VE414 Lecture 23

Jing Liu

UM-SJTU Joint Institute

November 26, 2019

Q: How can apply Bayesian analysis to data

$$X_1, X_2, \ldots, X_T$$

that is represented as a sequence of observations over time?

- For example,
  - The weather over time.
  - The location of a missile that has just been launched over time.
  - The words that someone spoke over time.

Q: What is the major difference between such data and the ones we considered?

$$\mathbf{X} \mid \mathbf{Y} \sim f_{\mathbf{X}|\mathbf{Y}} \qquad \text{where} \quad f_{\mathbf{X}|\mathbf{Y}} = \prod_{i=1}^{n} f_{X_i|\mathbf{Y}}$$

$$\mathbf{Y} \sim f_{\mathbf{Y}}$$

X_i's we have considered so far are independent and identically distributed.

- The easiest way to model this dependency is to use a Markov model.
- Given a set of possible states that we can observe over time, $t = 1, 2, \ldots, T$.

$$\mathcal{S} = \{s_1, s_2, \ldots, s_j, \ldots, s_{|\mathcal{S}|}\}$$

where $|\mathcal{S}|$ denotes the "size" of the state space.

- For example, suppose there are only three possible states for weather

$$\mathcal{S} = \{\text{sunny}, \text{cloudy}, \text{rainy}\}$$

then $|\mathcal{S}| = 3$, and a possible realisation over 5 days, i.e. our data, could be

$$\{x_1 = 1; \quad x_2 = 2 : \quad x_3 = 2; \quad x_4 = 3; \quad x_5 = 2\}$$

where $X_i = \begin{cases} 1 & \text{if sunny}, \\ 2 & \text{if cloudy}, \\ 3 & \text{if rainy}. \end{cases}$ is the random variable maps the weather to $\mathbb{Z}_+$

- The Markov chain in the MCMC section is the simplest Markov model.
- In our Markov model now, we make two assumptions:
  - the *Memoryless* assumption
  - the *Stationary* assumption

  which allow us to model the data in a more tractable fashion.
- *Memoryless* is about the probability of being in a state s_j at time $t+1$ only depends on the state at t, it means for all $1 \leq t < T$ and any $1 \leq j \leq |\mathcal{S}|$

$$\Pr\left(X_{t+1}=j \mid X_t, X_{t-1}, \ldots, X_1\right)=\Pr\left(X_{t+1}=j \mid X_t\right)$$

- The intuition behind this assumption is the that the state at time t contains "enough" information to learn about the past, and predict the future.
- *Stationary* is about the probability being invariant in addition to memoryless,

$$\Pr\left(X_{t+1}=j \mid X_t\right)=\Pr\left(X_2=j \mid X_1\right)$$

- The intuition behind this assumption is that the system is "stable".

- As a convention, we will assume there is an unknown but fixed initial state

$$\Pr(X_1 = j \mid X_0 = s_0) = f_S(s_j)$$

  that is, we assume all data sequences have the same initial state s_0.
- Recall previously we specify a scalar-valued likelihood function

$$f_{\mathbf{X}|\mathbf{Y}}$$

  for the data generating process, For a Markov model, a transition matrix

$$[p]_{kj} = \Pr(X_{t+1} = j \mid X_t = k)$$

- For example, the transition matrix $\mathbf{P}$ for the weather system could be

$$\begin{array}{cc} & \begin{array}{ccc} s_1 & s_2 & s_3 \end{array} \\ \begin{array}{c} s_1 \\ s_2 \\ s_3 \end{array} & \begin{bmatrix} 0.8 & 0.1 & 0.1 \\ 0.2 & 0.6 & 0.2 \\ 0.1 & 0.2 & 0.7 \end{bmatrix} \end{array}$$

- For notational convenience, we put $\mathbf{P}$ and f_S into a single matrix $\mathbf{A}$, e.g.

$$\mathbf{A} = \begin{bmatrix} 0 & 0.\dot{3} & 0.\dot{3} & 0.\dot{3} \\ 0 & 0.8 & 0.1 & 0.1 \\ 0 & 0.2 & 0.6 & 0.2 \\ 0 & 0.1 & 0.2 & 0.7 \end{bmatrix}$$

  where f_S is taken to be uniform in this case.
- Given the matrix $\mathbf{A}$, we can compute the probability of observing the data

$$\{x_1, \ldots, x_T\}$$

  for example, using the above transition matrix

$$\mathbf{x} = \begin{bmatrix} 1 \\ 2 \\ 2 \\ 3 \\ 2 \end{bmatrix} \implies \Pr(\mathbf{x}) = \frac{1}{3} \cdot \frac{1}{10} \cdot \frac{6}{10} \cdot \frac{2}{10} \cdot \frac{2}{10}$$

- Using this notation, the probability of observing an arbitrary data sequence is

$$
\begin{aligned}
\Pr\left(\mathbf{x} \mid \mathbf{A}\right) &= \Pr\left(x_T, x_{T-1} \ldots x_1 \mid \mathbf{A}\right) \\
&= \Pr\left(x_T \mid x_{T-1}, x_{T-2} \ldots, x_1, \mathbf{A}\right) \\
&\quad\ \Pr\left(x_{T-1} \mid x_{T-2}, x_{T-3} \ldots x_1, \mathbf{A}\right) \cdots \Pr\left(x_1 \mid x_0, \mathbf{A}\right) \\
&= \Pr\left(x_T \mid x_{T-1}, \mathbf{A}\right) \Pr\left(x_{T-1} \mid x_{T-2}, \mathbf{A}\right) \cdots \Pr\left(x_1 \mid x_0, \mathbf{A}\right) \\
&= \prod_{t=0}^{T-1} \Pr\left(x_{t+1} \mid x_t, \mathbf{A}\right) = \prod_{t=0}^{T-1} A_{x_t x_{t+1}}
\end{aligned}
$$

- Of course, in practice, we seek information about the matrix $\mathbf{A}$ given data

$$\{x_1, \ldots, x_T\}$$

Q: What would a frequentist do?

$$\arg\max_{\mathbf{A}} \mathcal{L}\left(\mathbf{A}; \mathbf{x}\right)$$

- Of course, the likelihood is simply

$$\mathcal{L}(\mathbf{A}; \mathbf{x}) = \prod_{t=0}^{T-1} A_{x_t x_{t+1}}$$

- If we denote the transition counts as

$$N_{kj} = n_{kj}$$

that is, the number of times in the data that the system goes from s_k to s_j,

$$\mathcal{L}(\mathbf{A}; \mathbf{x}) = \prod_{k=1}^{|\mathcal{S}|+1} \prod_{j=1}^{|\mathcal{S}|+1} A_{kj}^{n_{kj}}$$

- Taking log, and differentiating with respect to A_{kj}, we have

$$\ell = \sum_{k,j} n_{kj} \ln A_{kj} \implies \frac{\partial \ell}{\partial A_{kj}} = \frac{n_{kj}}{A_{kj}}$$

Q: What has gone wrong?

- We have failed to notice that $\mathbf{A}$ cannot be any matrix, e.g.

$$\mathbf{A} = \begin{bmatrix} 0 & 0.\dot{3} & 0.\dot{3} & 0.\dot{3} \\ 0 & 0.8 & 0.1 & 0.1 \\ 0 & 0.2 & 0.6 & 0.2 \\ 0 & 0.1 & 0.2 & 0.7 \end{bmatrix}$$

- By construction, the first column needs to zero, and for $k = 1, 2 \ldots, |\mathcal{S}| + 1$,

$$\sum_{j=2}^{|\mathcal{S}|+1} A_{kj} = 1 \qquad \text{and} \qquad 0 \leq A_{kj} \leq 1$$

which means, we need to pick one of the transition probabilities for each k,

$$A_{k2} = 1 - \sum_{j=3}^{|\mathcal{S}|+1} A_{kj}$$

to express in terms of the others in the same row.

- Therefore we should maximises the following

$$\ell = \sum_{k=1}^{|\mathcal{S}|+1} \left(n_{k2} \ln \left(1 - \sum_{j=3}^{|\mathcal{S}|+1} A_{kj} \right) + \sum_{j=3}^{|\mathcal{S}|+1} n_{kj} \ln A_{kj} \right)$$

with respect to A_{kj}, for $k = 1, \ldots, |\mathcal{S}| + 1$ and $j = 3, \ldots |\mathcal{S}| + 1$

$$\frac{\partial \ell}{\partial A_{kj}} = -\frac{n_{k2}}{1 - \sum_{j=3}^{|\mathcal{S}|+1} A_{kj}} + \frac{n_{kj}}{A_{kj}}$$

$$\implies \frac{n_{kj}}{\hat{A}_{kj}} = \frac{n_{k2}}{1 - \sum_{j=3}^{|\mathcal{S}|+1} \hat{A}_{kj}} \implies \frac{n_{kj}}{n_{k2}} = \frac{\hat{A}_{kj}}{1 - \sum_{j=3}^{|\mathcal{S}|+1} \hat{A}_{kj}}$$

- Thus $\hat{A}_{kj}$ is proportional to n_{kj}, and the MLE is given by $\hat{A}_{kj} = \dfrac{n_{kj}}{\sum_{j=2}^{|\mathcal{S}|+1} n_{kj}}$.

Q: How would a Bayesian estimate $\mathbf{A}$?

$$f_{\mathbf{A}|\mathbf{X}} \propto \mathcal{L} \cdot f_{\mathbf{A}} \qquad \text{where} \quad \mathcal{L} = \prod_{k=1}^{|\mathcal{S}|+1} \prod_{j=2}^{|\mathcal{S}|+1} A_{kj}^{n_{kj}}$$

Q: How about the prior?

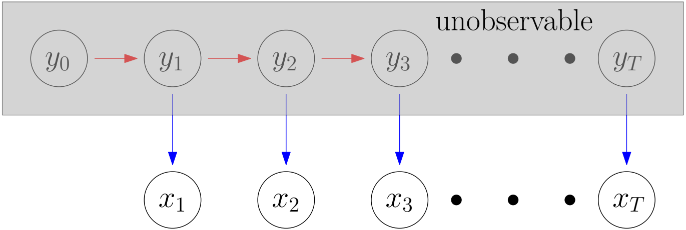
- One nature candidate is the generalisation of the beta distribution

$$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1} \qquad \text{which is for one} \qquad y \in [0,1]$$

namely, the Dirichlet distribution, one for each row of $\mathbf{A}$, with independence

$$f_{\mathbf{A}}(\mathbf{A}) = \prod_{k=1}^{|\mathcal{S}|+1} \left(\frac{\Gamma\left(\sum_{j=2}^{|\mathcal{S}|+1} \alpha_{kj}\right)}{\prod_{j=2}^{|\mathcal{S}|+1} \Gamma(\alpha_{kj})} \prod_{j=2}^{|\mathcal{S}|+1} A_{kj}^{\alpha_{kj}-1} \right)$$

- Markov Models are powerful in terms of modelling time series data, but we are often interested in a system that we cannot observe the states directly.



where we are interested in the unobservable Y_i instead of the observable X_i.

- For example,
  - Having data on the weather but interested in the location.
  - Having radar data on the missile but interested in the actual position.
  - Having predicted words but interested in the actual words.
- Such situations can be modelled by a Hidden Markov Model (HMM).

- In a HMM, there is an unobservable sequence of states:

$$\{y_1, y_2, \ldots, y_T\}$$

which is a process follows a Markov Model with a state transition matrix

$$\mathbf{A}$$

that is, the state space of Y_t is discrete and finite,

$$\mathcal{S} = \{s_1, s_2, \ldots, s_j, \ldots, s_{|\mathcal{S}|}\}$$

and the transition probability is given by the elements of $\mathbf{A}$

$$\Pr(Y_{t+1} = j \mid Y_t = k) = A_{kj}$$

- The observable X_t also has a discrete and finite state space

$$\mathcal{O} = \{o_1, o_2, \ldots, o_q, \ldots o_{|\mathcal{O}|}\}$$

- In addition to the two assumptions in a Markov model,
  - the *Memoryless* assumption
  - the *Stationary* assumption

  we make the *output independence* assumption in a HMM, formally

$$\Pr\left(X_t \mid Y_1, Y_2, \ldots, Y_t, X_1, X_2, \ldots, X_{t-1}\right) = \Pr\left(X_t \mid Y_t\right)$$

- At each time step t, the observed data is a random variable depends on Y_t,

$$X_t \mid Y_t \sim f_{X|Y}$$

  that is, it is only state-dependent but it is time-independent, that is

$$\Pr\left(X_t = q \mid Y_t = j\right) = B_{jq}$$

  where B_{jq} is the jth-row qth-column element of a matrix $\mathbf{B}$ that contains the probability of having output o_q given the hidden state s_j.

- The joint probability is given by

$$\Pr(\mathbf{x}, \mathbf{y} \mid \mathbf{A}, \mathbf{B}) = \Pr(\mathbf{x} \mid \mathbf{y}, \mathbf{A}, \mathbf{B}) \Pr(\mathbf{y} \mid \mathbf{A}, \mathbf{B})$$

- Since $\mathbf{y}$ is a discrete random variable, the likelihood of the observed data is

$$\Pr(\mathbf{x} \mid \mathbf{A}, \mathbf{B}) = \sum_{\mathbf{y}} \Pr(\mathbf{x} \mid \mathbf{y}, \mathbf{A}, \mathbf{B}) \Pr(\mathbf{y} \mid \mathbf{A}, \mathbf{B})$$

where the summation is over all possible sequence of states.

- Using the three assumptions of HMM, we have

$$\begin{aligned}\Pr(\mathbf{x} \mid \mathbf{A}, \mathbf{B}) &= \sum_{\mathbf{y}} \left[\prod_{t=1}^{T} \Pr(x_t \mid y_t, \mathbf{B})\right] \cdot \left[\prod_{t=1}^{T} \Pr(y_t \mid y_{t-1}, \mathbf{A})\right] \\ &= \sum_{\mathbf{y}} \left[\prod_{t=1}^{T} B_{y_t x_t}\right] \cdot \left[\prod_{t=1}^{T} A_{y_{t-1} y_t}\right]\end{aligned}$$

which gives the probability or the likelihood of observing $\mathbf{x}$ given $\mathbf{A}$ and $\mathbf{B}$.

- Computing the likelihood is simple but expensive if it is done naively,

$$\Pr\left(\mathbf{x} \mid \mathbf{A}, \mathbf{B}\right) = \sum_{\mathbf{y}} \left[\prod_{t=1}^{T} B_{y_t x_t}\right] \cdot \left[\prod_{t=1}^{T} A_{y_{t-1} y_t}\right]$$

since it is over all possible $\mathbf{y}$, which means $|\mathcal{S}|^T$ possibilities.

- Let us denote the probability of having the observations up to time t by

$$\alpha_j(t) = \Pr\left(x_1, x_2, \ldots, x_t, Y_t = j \mid \mathbf{A}, \mathbf{B}\right)$$

where the hidden state is at j.

- Notice those probabilities can be recursively computed,

$$\alpha_i(0) = A_{1i}; \qquad \alpha_j(t) = \sum_{i=1}^{|\mathcal{S}|} \alpha_i(t-1) A_{ij} B_{j x_t}$$

for $j = 1 \ldots |\mathcal{S}|$ and $t = 1 \ldots T$.

- Once we have those α's, the likelihood can be evaluated as

$$\begin{aligned}\Pr(\mathbf{x} \mid \mathbf{A}, \mathbf{B}) &= \Pr(x_1, x_2, \ldots, x_T \mid \mathbf{A}, \mathbf{B}) \\ &= \sum_{j=1}^{|\mathcal{S}|} \Pr(x_1, x_2, \ldots, x_T, Y_T = j \mid \mathbf{A}, \mathbf{B}) = \sum_{j=1}^{|\mathcal{S}|} \alpha_j(T)\end{aligned}$$

- This, which is linear in terms of $|\mathcal{S}| \cdot T$, is called the forward procedure.
- The most common quest in using HMM is to estimate $\mathbf{y}$ given $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{x}$.

$$\begin{aligned}\arg\max_{\mathbf{y}} \Pr(\mathbf{y} \mid \mathbf{x}, \mathbf{A}, \mathbf{B}) &= \arg\max_{\mathbf{y}} \frac{\Pr(\mathbf{y}, \mathbf{x} \mid \mathbf{A}, \mathbf{B})}{\Pr(\mathbf{x} \mid \mathbf{A}, \mathbf{B})} \\ &= \arg\max_{\mathbf{y}} \Pr(\mathbf{y}, \mathbf{x} \mid \mathbf{A}, \mathbf{B})\end{aligned}$$

Q: Is there a better way because the naive way is again very expensive?

$$\Pr(\mathbf{x} \mid \mathbf{A}, \mathbf{B}) = \sum_{\mathbf{y}} \Pr(\mathbf{x} \mid \mathbf{y}, \mathbf{A}, \mathbf{B}) \Pr(\mathbf{y} \mid \mathbf{A}, \mathbf{B})$$

Q: Last but not the least, how can we estimate $\mathbf{A}$ and $\mathbf{B}$?

$$\mathcal{L}(\mathbf{A}, \mathbf{B}; \mathbf{x}, \mathbf{y}) = \Pr(\mathbf{x}, \mathbf{y} \mid \mathbf{A}, \mathbf{B}) = \left[\prod_{t=1}^{T} B_{y_t x_t}\right] \cdot \left[\prod_{t=1}^{T} A_{y_{t-1} y_t}\right]$$

Q: How would a frequentist do it ?

$$\ell(\mathbf{A}, \mathbf{B}; \mathbf{x}, \mathbf{y}) = \ln \mathcal{L} = \sum_{t=1}^{T} \ln A_{y_{t-1} y_t} + \sum_{t=1}^{T} \ln B_{y_t x_t}$$

Q: Of course, the hidden states are never observed, so how can we solve it?

$$\mathbb{E}[\ell(\mathbf{A}, \mathbf{B}; \mathbf{x}, \mathbf{y})] = \sum_{\mathbf{y}} Q(\mathbf{y} \mid \mathbf{x}, \mathbf{A}^*, \mathbf{B}^*) \, \ell(\mathbf{A}, \mathbf{B}; \mathbf{x}, \mathbf{y})$$

where Q is the probability mass function of $\mathbf{y}$, i.e. $\Pr(\mathbf{y} \mid \mathbf{x}, \mathbf{A}^*, \mathbf{B}^*)$.

Q: Does it remind you of something we have seen?