Notes on Statistical and Machine Learning

Sequential Data

Chapter: *33* **Prepared by:** *Chenxi Zhou*

This note is prepared based on *Chapter 13, Sequential Data* in Bishop (2016).

I. Introduction

1. **Sequential Data:** We consider sequential data in this chapter, where data are not independent any more and but may come in a certain order and have correlation among them.

   Examples:

   - The rainfall measurements on successive days at a particular location;
   - The daily values of a currency exchange rate;
   - The sequence of nucleotide base pairs along a strand of DNA;
   - The sequence of characters in an English sentence.

2. **Stationary and Non-stationary Data:**

   (a) *Stationary Data:* In the stationary case, the data evolves in time, but the distribution from which it is generated remains the same;

   (b) *Non-stationary Data:* For the non-stationary case, the generative distribution itself is evolving with time.

   Remark. We focus on the stationary case.

II. Markov Model

1. **Motivation:**

   (a) We expect that *recent* observations are likely to be more informative than more historical observations in predicting future values;

   (b) It would be *impractical* to consider a general dependence of future observations on *all* previous observations.

   Hence, we consider *Markov models* in which we assume that future predictions are independent of all but the most recent observations.

2. **General Product Rule in Probability:** Let $X_1, X_2, \cdots, X_n$ be a sequence of data. Their joint density function can be expressed as

$$f_{X_1,X_2,\cdots,X_n}(x_1, x_2, \cdots, x_n) = f_{X_1}(x_1) \prod_{i=2}^{n} f_{X_i \,|\, X_{i-1},\cdots,X_1}(x_i \,|\, x_{i-1}, \cdots, x_1), \tag{1}$$

using the product rule.

3. **First-order Markov Model:** If we assume that each of the conditional distributions on the right-hand side of (1) is independent of all previous observations except the most recent, we obtain the *first-order Markov model*.

   In other words, the joint density function of $X_1, X_2, \cdots, X_n$ in the first-order Markov chain is factored as

$$f_{X_1,X_2,\cdots,X_n}(x_1, x_2, \cdots, x_n) = f_{X_1}(x_1) \prod_{i=2}^{n} f_{X_i \,|\, X_{i-1}}(x_i \,|\, x_{i-1}). \tag{2}$$

   Remark 1. Under the first-order Markov model assumption, the conditional distribution for the i-th observation, given all observations up to time i, is

$$f_{X_i \,|\, X_{i-1},\cdots,X_1}(x_i \,|\, x_{i-1}, x_{i-2}, \cdots, x_1) = f_{X_i \,|\, X_{i-1}}(x_i \,|\, x_{i-1}),$$

   for all $i = 2, 3, \cdots, n$.

   Remark 2. If we use a first-order Markov model to predict the next observation in a sequence, the distribution of predictions depends only on the value of the immediately preceding observation and is independent of all earlier observations.

4. **Higher-order Markov Model:** If we assume that the current observation depends the past few observations, we obtain the higher-order Markov model.

   For example, the density function of $X_1, X_2, \cdots, X_n$ under the second-order Markov model can be factored as

$$f_{X_1,X_2,\cdots,X_n}(x_1, x_2, \cdots, x_n) = f_{X_1}(x_1) f_{X_2 \,|\, X_1}(x_2 \,|\, x_1) \prod_{i=3}^{n} f_{X_i \,|\, X_{i-1},X_{i-2}}(x_i \,|\, x_{i-1}, x_{i-2}),$$

   where each observation is influenced by two previous observations.

5. **Model Complexity of m-th Order Markov Model:** Suppose the observation are discrete random variables having K states.

   (a) In a first-order Markov model, the conditional distribution $f_{X_i \,|\, X_{i-1}}$ is specified by a set of $K-1$ parameters for each of the K states of X_{i-1}. In total, we have $K(K-1)$ parameters in the model.

(b) Consider an m-th order Markov chain where the joint density function is built up from conditionals

$$f_{X_i \mid X_{i-1}, X_{i-2}, \cdots, X_{i-m}}.$$

If the conditional distributions are represented by general conditional probability tables, then the number of parameters in such a model will have $\mathcal{O}(K^m)$ parameters.

Remark. Because the number of parameters grows exponentially with m, it will often render this approach *impractical* for larger values of m.

6. **State-Space Model:** For each observation X_i that is observed at time i, we introduce the corresponding latent variable Z_i (which may be of different type or dimensionality to the observed variable) and assume these latent variables form a Markov chain, which gives rise to a *state-space model*. A diagram of the state-space model is shown in Figure 1.

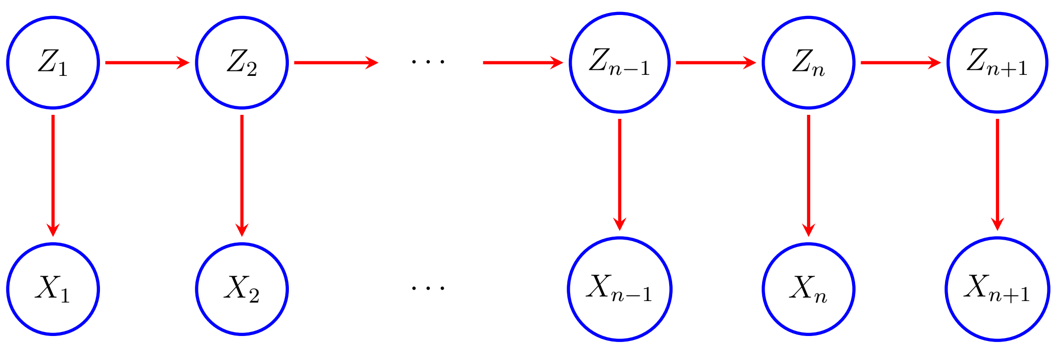

Figure 1: Diagram associated with the state-space model.

The join density function of $X_1, X_2, \cdots, X_n, Z_1, Z_2, \cdots, Z_n$ is given by

$$\begin{aligned} &f_{X_1, \cdots, X_n, Z_1, \cdots, Z_n}(x_1, \cdots, x_n, z_1, \cdots, z_n) \\ &\qquad = f_{Z_1}(z_1) \left[\prod_{i=2}^{n} f_{Z_i|Z_{i-1}}(z_i \mid z_{i-1})\right] \left[\prod_{i=1}^{n} f_{X_i|Z_i}(x_i \mid z_i)\right]. \end{aligned}$$

Remark 1. In such a state-space model, Z_{i-1} and Z_{i+1} are independent given Z_i, for all $i = 2, 3, \cdots, n$.

Remark 2. There is always a path connecting any two observed variables X_i and X_j via the latent variables, and that this path is never blocked.

III. Hidden Markov Models

1. **Hidden Markov Models:** If the latent variables in the state-space model are discrete random variables, we obtain the *hidden Markov model*.

2. **Homogeneity Assumption:** We assume that all parameters in the model (see below) are independent of the time stamp. In other words, for all $i \neq j$,
   - the parameters appearing in $f_{Z_i|Z_{i-1}}$ are identical to those appearing in $f_{Z_j|Z_{j-1}}$, and
   - the parameters appearing in $f_{X_i \,|\, Z_i}$ are identical to those appearing in $f_{X_j \,|\, Z_j}$.

3. **Model Specification — Distribution of Z_1:** We assume that Z_1 follows a multinomial distribution with K components and can take any values in $\{1, 2, \cdots, K\}$. We introduce the following notation

$$\widetilde{Z}_{1,k} = \begin{cases} 1, & \text{if } Z_1 = k, \\ 0, & \text{otherwise,} \end{cases}$$

   and let

$$\widetilde{Z}_1 := (\widetilde{Z}_{1,1}, \widetilde{Z}_{1,2}, \cdots, \widetilde{Z}_{1,K}) \in \mathbb{R}^K.$$

   Since the initial latent variable Z_1 does *not* have a parent node, we consider its marginal distribution and let

$$\pi_k := \mathbb{P}(Z_1 = k) = \mathbb{P}(\widetilde{Z}_{1,k} = 1), \qquad \text{for all } k = 1, 2, \cdots, K.$$

   Note that $\pi_1, \pi_2, \cdots, \pi_K$ must satisfy

$$\pi_k \geq 0 \qquad \text{for all } k = 1, 2, \cdots, K$$

   and

$$\sum_{k=1}^{K} \pi_k = 1. \tag{3}$$

   We let $\boldsymbol{\pi} := (\pi_1, \pi_2, \cdots, \pi_K)^\top \in \mathbb{R}^K$.

   Remark. Note that exactly one entry of $\widetilde{Z}_1$ is equal to 1 and all others are equal to 0.

4. **Model Specification — Conditional Distribution of Latent Variables:** We assume that, conditional on Z_{i-1}, Z_i follows a multinomial distribution with K components, for all $i = 2, 3, \cdots$, and can take any values in $\{1, 2, \cdots, K\}$. Similar to $\widetilde{Z}_1$, we introduce the following notation: conditional on Z_{i-1}, let

$$\widetilde{Z}_{i,k} = \begin{cases} 1, & \text{if } Z_i = k, \\ 0, & \text{otherwise,} \end{cases}$$

   for all $k = 1, 2, \cdots, K$ and $i = 2, 3, \cdots$, and collectively, let

$$\widetilde{Z}_i := (\widetilde{Z}_{i,1}, \widetilde{Z}_{i,2}, \cdots, \widetilde{Z}_{i,K}),$$

for all $i = 2, 3, \cdots$.

Under the multinomial distribution assumption, the conditional distribution of Z_i, given Z_{i-1}, corresponds to a table of numbers, denoted by $\mathbf{A} \in \mathbb{R}^{K \times K}$, where

$$A_{j,k} = \mathbb{P}(Z_i = k \mid Z_{i-1} = j) = \mathbb{P}(\widetilde{Z}_{i,k} = 1 \mid \widetilde{Z}_{i-1,j} = 1),$$

for all $j, k = 1, 2, \cdots, K$. Note that $A_{j,k}$'s satisfy

$$A_{j,k} \geq 0 \qquad \text{and} \qquad \sum_{k=1}^{K} A_{j,k} = 1. \tag{4}$$

The elements of $\mathbf{A}$, $A_{j,k}$'s, are called the *transition probabilities*.

Remark. Due to the constraints in (4), the matrix $\mathbf{A}$ has $K(K-1)$ independent parameters.

5. **Model Specification — Conditional Distribution of Observed Variables:** Conditioning on $Z_i = k$, we let the density function of X_i be

$$f_{X_i \mid Z_i = k}(\cdot \mid \phi_k),$$

where ϕ_k is a set of parameters governing the density function. Collectively, we write $\boldsymbol{\phi} := (\phi_1, \phi_2, \cdots, \phi_K)$.

6. **Joint Density Function of Latent and Observed Variables:** Under the assumptions and notation above, the joint probability density function of the observed variables, $X_1, X_2, \cdots, X_n$, and the latent variables, $Z_1, Z_2, \cdots, Z_n$, is given by

$$\begin{aligned} & f_{X_1, \cdots, X_n, \widetilde{Z}_1, \cdots, \widetilde{Z}_n}(x_1, x_2, \cdots, x_n, \tilde{\mathbf{z}}_1, \tilde{\mathbf{z}}_2, \cdots, \tilde{\mathbf{z}}_n \mid \boldsymbol{\theta}) \\ & \qquad = \left[\prod_{k=1}^{K} \pi_k^{\tilde{z}_{1,k}}\right] \left[\prod_{i=2}^{n} \prod_{k=1}^{K} \prod_{j=1}^{K} A_{j,k}^{\tilde{z}_{i-1,j} \tilde{z}_{i,k}}\right] \left[\prod_{i=1}^{n} \prod_{k=1}^{K} \big(f_{X_i \mid Z_i = k}(x_i \mid \phi_k)\big)^{\tilde{z}_{i,k}}\right], \end{aligned} \tag{5}$$

where $\boldsymbol{\theta} := (\mathbf{A}, \boldsymbol{\pi}, \boldsymbol{\phi})$.

Remark 1. We used the homogeneity assumption above; more explicitly,

- all conditional distributions governing the latent variables share the same parameters $\mathbf{A}$, and
- all conditional distributions governing the observed variables share the same parameters ϕ_k's.

These parameters only depends on the appropriate states but not the time stamps.

Remark 2. The model here is tractable for a wide range of $f_{X_i \mid Z_i = k}(\cdot \mid \phi_k)$ including discrete tables, Gaussians, mixtures of Gaussians, or even neural networks.

7. **Sampling from Hidden Markov Model:** We can obtain samples from a hidden Markov model as follows:

(a) Choose the initial latent variable Z_1 with probabilities governed by the parameters $\pi_1, \pi_2, \cdots, \pi_K$, and then sample the corresponding observation X_1;

(b) Choose the state of the variable Z_2 according to the transition probabilities $\mathbb{P}(Z_2 \mid Z_1)$ using Z_1: supposing $Z_1 = j$ for some $j \in \{1, 2, \cdots, K\}$, we sample Z_2 according to probabilities $(A_{j,1}, A_{j,2}, \cdots, A_{j,K})$;

(c) Once we know Z_2, we can draw a sample for X_2 and also sample the next latent variable Z_3 and so on.

Remark. The sampling procedure outlined here is an example of *ancestral sampling* for a directed graphical model.

8. **Difficulties in Parameter Estimation Using Maximum Likelihood:** Suppose the observed data are given as $\mathbf{X} := \{x_1, x_2, \cdots, x_n\}$ and the hidden data are given as $\widetilde{\mathbf{Z}} := \{\widetilde{\mathbf{z}}_1, \widetilde{\mathbf{z}}_2, \cdots, \widetilde{\mathbf{z}}_n\}$. The likelihood function can be obtained by marginalizing (5) over the latent variables

$$L(\boldsymbol{\theta} \mid \mathbf{X}) = \sum_{\widetilde{\mathbf{Z}}} f_{X_1, \cdots, X_n, \widetilde{Z}_1, \cdots, \widetilde{Z}_n}(x_1, x_2, \cdots, x_n, \tilde{\mathbf{z}}_1, \tilde{\mathbf{z}}_2, \cdots, \tilde{\mathbf{z}}_n \mid \boldsymbol{\theta}). \tag{6}$$

Direct maximization of L above is intractable by noting that the summation in (6) corresponds to summing over K^n terms.

9. **EM Algorithm to Estimate $\boldsymbol{\theta}$ — Overview:** We use the expectation-maximization (EM) algorithm to maximize the likelihood function in hidden Markov models.

The EM algorithm starts with some initial selection for the model parameters, which we denote by $\boldsymbol{\theta}^{\text{old}}$. Then,

(a) in the E step, we take $\boldsymbol{\theta}^{\text{old}}$ and find the conditional distribution of the latent variables $f(\widetilde{\mathbf{Z}} \mid \mathbf{X}, \boldsymbol{\theta}^{\text{old}})$, using which we evaluate the expectation of the logarithm of the complete-data likelihood function;

(b) in the M step, we maximize the expectation with respect to the parameters.

10. **EM Algorithm to Estimate $\boldsymbol{\theta}$ — E Step:** With the current value of the parameters $\boldsymbol{\theta}^{\text{old}}$, in the E step, we evaluate the expectation of the logarithm of the complete data likelihood function as a function of the parameters $\boldsymbol{\theta}$; that is, we compute

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) := \sum_{\widetilde{\mathbf{Z}}} f(\widetilde{\mathbf{Z}} \mid \mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \log f(\mathbf{X}, \widetilde{\mathbf{Z}} \mid \boldsymbol{\theta}), \tag{7}$$

where we omit the subscripts for density functions. Note that

$$\begin{aligned}\log f(\mathbf{X}, \widetilde{\mathbf{Z}} \mid \boldsymbol{\theta}) = & \sum_{k=1}^{K} \tilde{z}_{1,k} \log \pi_k + \sum_{i=2}^{n} \sum_{k=1}^{K} \sum_{j=1}^{K} \tilde{z}_{i-1,j} \tilde{z}_{i,k} \log A_{j,k} \\ & + \sum_{i=1}^{n} \sum_{k=1}^{K} \tilde{z}_{i,k} \log f_{X_i \mid Z_i = k}(x_i \mid \phi_k).\end{aligned}$$

Then, it can be shown that

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \sum_{k=1}^{K} \gamma(\widetilde{Z}_{1,k}) \log \pi_k + \sum_{i=2}^{n} \sum_{k=1}^{K} \sum_{j=1}^{K} \xi(\widetilde{Z}_{i-1,j}, \widetilde{Z}_{i,k}) \log A_{j,k} + \sum_{i=1}^{n} \sum_{k=1}^{K} \gamma(\widetilde{Z}_{i,k}) \log f_{X_i \mid Z_i = k}(x_i \mid \phi_k),$$

where

$$\begin{aligned} \gamma(\widetilde{Z}_{i,k}) &= \mathbb{P}(\widetilde{Z}_{i,k} = 1 \mid \mathbf{X}, \boldsymbol{\theta}^{\text{old}}), \\ \xi(\widetilde{Z}_{i-1,j}, \widetilde{Z}_{i,k}) &= \mathbb{P}(\tilde{Z}_{i-1,j} = 1, \widetilde{Z}_{i,k} = 1 \mid \mathbf{X}, \boldsymbol{\theta}^{\text{old}}). \end{aligned}$$

11. **EM Algorithm to Estimate $\boldsymbol{\theta}$ — M Step:** In the M step, we maximize $Q(\,\cdot\,, \boldsymbol{\theta}^{\text{old}})$ with respect to the first argument by treating the quantities $\gamma(\widetilde{Z}_{i,k})$'s and $\xi(\widetilde{Z}_{i-1,j}, \widetilde{Z}_{i,k})$'s as known. Recall that $\boldsymbol{\theta}$ involves three parts, namely, $\boldsymbol{\pi}$, $\mathbf{A}$, and $\boldsymbol{\phi}$.

   (a) *Maximizing Over $\boldsymbol{\pi}$:* Maximizing over $\boldsymbol{\pi}$ under the constraint $\sum_{k=1}^{K} \pi_k = 1$ yields

   $$\widehat{\pi}_k = \frac{\gamma(\widetilde{Z}_{1,k})}{\sum_{j=1}^{K} \gamma(\widetilde{Z}_{1,j})}, \qquad \text{for all } k = 1, 2, \cdots, K.$$

   (b) *Maximizing Over $\mathbf{A}$:* Maximizing over $\mathbf{A}$ under the constraint $\sum_{k=1}^{K} A_{j,k} = 1$, for all $j = 1, 2, \cdots, K$, yields

   $$\widehat{A}_{j,k} = \frac{\sum_{i=2}^{n} \xi(\widetilde{Z}_{i-1,j}, \widetilde{Z}_{i,k})}{\sum_{\ell=1}^{K} \sum_{i=2}^{n} \xi(\widetilde{Z}_{i-1,j}, \widetilde{Z}_{i,\ell})}, \qquad \text{for all } j, k = 1, 2, \cdots, K.$$

   (c) *Maximizing Over $\boldsymbol{\phi}$:* Maximizing over $\boldsymbol{\phi}$ is data dependent and depends on the specific form of the conditional density functions p_{ϕ_k}'s.

   For example, if $f_{X_i \mid Z_i = k}(\,\cdot \mid \phi_k)$ is the density function of the normal distribution with mean $\boldsymbol{\mu}_k$ and covariance matrix $\boldsymbol{\Sigma}_k$, maximizing $Q(\,\cdot\,, \boldsymbol{\theta}^{\text{old}})$ with respect to $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ yields the following estimators

   $$\begin{aligned} \widehat{\boldsymbol{\mu}}_k &= \frac{\sum_{i=1}^{n} \gamma(\widetilde{Z}_{i,k}) x_i}{\sum_{i=1}^{n} \gamma(\widetilde{Z}_{i,k})}, \\ \widehat{\boldsymbol{\Sigma}}_k &= \frac{\sum_{i=1}^{n} \gamma(\widetilde{Z}_{i,k})(x_i - \widehat{\boldsymbol{\mu}}_k)(x_i - \widehat{\boldsymbol{\mu}}_k)^\top}{\sum_{i=1}^{n} \gamma(\widetilde{Z}_{i,k})}, \end{aligned}$$

   respectively, for all $k = 1, 2, \cdots, K$.

Remark. The parameters $\boldsymbol{\pi}$ and $\mathbf{A}$ must be initialized in a way such that the constraints (3) and (4) are satisfied.

Any elements of $\boldsymbol{\pi}$ or $\mathbf{A}$ that are set to zero initially will remain zero in subsequent EM updates.

References

Bishop, Christopher M (Aug. 2016). *Pattern Recognition and Machine Learning*. en. Springer New York.