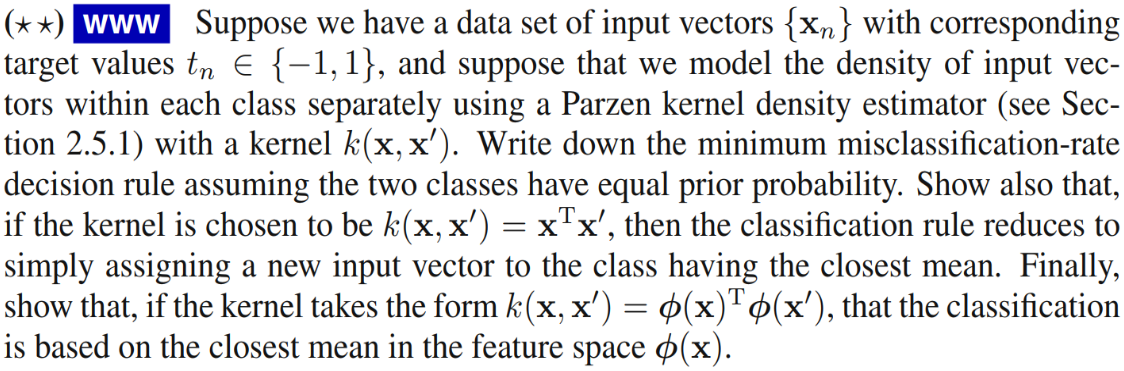
ECE1513
Tutorial 6:
Selected Exercises from Chapters 7 and 8

TA: Faye Pourghasem, fateme.pourghasem@mail.utoronto.ca

October 24, 2023

1 Example 7.1



($\star\star$) www Suppose we have a data set of input vectors $\{\mathbf{x}_n\}$ with corresponding target values $t_n \in \{-1, 1\}$, and suppose that we model the density of input vectors within each class separately using a Parzen kernel density estimator (see Section 2.5.1) with a kernel $k(\mathbf{x}, \mathbf{x}')$. Write down the minimum misclassification-rate decision rule assuming the two classes have equal prior probability. Show also that, if the kernel is chosen to be $k(\mathbf{x}, \mathbf{x}') = \mathbf{x}^{\mathrm{T}}\mathbf{x}'$, then the classification rule reduces to simply assigning a new input vector to the class having the closest mean. Finally, show that, if the kernel takes the form $k(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x})^{\mathrm{T}}\phi(\mathbf{x}')$, that the classification is based on the closest mean in the feature space $\phi(\mathbf{x})$.

Solution

By analogy to Eq (2.249) which represents the probability of the estimated density at x

$$p(\mathbf{x}) = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{h^D}k\left(\frac{\mathbf{x}-\mathbf{x}_n}{h}\right) \tag{2.249}$$

we can have:

$$p(\mathbf{x}|t) = \begin{cases} \dfrac{1}{N_{+1}}\displaystyle\sum_{n=1}^{N_{+1}}\dfrac{1}{Z_k}\cdot k(\mathbf{x}, \mathbf{x}_n) & t = +1 \\ \dfrac{1}{N_{-1}}\displaystyle\sum_{n=1}^{N_{-1}}\dfrac{1}{Z_k}\cdot k(\mathbf{x}, \mathbf{x}_n) & t = -1 \end{cases}$$

where N_{+1} represents the number of samples with label $t = +1$ and it is the same for N_{-1}. Z_k is a normalization constant representing the volume of the hypercube. Since we have equal prior for the class, i.e.,

$$p(t) = \begin{cases} 0.5 & t = +1 \\ 0.5 & t = -1 \end{cases}$$

Based on Bayes' Theorem, we have $p(t|\mathbf{x}) \propto \ p(\mathbf{x}|t) \cdot p(t)$, yielding:

$$p(t|\mathbf{x}) = \begin{cases} \frac{1}{Z} \cdot \frac{1}{N_{+1}} \sum_{n=1}^{N_{+1}} \cdot k(\mathbf{x}, \mathbf{x}_n) & t = +1 \\ \frac{1}{Z} \cdot \frac{1}{N_{-1}} \sum_{n=1}^{N_{-1}} \cdot k(\mathbf{x}, \mathbf{x}_n) & t = -1 \end{cases}$$

Where $1/Z$ is a normalization constant to guarantee the integration of the posterior equal to 1. To classify a new sample $\mathbf{x}^\star$, we try to find the value $t^\star$ that can maximize $p(t|\mathbf{x})$. Therefore, we can obtain:

$$t^\star = \begin{cases} +1 & \text{if} \frac{1}{N_{+1}} \sum_{n=1}^{N_{+1}} \cdot k(\mathbf{x}, \mathbf{x}_n) \geq \frac{1}{N_{-1}} \sum_{n=1}^{N_{-1}} \cdot k(\mathbf{x}, \mathbf{x}_n) \\ -1 & \text{if} \frac{1}{N_{+1}} \sum_{n=1}^{N_{+1}} \cdot k(\mathbf{x}, \mathbf{x}_n) \leq \frac{1}{N_{-1}} \sum_{n=1}^{N_{-1}} \cdot k(\mathbf{x}, \mathbf{x}_n) \end{cases} \qquad (*)$$

If we now choose the kernel function as $k(\mathbf{x}, \mathbf{x}') = \mathbf{x}^T\mathbf{x}'$,we have:

$$\frac{1}{N_{+1}} \sum_{n=1}^{N_{+1}} k(\mathbf{x}, \mathbf{x}_n) = \frac{1}{N_{+1}} \sum_{n=1}^{N_{+1}} \mathbf{x}^T \mathbf{x}_n = \mathbf{x}^T \tilde{\mathbf{x}}_{+1}$$

Where we have denoted:

$$\tilde{\mathbf{x}}_{+1} = \frac{1}{N_{+1}} \sum_{n=1}^{N_{+1}} \mathbf{x}_n$$

and similarly for $\tilde{\mathbf{x}}_{-1}$. Therefore, the classification criterion $(*)$ can be written as:

$$t^\star = \begin{cases} +1 & \text{if} \quad \tilde{\mathbf{x}}_{+1} \geq \tilde{\mathbf{x}}_{-1} \\ -1 & \text{if} \quad \tilde{\mathbf{x}}_{+1} \leq \tilde{\mathbf{x}}_{-1} \end{cases}$$

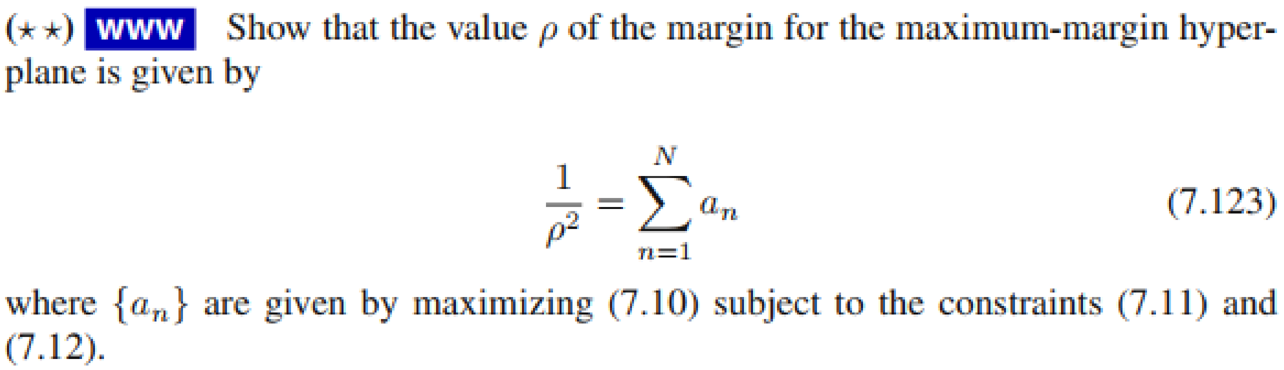
When we choose the kernel function as $k(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x})^T \phi(\mathbf{x}')$, we can similarly obtain the classification criterion:

$$t^\star = \begin{cases} +1 & \text{if} \quad \tilde{\phi}(\mathbf{x}_{+1}) \geq \tilde{\phi}(\mathbf{x}_{-1}) \\ -1 & \text{if} \quad \tilde{\phi}(\mathbf{x}_{+1}) \leq \tilde{\phi}(\mathbf{x}_{-1}) \end{cases}$$

Where we have defined:

$$\tilde{\phi}(\mathbf{x}_{+1}) = \frac{1}{N_{+1}} \sum_{n=1}^{N_{+1}} \phi(\mathbf{x}_n)$$

2 Example 7.4



(⋆⋆) WWW Show that the value ρ of the margin for the maximum-margin hyperplane is given by

$$\frac{1}{\rho^2} = \sum_{n=1}^{N} a_n \tag{7.123}$$

where $\{a_n\}$ are given by maximizing (7.10) subject to the constraints (7.11) and (7.12).

Solution

Problem 7.4 Solution

Since we know that

$$\rho = \frac{1}{||\mathbf{w}||}$$

Therefore, we have:

$$\frac{1}{\rho^2} = ||\mathbf{w}||^2$$

In other words, we only need to prove that

$$||\mathbf{w}||^2 = \sum_{n=1}^{N} a_n$$

When we find th optimal solution, the second term on the right hand side of Eq (7.7) vanishes. Based on Eq (7.8) and Eq (7.10), we also observe that its dual is given by:

$$\tilde{L}(\mathbf{a}) = \sum_{n=1}^{N} a_n - \frac{1}{2}||\mathbf{w}||^2$$

Therefore, we have:

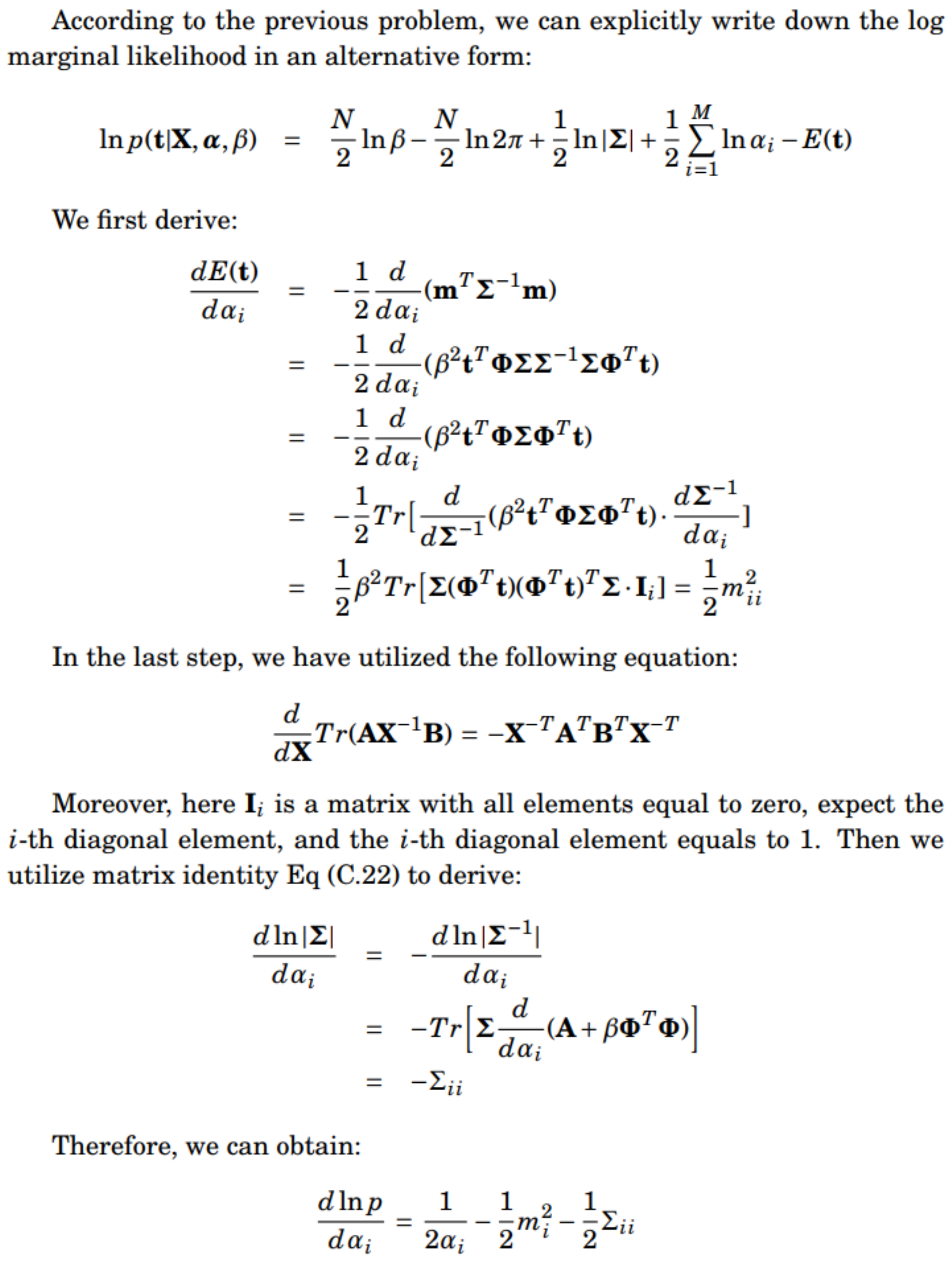
$$\frac{1}{2}||\mathbf{w}||^2 = L(\mathbf{a}) = \tilde{L}(\mathbf{a}) = \sum_{n=1}^{N} a_n - \frac{1}{2}||\mathbf{w}||^2$$

Rearranging it, we will obtain what we are required.

3 Example 7.12

($\star\star$) **www** Show that direct maximization of the log marginal likelihood (7.85) for the regression relevance vector machine leads to the re-estimation equations (7.87) and (7.88) where γ_i is defined by (7.89).

Solution



According to the previous problem, we can explicitly write down the log marginal likelihood in an alternative form:

$$\ln p(\mathbf{t}|\mathbf{X},\boldsymbol{\alpha},\beta) = \frac{N}{2}\ln\beta - \frac{N}{2}\ln 2\pi + \frac{1}{2}\ln|\boldsymbol{\Sigma}| + \frac{1}{2}\sum_{i=1}^{M}\ln\alpha_i - E(\mathbf{t})$$

We first derive:

$$\begin{aligned}
\frac{dE(\mathbf{t})}{d\alpha_i} &= -\frac{1}{2}\frac{d}{d\alpha_i}(\mathbf{m}^T\boldsymbol{\Sigma}^{-1}\mathbf{m}) \\
&= -\frac{1}{2}\frac{d}{d\alpha_i}(\beta^2\mathbf{t}^T\boldsymbol{\Phi}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}\boldsymbol{\Phi}^T\mathbf{t}) \\
&= -\frac{1}{2}\frac{d}{d\alpha_i}(\beta^2\mathbf{t}^T\boldsymbol{\Phi}\boldsymbol{\Sigma}\boldsymbol{\Phi}^T\mathbf{t}) \\
&= -\frac{1}{2}Tr[\frac{d}{d\boldsymbol{\Sigma}^{-1}}(\beta^2\mathbf{t}^T\boldsymbol{\Phi}\boldsymbol{\Sigma}\boldsymbol{\Phi}^T\mathbf{t})\cdot\frac{d\boldsymbol{\Sigma}^{-1}}{d\alpha_i}] \\
&= \frac{1}{2}\beta^2 Tr[\boldsymbol{\Sigma}(\boldsymbol{\Phi}^T\mathbf{t})(\boldsymbol{\Phi}^T\mathbf{t})^T\boldsymbol{\Sigma}\cdot\mathbf{I}_i] = \frac{1}{2}m_{ii}^2
\end{aligned}$$

In the last step, we have utilized the following equation:

$$\frac{d}{d\mathbf{X}}Tr(\mathbf{A}\mathbf{X}^{-1}\mathbf{B}) = -\mathbf{X}^{-T}\mathbf{A}^T\mathbf{B}^T\mathbf{X}^{-T}$$

Moreover, here $\mathbf{I}_i$ is a matrix with all elements equal to zero, expect the i-th diagonal element, and the i-th diagonal element equals to 1. Then we utilize matrix identity Eq (C.22) to derive:

$$\begin{aligned}
\frac{d\ln|\boldsymbol{\Sigma}|}{d\alpha_i} &= -\frac{d\ln|\boldsymbol{\Sigma}^{-1}|}{d\alpha_i} \\
&= -Tr\left[\boldsymbol{\Sigma}\frac{d}{d\alpha_i}(\mathbf{A}+\beta\boldsymbol{\Phi}^T\boldsymbol{\Phi})\right] \\
&= -\Sigma_{ii}
\end{aligned}$$

Therefore, we can obtain:

$$\frac{d\ln p}{d\alpha_i} = \frac{1}{2\alpha_i} - \frac{1}{2}m_i^2 - \frac{1}{2}\Sigma_{ii}$$

Set it to zero and obtain:

$$\alpha_i = \frac{1-\alpha_i\Sigma_{ii}}{m_i} = \frac{\gamma_i}{m_i^2}$$

Then we calculate the derivatives of $\ln p$ with respect to β beginning by:

$$\begin{aligned}
\frac{d\ln|\boldsymbol{\Sigma}|}{d\beta} &= -\frac{d\ln|\boldsymbol{\Sigma}^{-1}|}{d\beta} \\
&= -Tr\Big[\boldsymbol{\Sigma}\frac{d}{d\beta}(\mathbf{A}+\beta\boldsymbol{\Phi}^T\boldsymbol{\Phi})\Big] \\
&= -Tr\Big[\boldsymbol{\Sigma}\boldsymbol{\Phi}^T\boldsymbol{\Phi}\Big]
\end{aligned}$$

Then we continue:

$$\begin{aligned}
\frac{dE(\mathbf{t})}{d\beta} &= \frac{1}{2}\mathbf{t}^T\mathbf{t} - \frac{1}{2}\frac{d}{d\beta}(\mathbf{m}^T\boldsymbol{\Sigma}^{-1}\mathbf{m}) \\
&= \frac{1}{2}\mathbf{t}^T\mathbf{t} - \frac{1}{2}\frac{d}{d\beta}(\beta^2\mathbf{t}^T\boldsymbol{\Phi}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}\boldsymbol{\Phi}^T\mathbf{t}) \\
&= \frac{1}{2}\mathbf{t}^T\mathbf{t} - \frac{1}{2}\frac{d}{d\beta}(\beta^2\mathbf{t}^T\boldsymbol{\Phi}\boldsymbol{\Sigma}\boldsymbol{\Phi}^T\mathbf{t}) \\
&= \frac{1}{2}\mathbf{t}^T\mathbf{t} - \beta\mathbf{t}^T\boldsymbol{\Phi}\boldsymbol{\Sigma}\boldsymbol{\Phi}^T\mathbf{t} - \frac{1}{2}\beta^2\frac{d}{d\beta}(\mathbf{t}^T\boldsymbol{\Phi}\boldsymbol{\Sigma}\boldsymbol{\Phi}^T\mathbf{t}) \\
&= \frac{1}{2}\Big\{\mathbf{t}^T\mathbf{t} - 2\beta\mathbf{t}^T\boldsymbol{\Phi}\boldsymbol{\Sigma}\boldsymbol{\Phi}^T\mathbf{t} - \beta^2\frac{d}{d\beta}(\mathbf{t}^T\boldsymbol{\Phi}\boldsymbol{\Sigma}\boldsymbol{\Phi}^T\mathbf{t})\Big\} \\
&= \frac{1}{2}\Big\{\mathbf{t}^T\mathbf{t} - 2\mathbf{t}^T(\boldsymbol{\Phi}\mathbf{m}) - \beta^2\frac{d}{d\beta}(\mathbf{t}^T\boldsymbol{\Phi}\boldsymbol{\Sigma}\boldsymbol{\Phi}^T\mathbf{t})\Big\} \\
&= \frac{1}{2}\Big\{\mathbf{t}^T\mathbf{t} - 2\mathbf{t}^T(\boldsymbol{\Phi}\mathbf{m}) - \beta^2 Tr\big[\frac{d}{d\boldsymbol{\Sigma}^{-1}}(\mathbf{t}^T\boldsymbol{\Phi}\boldsymbol{\Sigma}\boldsymbol{\Phi}^T\mathbf{t})\cdot\frac{d\boldsymbol{\Sigma}^{-1}}{d\beta}\big]\Big\} \\
&= \frac{1}{2}\Big\{\mathbf{t}^T\mathbf{t} - 2\mathbf{t}^T(\boldsymbol{\Phi}\mathbf{m}) + \beta^2 Tr\big[\boldsymbol{\Sigma}(\boldsymbol{\Phi}^T\mathbf{t})(\boldsymbol{\Phi}^T\mathbf{t})^T\boldsymbol{\Sigma}\cdot\boldsymbol{\Phi}^T\boldsymbol{\Phi}\big]\Big\} \\
&= \frac{1}{2}\Big\{\mathbf{t}^T\mathbf{t} - 2\mathbf{t}^T(\boldsymbol{\Phi}\mathbf{m}) + Tr\big[\mathbf{m}\mathbf{m}^T\cdot\boldsymbol{\Phi}^T\boldsymbol{\Phi}\big]\Big\} \\
&= \frac{1}{2}\Big\{\mathbf{t}^T\mathbf{t} - 2\mathbf{t}^T(\boldsymbol{\Phi}\mathbf{m}) + Tr\big[\boldsymbol{\Phi}\mathbf{m}\mathbf{m}^T\cdot\boldsymbol{\Phi}^T\big]\Big\} \\
&= \frac{1}{2}||\mathbf{t}-\boldsymbol{\Phi}\mathbf{m}||^2
\end{aligned}$$

Therefore, we have obtained:

$$\frac{d\ln p}{d\beta} = \frac{1}{2}\Big(\frac{N}{\beta} - ||\mathbf{t}-\boldsymbol{\Phi}\mathbf{m}||^2 - Tr[\boldsymbol{\Sigma}\boldsymbol{\Phi}^T\boldsymbol{\Phi}]\Big)$$

$$\mathbf{\Sigma} = \left(\mathbf{A} + \beta\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\right)^{-1} \tag{7.83}$$

Using Eq (7.83), we can obtain:

$$\begin{aligned}
\mathbf{\Sigma\Phi}^T\mathbf{\Phi} &= \mathbf{\Sigma\Phi}^T\mathbf{\Phi} + \beta^{-1}\mathbf{\Sigma A} - \beta^{-1}\mathbf{\Sigma A} \\
&= \mathbf{\Sigma}(\beta\mathbf{\Phi}^T\mathbf{\Phi} + \mathbf{A})\beta^{-1} - \beta^{-1}\mathbf{\Sigma A} \\
&= \mathbf{I}\beta^{-1} - \beta^{-1}\mathbf{\Sigma A} \\
&= (\mathbf{I} - \mathbf{\Sigma A})\beta^{-1}
\end{aligned}$$

Setting the derivative equal to zero, we can obtain:

$$\beta^{-1} = \frac{||\mathbf{t} - \mathbf{\Phi m}||^2}{N - Tr(\mathbf{I} - \mathbf{\Sigma A})} = \frac{||\mathbf{t} - \mathbf{\Phi m}||^2}{N - \sum_i \gamma_i}$$

Just as required.

4 Example 7.15

($\star\star$) www Using the results (7.94) and (7.95), show that the marginal likelihood (7.85) can be written in the form (7.96), where $\lambda(\alpha_n)$ is defined by (7.97) and the sparsity and quality factors are defined by (7.98) and (7.99), respectively.

Solution

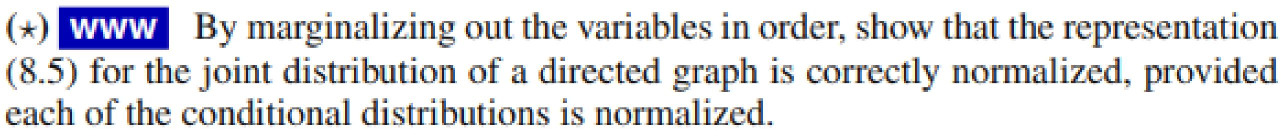
Problem 7.15 Solution

We just follow the hint.

$$
\begin{aligned}
L(\boldsymbol{\alpha}) &= -\frac{1}{2}\{N\ln 2\pi + \ln|\mathbf{C}| + \mathbf{t}^T\mathbf{C}^{-1}\mathbf{t}\} \\
&= -\frac{1}{2}\Big\{N\ln 2\pi + \ln|\mathbf{C}_{-i}| + \ln|1+\alpha_i^{-1}\boldsymbol{\varphi}_i^T\mathbf{C}_{-i}^{-1}\boldsymbol{\varphi}_i| \\
&\quad +\mathbf{t}^T(\mathbf{C}_{-i}^{-1} - \frac{\mathbf{C}_{-i}^{-1}\boldsymbol{\varphi}_i\boldsymbol{\varphi}_i^T\mathbf{C}_{-i}^{-1}}{\alpha_i+\boldsymbol{\varphi}_i^T\mathbf{C}_{-i}^{-1}\boldsymbol{\varphi}_i})\mathbf{t}\Big\} \\
&= L(\boldsymbol{\alpha}_{-i}) - \frac{1}{2}\ln|1+\alpha_i^{-1}\boldsymbol{\varphi}_i^T\mathbf{C}_{-i}^{-1}\boldsymbol{\varphi}_i| + \frac{1}{2}\mathbf{t}^T\frac{\mathbf{C}_{-i}^{-1}\boldsymbol{\varphi}_i\boldsymbol{\varphi}_i^T\mathbf{C}_{-i}^{-1}}{\alpha_i+\boldsymbol{\varphi}_i^T\mathbf{C}_{-i}^{-1}\boldsymbol{\varphi}_i}\mathbf{t} \\
&= L(\boldsymbol{\alpha}_{-i}) - \frac{1}{2}\ln|1+\alpha_i^{-1}s_i| + \frac{1}{2}\frac{q_i^2}{\alpha_i+s_i} \\
&= L(\boldsymbol{\alpha}_{-i}) - \frac{1}{2}\ln\frac{\alpha_i+s_i}{\alpha_i} + \frac{1}{2}\frac{q_i^2}{\alpha_i+s_i} \\
&= L(\boldsymbol{\alpha}_{-i}) + \frac{1}{2}\Big[\ln\alpha_i - \ln(\alpha_i+s_i) + \frac{q_i^2}{\alpha_i+s_i}\Big] = L(\boldsymbol{\alpha}_{-i}) + \lambda(\alpha_i)
\end{aligned}
$$

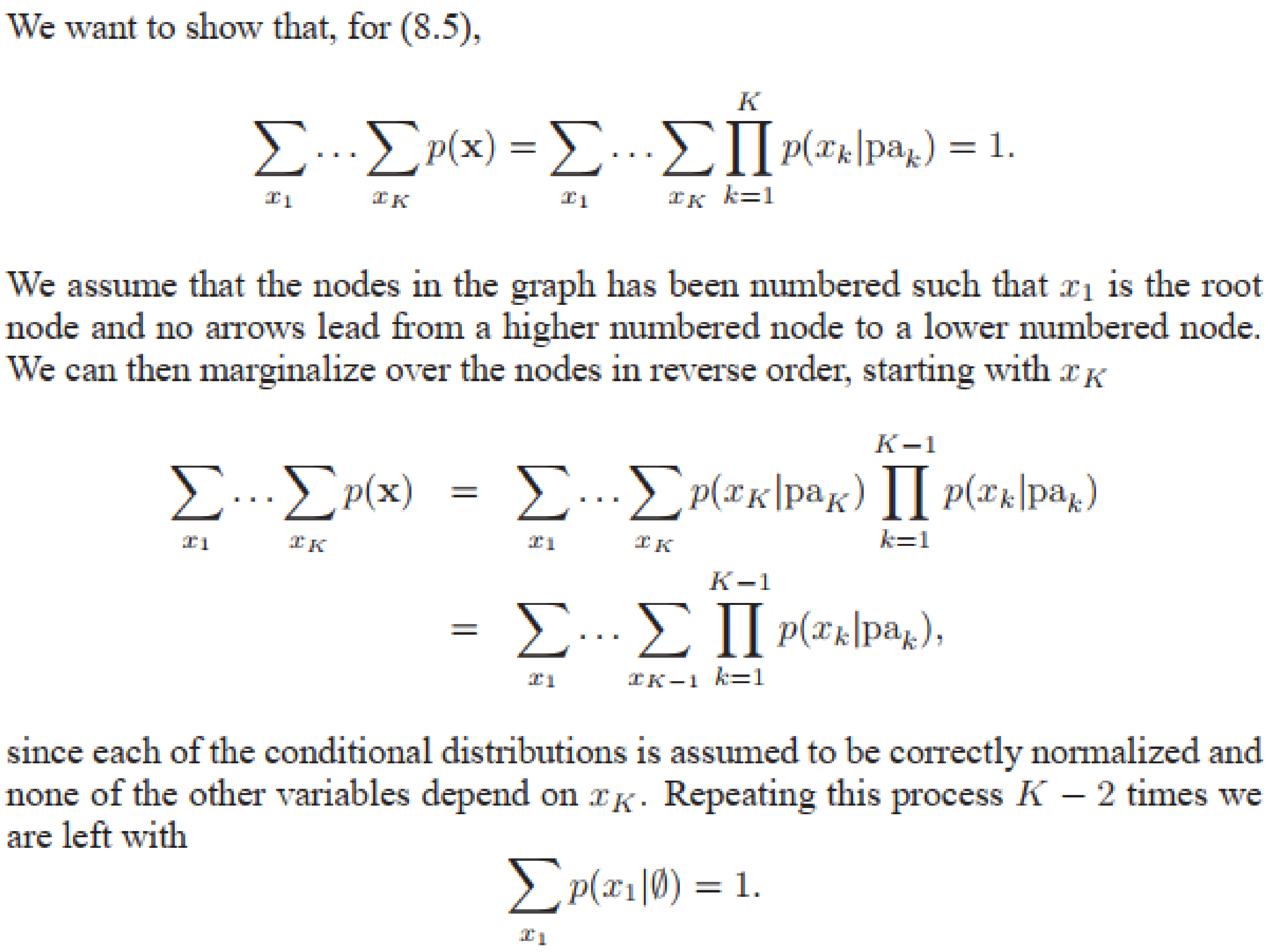
Where we have defined $\lambda(\alpha_i)$, s_i and q_i as shown in Eq (7.97)-(7.99).

5 Example 8.1



(⋆) www By marginalizing out the variables in order, show that the representation (8.5) for the joint distribution of a directed graph is correctly normalized, provided each of the conditional distributions is normalized.

Solution



We want to show that, for (8.5),

$$\sum_{x_1} \cdots \sum_{x_K} p(\mathbf{x}) = \sum_{x_1} \cdots \sum_{x_K} \prod_{k=1}^{K} p(x_k|\mathrm{pa}_k) = 1.$$

We assume that the nodes in the graph has been numbered such that x_1 is the root node and no arrows lead from a higher numbered node to a lower numbered node. We can then marginalize over the nodes in reverse order, starting with x_K

$$\begin{aligned} \sum_{x_1} \cdots \sum_{x_K} p(\mathbf{x}) &= \sum_{x_1} \cdots \sum_{x_K} p(x_K|\mathrm{pa}_K) \prod_{k=1}^{K-1} p(x_k|\mathrm{pa}_k) \\ &= \sum_{x_1} \cdots \sum_{x_{K-1}} \prod_{k=1}^{K-1} p(x_k|\mathrm{pa}_k), \end{aligned}$$

since each of the conditional distributions is assumed to be correctly normalized and none of the other variables depend on x_K. Repeating this process $K-2$ times we are left with

$$\sum_{x_1} p(x_1|\emptyset) = 1.$$

6 Example 8.5

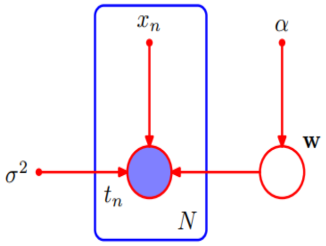
(⋆) **www** Draw a directed probabilistic graphical model corresponding to the relevance vector machine described by (7.79) and (7.80).

Solution

$$p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \prod_{n=1}^{N} p(t_n|\mathbf{x}_n, \mathbf{w}, \beta^{-1}). \tag{7.79}$$

$$p(\mathbf{w}|\boldsymbol{\alpha}) = \prod_{i=1}^{M} \mathcal{N}(w_i|0, \alpha_i^{-1}) \tag{7.80}$$

Figure 8.6 As in Figure 8.5 but with the nodes $\{t_n\}$ shaded to indicate that the corresponding random variables have been set to their observed (training set) values.



It looks quite like Figure 8.6. The difference is that we introduce α_i for each w_i, where $i = 1, 2, ..., M$.

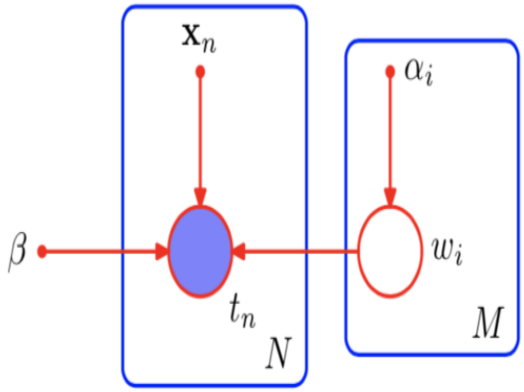
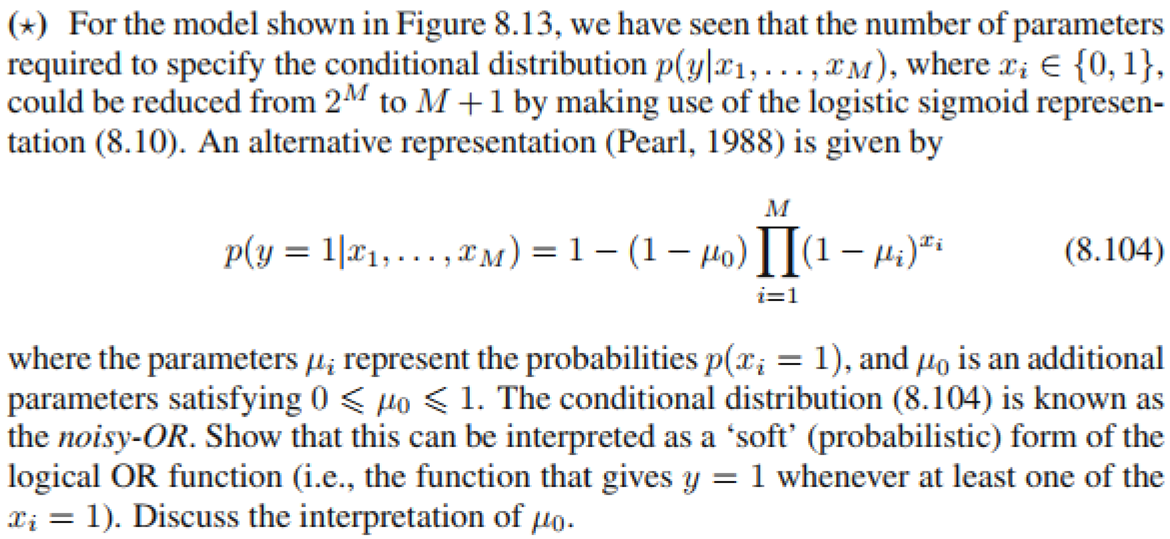


Figure 1: probabilistic graphical model corresponding to the RVM described in (7.79) and (7.80).

7 Example 8.6

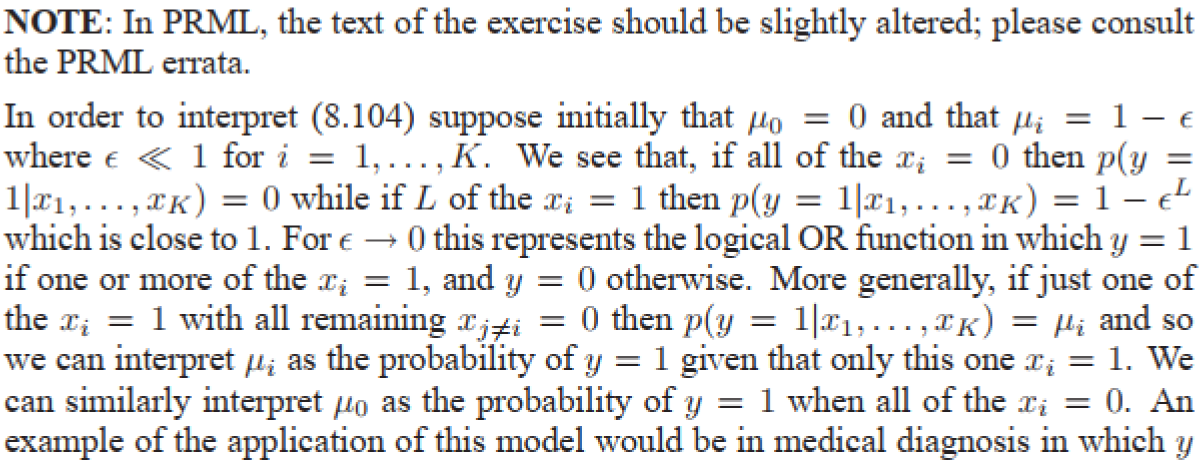


($\star$) For the model shown in Figure 8.13, we have seen that the number of parameters required to specify the conditional distribution $p(y|x_1, \ldots, x_M)$, where $x_i \in \{0, 1\}$, could be reduced from 2^M to $M+1$ by making use of the logistic sigmoid representation (8.10). An alternative representation (Pearl, 1988) is given by

$$p(y = 1|x_1, \ldots, x_M) = 1 - (1 - \mu_0) \prod_{i=1}^{M} (1 - \mu_i)^{x_i} \tag{8.104}$$

where the parameters μ_i represent the probabilities $p(x_i = 1)$, and μ_0 is an additional parameters satisfying $0 \leqslant \mu_0 \leqslant 1$. The conditional distribution (8.104) is known as the *noisy-OR*. Show that this can be interpreted as a 'soft' (probabilistic) form of the logical OR function (i.e., the function that gives $y = 1$ whenever at least one of the $x_i = 1$). Discuss the interpretation of μ_0.

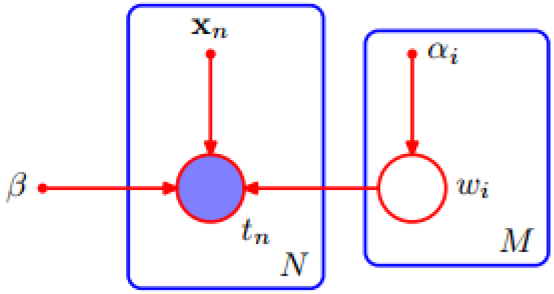
Solution



NOTE: In PRML, the text of the exercise should be slightly altered; please consult the PRML errata.

In order to interpret (8.104) suppose initially that $\mu_0 = 0$ and that $\mu_i = 1 - \epsilon$ where $\epsilon \ll 1$ for $i = 1, \ldots, K$. We see that, if all of the $x_i = 0$ then $p(y = 1|x_1, \ldots, x_K) = 0$ while if L of the $x_i = 1$ then $p(y = 1|x_1, \ldots, x_K) = 1 - \epsilon^L$ which is close to 1. For $\epsilon \to 0$ this represents the logical OR function in which $y = 1$ if one or more of the $x_i = 1$, and $y = 0$ otherwise. More generally, if just one of the $x_i = 1$ with all remaining $x_{j \neq i} = 0$ then $p(y = 1|x_1, \ldots, x_K) = \mu_i$ and so we can interpret μ_i as the probability of $y = 1$ given that only this one $x_i = 1$. We can similarly interpret μ_0 as the probability of $y = 1$ when all of the $x_i = 0$. An example of the application of this model would be in medical diagnosis in which y

Figure 5 The graphical representation of the relevance vector machine (RVM); Solution 8.5.



represents the presence or absence of a symptom, and each of the x_i represents the presence or absence of some disease. For the i^{th} disease there is a probability μ_i that it will give rise to the symptom. There is also a background probability μ_0 that the symptom will be observed even in the absence of disease. In practice we might observe that the symptom is indeed present (so that $y = 1$) and we wish to infer the posterior probability for each disease. We can do this using Bayes' theorem once we have defined prior probabilities $p(x_i)$ for the diseases.

8 Example 8.12

(⋆) www Show that there are $2^{M(M-1)/2}$ distinct undirected graphs over a set of M distinct random variables. Draw the 8 possibilities for the case of $M = 3$.

Solution

Problem 8.12 Solution

An intuitive solution is that we construct a matrix $\mathbf{A}$ with size of $M \times M$. If there is a link from node i to node j, the entry on the i-th row and j-th column of matrix $\mathbf{A}$, i.e., $A_{i,j}$, will equal to 1. Otherwise, it will equal to 0. Since the graph is undirected, the matrix $\mathbf{A}$ will be symmetric. What's more, the element on the diagonal is 0 by definition. For a undirected graph, we can use a matrix $\mathbf{A}$ to represent it. It is also a one-to-one mapping.

In other words, we equivalently count the number of possible matrix $\mathbf{A}$ satisfying the following criteria: (i) each of the entry is either 0 or 1, (ii) it is symmetric, and (iii) all of the entries on the diagonal are already determined (i.e., they all equal 0).

Using the property of symmetry, we only need to count the free variables on the lower triangle of the matrix. In the first column, there are $(M-1)$ free variables. In the second column, there are $(M-2)$ free variables. Therefore, the total free variables are given by:

$$(M-1)+(M-2)+...+0 = \frac{M(M-1)}{2}$$

Each value of these free variables has two choices, i.e., 1 or 0. Therefore, the total number of such matrix is $2^{M(M-1)/2}$. In the case of $M = 3$, there are 8 possible undirected graphs:

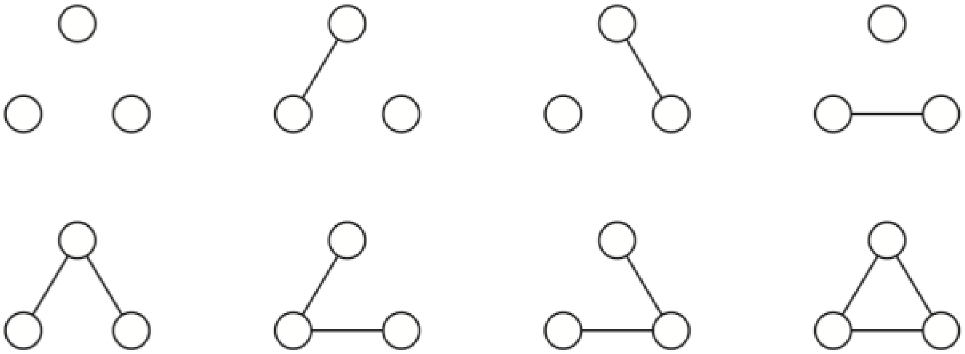
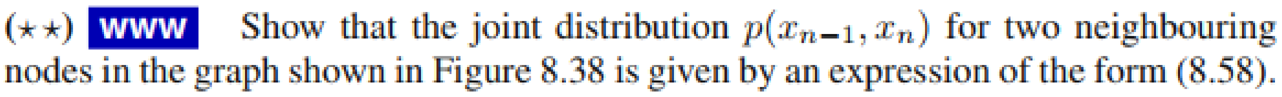


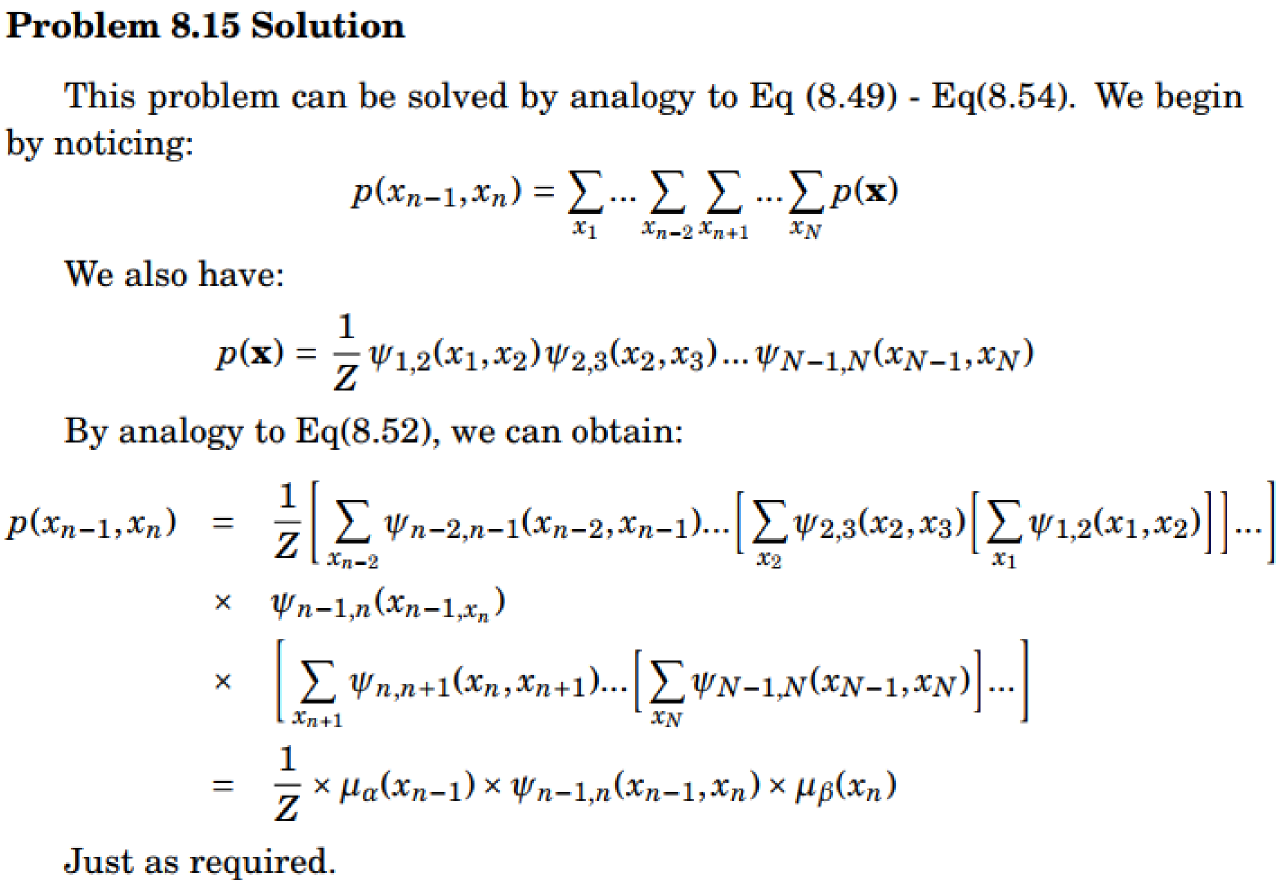
Figure 3: the undirected graph when $M = 3$

9 Example 8.15



($\star\star$) www Show that the joint distribution $p(x_{n-1}, x_n)$ for two neighbouring nodes in the graph shown in Figure 8.38 is given by an expression of the form (8.58).

Solution



Problem 8.15 Solution

This problem can be solved by analogy to Eq (8.49) - Eq(8.54). We begin by noticing:

$$p(x_{n-1}, x_n) = \sum_{x_1} \dots \sum_{x_{n-2}} \sum_{x_{n+1}} \dots \sum_{x_N} p(\mathbf{x})$$

We also have:

$$p(\mathbf{x}) = \frac{1}{Z} \psi_{1,2}(x_1, x_2)\psi_{2,3}(x_2, x_3) \dots \psi_{N-1,N}(x_{N-1}, x_N)$$

By analogy to Eq(8.52), we can obtain:

$$\begin{aligned} p(x_{n-1}, x_n) &= \frac{1}{Z}\left[\sum_{x_{n-2}} \psi_{n-2,n-1}(x_{n-2}, x_{n-1}) \dots \left[\sum_{x_2} \psi_{2,3}(x_2, x_3)\left[\sum_{x_1} \psi_{1,2}(x_1, x_2)\right]\right] \dots\right] \\ &\times \quad \psi_{n-1,n}(x_{n-1,x_n}) \\ &\times \quad \left[\sum_{x_{n+1}} \psi_{n,n+1}(x_n, x_{n+1}) \dots \left[\sum_{x_N} \psi_{N-1,N}(x_{N-1}, x_N)\right] \dots\right] \\ &= \frac{1}{Z} \times \mu_\alpha(x_{n-1}) \times \psi_{n-1,n}(x_{n-1}, x_n) \times \mu_\beta(x_n) \end{aligned}$$

Just as required.

10 Example 8.18

($\star\star$) **www** Show that a distribution represented by a directed tree can trivially be written as an equivalent distribution over the corresponding undirected tree. Also show that a distribution expressed as an undirected tree can, by suitable normalization of the clique potentials, be written as a directed tree. Calculate the number of distinct directed trees that can be constructed from a given undirected tree.

Solution

First, the distribution represented by a directed tree can be trivially be written as an equivalent distribution over an undirected tree by moralization. You can find more details in section 8.4.2.

Alternatively, now we want to represent a distribution, which is given by a directed graph, via a directed graph. For example, the distribution defined by the undirected tree in Fig.4 can be written as:

$$p(\mathbf{x}) = \frac{1}{Z}\psi_{1,3}(x_1,x_3)\,\psi_{2,3}(x_2,x_3)\,\psi_{3,4}(x_3,x_4)\,\psi_{4,5}(x_4,x_5)$$

We simply choose x_4 as the root and the corresponding directed tree is well defined by working outwards. In this case, the distribution defined by the directed tree is:

$$p(\mathbf{x}) = p(x_4)\,p(x_5|x_4)\,p(x_3|x_4)\,p(x_1|x_3)\,p(x_2|x_3)$$

Thus it is not difficult to change an undirected tree to a directed on if performing:

$$p(x_4)p(x_5|x_4) \propto \psi_{5,4},\ p(x_3|x_4) \propto \psi_{3,4},\ p(x_2|x_3) \propto \psi_{2,3},\ p(x_1|x_3) \propto \psi_{1,3},$$

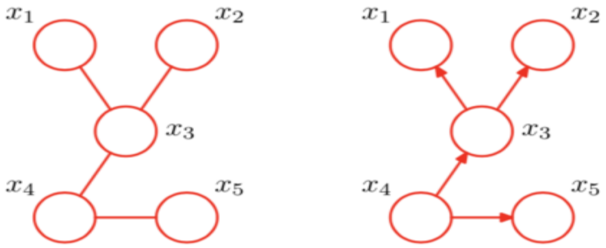


Figure 4: Example of changing an undirected tree to a directed one x_i

The symbol $\propto$ is used to represent a normalization term, which is used to guarantee the integral of PDF equal to 1. In summary, in the particular case of an undirected tree, there is only one path between any pair of nodes, and thus the maximal clique is given by a pair of two nodes in an undirected tree. This is because if we choose any three nodes x_1, x_2, x_3, according to the definition there cannot exist a loop. Otherwise there are two paths between x_1 and x_3: (i) $x_1 -> x_3$ and (ii) $x_1 -> x_2 -> x_3$. In the directed tree, each node