COMP9418 Assignment 2

Advanced Topics in Statistical Machine Learning, 18s2, UNSW Sydney

Last Update: Monday 10$^{\text{th}}$ September, 2018 at 09:09

Submission deadline: Friday September 28th, 2018 at 23:59:59

Late Submission Policy: 20% marks will be deducted from the total for each day late, up to a total of four days. If five or more days late, a zero mark will be given.

Form of Submission: You should submit your solution with the following files:

1. `solution.pdf`: Theory part;
2. `solution.ipynb`: Jupyter notebook; and
3. `model.npz`: The model in compressed .npz format.

No other formats will be accepted (scanned versions of legible handwritten answers are accepted for the theory part). There is a maximum file size cap of 20MB so make sure your submission does not exceed this size.

Submit your files using give. On a CSE Linux machine, type the following on the command-line:
`$ give cs9418 ass2 solution.pdf solution.ipynb model.npz`
Alternative, you can submit your solution via the course website
https://webcms3.cse.unsw.edu.au/COMP9418/18s2/resources/20892

Recall the guidance regarding plagiarism in the course introduction: this applies to this homework and if evidence of plagiarism is detected it may result in penalties ranging from loss of marks to suspension.

1 [50 Marks] Expectation Maximisation

Consider a model with continuous observed variables $\mathbf{x} \in \mathbb{R}^D$ and hidden variables $\mathbf{t} \in \{0,1\}^K$ and $\mathbf{z} \in \mathbb{R}^Q$. The hidden variable $\mathbf{t}$ is a K-dimensional binary random variable with a 1-of-K representation, where $t_k \in \{0,1\}$ and $\sum_k t_k = 1$, i.e. exactly one component of t_k is equal to 1 while all others are equal to 0. The prior distribution over $\mathbf{t}$ is given by

$$p(t_k = 1|\boldsymbol{\theta}) = \pi_k, \tag{1}$$

where mixing weights $\boldsymbol{\pi} = \{\pi_k\}_{k=1}^K$ satisfy $0 \leq \pi_k \leq 1$ and $\sum_{k=1}^K \pi_k = 1$. This can also be written in the form

$$p(\mathbf{t}|\boldsymbol{\theta}) = \prod_{k=1}^K \pi_k^{t_k}. \tag{2}$$

Hidden variable $\mathbf{z}$ is a Q-dimensional continuous random variable with prior distribution

$$p(\mathbf{z}|\boldsymbol{\theta}) = p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{3}$$

The conditional likelihood of $\mathbf{x}$ given $\mathbf{z}$ and $t_k = 1$ is a Gaussian defined as

$$p(\mathbf{x}|\mathbf{z}, t_k = 1, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{x}|\mathbf{W}_k\mathbf{z} + \mathbf{b}_k, \boldsymbol{\Psi}), \tag{4}$$

where $\mathbf{W}_k \in \mathbb{R}^{D\times Q}$, $\mathbf{b}_k \in \mathbb{R}^D$ and $\boldsymbol{\Psi} \in \mathbb{R}^{D\times D}$ is a *diagonal* covariance matrix. Another way to express this is

$$p(\mathbf{x}|\mathbf{z}, \mathbf{t}, \boldsymbol{\theta}) = \prod_{k=1}^{K} \mathcal{N}(\mathbf{x}|\mathbf{W}_k\mathbf{z} + \mathbf{b}_k, \boldsymbol{\Psi})^{t_k}. \tag{5}$$

Let us collectively denote the set of all observed variables by $\mathbf{X} = \{\mathbf{x}_n\}_{n=1}^N$ and hidden variables by $\mathbf{Z} = \{\mathbf{z}_n\}_{n=1}^N$ and $\mathbf{T} = \{t_n\}_{n=1}^N$. The joint distribution is denoted by $p(\mathbf{Z}, \mathbf{T}, \mathbf{X}|\boldsymbol{\theta})$, and is governed by the set of model parameters $\boldsymbol{\theta} = \{\boldsymbol{\pi}, \boldsymbol{\Psi}, (\mathbf{W}_k, \mathbf{b}_k)_{k=1}^K\}$.
In the questions below, unless otherwise stated explicitly, you must **show all your working**. Omission of details or derivations may yield a reduction in the corresponding marks.

a) [5 marks] Draw the graphical representation for this probabilistic model, making sure to include the parameters $\boldsymbol{\theta}$ in the graph. (Non-random variables can be included similarly to random variables, except that circles are not drawn around them).

b) [5 marks] In terms of K, D, Q, give an expression for the number of parameters we are required to estimate under this model.

c) [10 marks] In the E-step of the expectation maximization (EM) algorithm, we are required to compute the expected sufficient statistics of the posterior over hidden variables. The posterior responsibility of mixture component k for a data-point n is expressed as

$$r_{nk} \stackrel{\text{def}}{=} p(t_{nk} = 1|\mathbf{x}_n, \boldsymbol{\theta}^{\text{old}}) = \mathbb{E}_{p(t_{nk}|\mathbf{x}, \boldsymbol{\theta}^{\text{old}})}[t_{nk}]. \tag{6}$$

The conditional posterior over local hidden factor $\mathbf{z}_n$ is a Gaussian with mean $\mathbf{m}_{nk}$ and covariance $\mathbf{C}_{nk}$,

$$p(\mathbf{z}_n|t_{nk} = 1, \mathbf{x}_n, \boldsymbol{\theta}^{\text{old}}) = \mathcal{N}(\mathbf{z}_n|\mathbf{m}_{nk}, \mathbf{C}_{nk}). \tag{7}$$

The covariance is given by

$$\mathbf{C}_{nk} \stackrel{\text{def}}{=} \mathbf{S}_{nk} - \mathbf{m}_{nk}\mathbf{m}_{nk}^T, \tag{8}$$

where

$$\mathbf{m}_{nk} \stackrel{\text{def}}{=} \mathbb{E}_{p(\mathbf{z}_n|t_{nk}=1,\mathbf{x}_n,\boldsymbol{\theta}^{\text{old}})}[\mathbf{z}_n], \qquad \text{and} \qquad \mathbf{S}_{nk} \stackrel{\text{def}}{=} \mathbb{E}_{p(\mathbf{z}_n|t_{nk}=1,\mathbf{x}_n,\boldsymbol{\theta}^{\text{old}})}\left[\mathbf{z}_n\mathbf{z}_n^T\right]. \tag{9}$$

i) [5 marks] Give analytical expressions for the responsibilities r_{nk} and the expected sufficient statistics $\mathbf{m}_{nk}$ and $\mathbf{S}_{nk}$ in terms of the old model parameters $\boldsymbol{\theta}^{\text{old}}$.

ii) [1 marks] To de-clutter notation and simplify subsequent analysis, it is helpful to introduce *augmented* factor loading matrix and hidden factor vector,

$$\tilde{\mathbf{W}}_k \stackrel{\text{def}}{=} \begin{bmatrix}\mathbf{W}_k & \mathbf{b}_k\end{bmatrix}, \qquad \text{and} \qquad \tilde{\mathbf{z}} \stackrel{\text{def}}{=} \begin{bmatrix}\mathbf{z} & 1\end{bmatrix}^T. \tag{10}$$

Accordingly, give expressions for the sufficient statistics of the conditional posterior on augmented hidden factor vectors,

$$\tilde{\mathbf{m}}_{nk} \stackrel{\text{def}}{=} \mathbb{E}_{p(\tilde{\mathbf{z}}_n|t_{nk}=1,\mathbf{x}_n,\boldsymbol{\theta}^{\text{old}})}[\tilde{\mathbf{z}}_n], \qquad \text{and} \qquad \tilde{\mathbf{S}}_{nk} \stackrel{\text{def}}{=} \mathbb{E}_{p(\tilde{\mathbf{z}}_n|t_{nk}=1,\mathbf{x}_n,\boldsymbol{\theta}^{\text{old}})}\left[\tilde{\mathbf{z}}_n\tilde{\mathbf{z}}_n^T\right]. \tag{11}$$

Note you need only express this in terms of $\mathbf{m}_{nk}$ and $\mathbf{S}_{nk}$.

iii) [4 marks] Show that the sufficient statistics of the joint posterior factorise as follows,

$$\mathbb{E}_{p(\tilde{\mathbf{z}}_n, t_{nk}|\mathbf{x}_n, \boldsymbol{\theta}^{\text{old}})}[t_{nk}\tilde{\mathbf{z}}_n] = \mathbb{E}_{p(t_{nk}|\mathbf{x}_n, \boldsymbol{\theta}^{\text{old}})}[t_{nk}]\mathbb{E}_{p(\tilde{\mathbf{z}}_n|t_{nk}=1, \mathbf{x}_n, \boldsymbol{\theta}^{\text{old}})}[\tilde{\mathbf{z}}_n] = r_{nk}\tilde{\mathbf{m}}_{nk},$$
$$\mathbb{E}_{p(\tilde{\mathbf{z}}_n, t_{nk}|\mathbf{x}_n, \boldsymbol{\theta}^{\text{old}})}\left[t_{nk}\tilde{\mathbf{z}}_n\tilde{\mathbf{z}}_n^T\right] = \mathbb{E}_{p(t_{nk}|\mathbf{x}_n, \boldsymbol{\theta}^{\text{old}})}[t_{nk}]\mathbb{E}_{p(\tilde{\mathbf{z}}_n|t_{nk}=1, \mathbf{x}_n, \boldsymbol{\theta}^{\text{old}})}\left[\tilde{\mathbf{z}}_n\tilde{\mathbf{z}}_n^T\right] = r_{nk}\tilde{\mathbf{S}}_{nk}.$$

d) [10 marks] Write down the full expression for the *expected complete-data log likelihood* (also known as *auxiliary function*) for this model,

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) \stackrel{\text{def}}{=} \mathbb{E}_{p(\mathbf{Z}, \mathbf{T}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}})}[\log p(\mathbf{Z}, \mathbf{T}, \mathbf{X}|\boldsymbol{\theta})]. \tag{12}$$

e) [20 marks] Optimize the auxiliary function Q w.r.t. model parameters $\boldsymbol{\theta}$ to obtain M-step updates. Show all your working and highlight each individual update equation.

2 [50 Marks] Practical Part

See Jupyter notebook `comp9418_ass2.ipynb`.