PRML chapter13

1. Markov Models
   - definition of Markov model
   - the number of the considered node and the parameters
2. Hidden Markov Models
   - let z be a letent discrete variable, and A be a transition probability.
   - then we can write conditional distribution as

$$p\left(\mathbf{z}_n \mid \mathbf{z}_{n-1}, \mathbf{A}\right)=\prod_{k=1}^K \prod_{j=1}^K A_{jk}^{z_{n-1,j} z_{nk}}$$

   - define the conditional distributions of the observed variables

$$p\left(\mathbf{x}_n \mid \mathbf{z}_n, \boldsymbol{\phi}\right)=\prod_{k=1}^K p\left(\mathbf{x}_n \mid \boldsymbol{\phi}_k\right)^{z_{nk}}$$

   - Also, we can consider the homogeneous model as

$$p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta})=p\left(\mathbf{z}_1 \mid \boldsymbol{\pi}\right)\left[\prod_{n=2}^N p\left(\mathbf{z}_n \mid \mathbf{z}_{n-1}, \mathbf{A}\right)\right] \prod_{m=1}^N p\left(\mathbf{x}_m \mid \mathbf{z}_m, \boldsymbol{\phi}\right)$$

   - We can gain a better understanding of the hidden Markov model by considering it from a generative point of view
   - One of the most powerful properties of hidden Markov models is their ability to exhibit some degree of invariance to local warping of the time axis.

   (a) Maximum likelihood for the HMM
      - The likelihood function is obtained from joint distirbution $p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta})=p\left(\mathbf{z}_1 \mid \boldsymbol{\pi}\right)\left[\prod_{n=2}^N p\left(\mathbf{z}_n \mid \mathbf{z}_{n-1}, \mathbf{A}\right)\right] \prod_{m=1}^N p\left(\mathbf{x}_m \mid \mathbf{z}_m, \phi\right)$,

$$p(\mathbf{X} \mid \boldsymbol{\theta})=\sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta})$$

      - What we want to do is to maximize the likelihood function by using EM algorithms
      - E Step
        - initialize $\boldsymbol{\theta}^{\text{old}}$
        - $Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}\right)=\sum_{\mathbf{Z}} p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta}^{old}) \ln p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta})$
        - introduce the following notations

$$\gamma\left(\mathbf{z}_n\right)=p\left(\mathbf{z}_n \mid \mathbf{X}, \boldsymbol{\theta}^{old}\right)$$

$$\xi\left(\mathbf{z}_{n-1}, \mathbf{z}_n\right)=p\left(\mathbf{z}_{n-1}, \mathbf{z}_n \mid \mathbf{X}, \boldsymbol{\theta}^{old}\right)$$

- we obtain

$$Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}\right) = \sum_{k=1}^{K} \gamma\left(z_{1k}\right) \ln \pi_k + \sum_{n=2}^{N} \sum_{j=1}^{K} \sum_{k=1}^{K} \xi\left(z_{n-1,j}, z_{nk}\right) \ln A_{jk} + \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma\left(z_{nk}\right) \ln p\left(\mathbf{x}_n | \boldsymbol{\phi}_k\right)$$

- The goal is to evaluate $\gamma\left(\mathbf{z}_n\right) and \xi\left(\mathbf{z}_{n-1}, \mathbf{z}_n\right)$, which will discuss later

- M Step
  - maximize $Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}\right)$ with respect to π and A
  - we obtain

$$\pi_k = \frac{\gamma\left(z_{1k}\right)}{\sum_{j=1}^{K} \gamma\left(z_{1j}\right)}$$

$$A_{jk} = \frac{\sum_{n=2}^{N} \xi\left(z_{n-1,j}, z_{nk}\right)}{\sum_{l=1}^{K} \sum_{n=2}^{N} \xi\left(z_{n-1,j}, z_{nl}\right)}$$

  - At the initialization, we have to choose the value that obeys the controll condition.

(b) The forward-backward algorithm

- an efficient procedure for E step
- In this section, D separation and sum and product rule of probability realizes the efficiency
- First we want to evaluate $\gamma\left(z_{nk}\right)$
- $\gamma\left(\mathbf{z}_n\right) = p\left(\mathbf{z}_n | \mathbf{X}\right) = \frac{p(\mathbf{X}|\mathbf{z}_n)p(\mathbf{z}_n)}{p(\mathbf{X})}$

$$\gamma\left(\mathbf{z}_n\right) = \frac{p\left(\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{z}_n\right) p\left(\mathbf{x}_{n+1}, \ldots, \mathbf{x}_N | \mathbf{z}_n\right)}{p(\mathbf{X})} = \frac{\alpha\left(\mathbf{z}_n\right) \beta\left(\mathbf{z}_n\right)}{p(\mathbf{X})}$$

  under

$$\alpha\left(\mathbf{z}_n\right) \equiv p\left(\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{z}_n\right)$$

$$\beta\left(\mathbf{z}_n\right) \equiv p\left(\mathbf{x}_{n+1}, \ldots, \mathbf{x}_N | \mathbf{z}_n\right)$$

- derive the recursion relation to value α and β efficiently
- In terms of α,

$$\alpha\left(\mathbf{z}_n\right) = p\left(\mathbf{x}_n | \mathbf{z}_n\right) \sum_{\mathbf{z}_{n-1}} \alpha\left(\mathbf{z}_{n-1}\right) p\left(\mathbf{z}_n | \mathbf{z}_{n-1}\right)$$

  and

$$\alpha\left(\mathbf{z}_1\right) = p\left(\mathbf{x}_1, \mathbf{z}_1\right) = p\left(\mathbf{z}_1\right) p\left(\mathbf{x}_1 | \mathbf{z}_1\right) = \prod_{k=1}^{K} \left\{\pi_k p\left(\mathbf{x}_1 | \boldsymbol{\phi}_k\right)\right\}^{z_{1k}}$$

- In terms of β, we could obtain the following

$$\beta\left(\mathbf{z}_n\right) = \sum_{\mathbf{z}_{n+1}} \beta\left(\mathbf{z}_{n+1}\right) p\left(\mathbf{x}_{n+1} | \mathbf{z}_{n+1}\right) p\left(\mathbf{z}_{n+1} | \mathbf{z}_n\right)$$

- Second, we want to evaluate $\xi\left(\mathbf{z}_{n-1}, \mathbf{z}_n\right)$,

$$\xi\left(\mathbf{z}_{n-1}, \mathbf{z}_n\right) = p\left(\mathbf{z}_{n-1}, \mathbf{z}_n | \mathbf{X}\right)$$

$$= \frac{p(\mathbf{X}|\mathbf{z}_{n-1}, \mathbf{z}_n)p\left(\mathbf{z}_{n-1}, \mathbf{z}_n\right)}{p(\mathbf{X})}$$

$$= \frac{p\left(\mathbf{x}_1, \ldots, \mathbf{x}_{n-1} | \mathbf{z}_{n-1}\right) p\left(\mathbf{x}_n | \mathbf{z}_n\right) p\left(\mathbf{x}_{n+1}, \ldots, \mathbf{x}_N | \mathbf{z}_n\right) p\left(\mathbf{z}_n | \mathbf{z}_{n-1}\right) p\left(\mathbf{z}_{n-1}\right)}{p(\mathbf{X})}$$

$$= \frac{\alpha\left(\mathbf{z}_{n-1}\right) p\left(\mathbf{x}_n | \mathbf{z}_n\right) p\left(\mathbf{z}_n | \mathbf{z}_{n-1}\right) \beta\left(\mathbf{z}_n\right)}{p(\mathbf{X})}$$

(c) The sum-product algorithm for the HMM

- This session provides us the efficient procedure for E step, using sum-product algorithm.

(d) Scaling factors

- The problem we heve to discuss before using forward-backward algorithms.
- The calculator cannot treat very small number.
- We cannot calculate the data with loglikelihood like i.i.d data
- scale α and β 0 to 1

(e) Viterbi algorithm

- find most probable sequence of hidden states using max-sum algorithms

(f) Extensions of the hidden Markov model

- HMM is quite poor at generating models for the data
- A significant weakness of HMM is the way in which it represents the distribution of times for which the system remains in a given state
- HMM is poor at capturing long-range correlation
- input-output hidden markov model
- factorial hidden markov model

3. Linear Dynamical Systems

- HMM can be viewed as an extension of the mixture models of Chapter 9 to allow for sequential correlations in the data.
- we can view the linear dynamical system as a generalization of the continuous latent variable models such as probabilistic PCA and factor analysis. (z is not independent each other)
- The conditional distribution, marginal distribution, and joint distribution of variables are all Gaussian distribution

$$p\left(z_1\right)=\mathcal{N}\left(z_1|\mu_0,\mathbf{P}_0\right)$$
$$p\left(\mathbf{z_n}|\mathbf{z}_{n-1}\right)=\mathcal{N}\left(\mathbf{z_n}|\mathbf{A}\mathbf{z}_{n-1},\mathbf{\Gamma}\right)$$
$$p\left(\mathbf{x}_n|\mathbf{z}_n\right)=\mathcal{N}\left(\mathbf{x}_n|\mathbf{C}\mathbf{z}_n,\mathbf{\Sigma}\right)$$

(a) Inference in LDS

- solve the inference problem of finding the posterior marginal for a node zn given observations from x1 up to xn
- calculate the integration

$$c_n\mathcal{N}\left(\mathbf{z}_n|\boldsymbol{\mu}_n,\mathbf{V}_n\right)=\mathcal{N}\left(\mathbf{x}_n|\mathbf{C}\mathbf{z}_n,\mathbf{\Sigma}\right)\int\mathcal{N}\left(\mathbf{z}_n|\mathbf{A}\mathbf{z}_{n-1},\mathbf{\Gamma}\right)\mathcal{N}\left(\mathbf{z}_{n-1}|\boldsymbol{\mu}_{n-1},\mathbf{V}_{n-1}\right)\mathrm{d}\mathbf{z}_{n-1}$$

$$\textstyle\int\mathcal{N}\left(\mathbf{z}_n|\mathbf{A}\mathbf{z}_{n-1},\mathbf{\Gamma}\right)\mathcal{N}\left(\mathbf{z}_{n-1}|\boldsymbol{\mu}_{n-1},\mathbf{V}_{n-1}\right)\mathrm{d}\mathbf{z}_{n-1}=\mathcal{N}\left(\mathbf{z}_n|\mathbf{A}\boldsymbol{\mu}_{n-1},\mathbf{P}_{n-1}\right)$$

- under $\mathbf{P}_{n-1}=\mathbf{A}\mathbf{V}_{n-1}\mathbf{A}^{\mathrm{T}}+\mathbf{\Gamma}$,

$$\boldsymbol{\mu}_n=\mathbf{A}\boldsymbol{\mu}_{n-1}+\mathbf{K}_n\left(\mathbf{x}_n-\mathbf{C}\mathbf{A}_{n-1}\right)\mathbf{V}_n=\left(\mathbf{I}-\mathbf{K}_n\mathbf{C}\right)\mathbf{P}_{n-1}c_n=\mathcal{N}\left(\mathbf{x}_n|\mathbf{C}\mathbf{A}_{n-1},\mathbf{C}\mathbf{P}_{n-1}\mathbf{C}^{\mathrm{T}}+\mathbf{\Sigma}\right)$$

(b) Learning in LDS

- we consider the determination of these parameters using maximum likelihood