Stochastic matrices

A/Prof Richard Yi Da Xu
Yida.Xu@uts.edu.au
Wechat: aubedata
https://github.com/roboticcam/machine-learning-notes

University of Technology Sydney (UTS)

February 18, 2018

Stochastic matrices

- **Right stochastic matrix** (or row stochastic matrix) is a real square matrix, with **each row** summing to 1.

$$\begin{bmatrix} K_{1\to 1} & \dots & K_{1\to n} \\ \dots & \dots & \dots \\ K_{d\to 1} & \dots & K_{d\to n} \\ \dots & \dots & \dots \\ K_{n\to 1} & \dots & K_{n\to n} \end{bmatrix}$$

- **Left stochastic matrix** (or column stochastic matrix) is a real square matrix, with **each column** summing to 1

$$\begin{bmatrix} K_{1\to 1} & \dots & K_{n\to 1} \\ \dots & \dots & \dots \\ K_{1\to d} & \dots & K_{n\to d} \\ \dots & \dots & \dots \\ K_{1\to n} & \dots & K_{n\to n} \end{bmatrix}$$

- **doubly stochastic matrices**: is a real square matrix, where both **each column** and **each row** summing to 1.

Product of two stochastic matrix is still stochastic

- Each entry in the product AB is a dot product of a row from A and a column from B.

$$(AB)_{ij} = \sum_{k=1}^{n} A_{ik} B_{kj}$$

- We need to prove, for a single row of product $(AB)_{i,:}$

$$\sum_{j=1}^{n} (AB)_{ij} = \sum_{j=1}^{n} \sum_{k=1}^{n} A_{ik} B_{kj} = \sum_{k=1}^{n} (A_{ik} \sum_{j=1}^{n} B_{kj})$$

- Because B is stochastic, $\sum_{j=1}^{n} B_{kj} = 1$
- Because A is stochastic, $\sum_{k=1}^{n} A_{ik} = 1$

Perron-Frobenius Theorem:

If K is a **positive**, left stochastic matrix, then:

- 1 is an eigenvalue of multiplicity one.
- 1 is the largest eigenvalue: all the other eigenvalues have absolute value smaller than 1.
- the eigenvectors corresponding to the eigenvalue 1 have either only positive entries or only negative entries.
- Note that K is a **positive** means, $K_{ij} \geq 0 \;\; \forall i, j$. It's NOT **positive definite matrix**

Power Method Convergence Theorem

- Let K be a positive, left (i.e., column) stochastic $n \times n$ matrix.
- π^* its **probabilistic eigenvector** corresponding to the eigenvalue 1.

$$\begin{bmatrix} K_{1\to 1} & \dots & K_{n\to 1} \\ \dots & \dots & \dots \\ K_{1\to n} & \dots & K_{n\to n} \end{bmatrix} \begin{bmatrix} \pi_1^* \\ \dots \\ \pi_n^* \end{bmatrix} = \begin{bmatrix} \pi_1^* \\ \dots \\ \pi_n^* \end{bmatrix}$$

- Let $\pi^{(1)}$ be the column vector with all entries equal to some arbitrary stochastic vector.
- Then sequence $\{\pi^{(1)}, K\pi^{(1)}, K^2\pi^{(1)}, \dots, K^t\pi^{(1)} \dots, K^\infty\pi^{(1)}\}$ converges to the vector π^*

$$\lim_{t\to\infty} K^t = K^\infty \implies \lim_{t\to\infty} K^t\pi^{(1)} = \pi^*$$

- **Exercise** Generate some random matrix in MATLAB and to show an example of the above.

Extend to continous case

- in the **discrete** case:

$$\begin{bmatrix} K_{1\to 1} & K_{2\to 1} & \dots & K_{n\to 1} \\ \dots & \dots & \dots & \dots \\ K_{1\to d} & K_{2\to d} & \dots & K_{n\to d} \\ \dots & \dots & \dots & \dots \\ K_{1\to n} & K_{2\to n} & \dots & K_{n\to n} \end{bmatrix} \begin{bmatrix} \pi_1^* \\ \dots \\ \pi_d^* \\ \dots \\ \pi_n^* \end{bmatrix} = \begin{bmatrix} \pi_1^* \\ \dots \\ \pi_d^* \\ \dots \\ \pi_n^* \end{bmatrix} \implies \pi_d^* = \sum_{i=1}^{n} \pi_i^* K_{i\to d}$$

- in the **continous** case, let $\pi(x)$ be the target distribution:

$$\pi(x^{(n+1)}) = \int_{x_n} \pi\left(x^{(n)}\right) K\left(x^{(n)} \to x^{(n+1)}\right)$$

- A transition kernel K contains element-wise entries: $\left\{K\left(x^{(n)} \to x^{(n+1)}\right)\right\} \quad \forall x^{(n)}, x^{(n+1)}$
- Sometimes we prefer to write $\left(x^{(n)} \text{ as } x\right)$ and $\left(x^{(n+1)} \text{ as } x^*\right)$.
- $K(x \to x^*)$ is the probability a process at state x moves to state x^* in a **one step**
- $K^n(x \to x^*)$ is the probability a process at state x moves to state x^* in **n steps**

Power Method Convergence in continuous case

- One may have first sample $x^{(1)}$ distributed from an arbitrary distribution:

$$x^{(1)} \sim \pi^{(1)}$$

- by applying K function, to obtain $x^{(2)}$ given $x^{(1)}$ with probability:

$$\pi^{(2)}\left(x^{(2)}\right) = \int_{x^{(1)}} \pi^{(1)}\left(x^{(1)}\right) K\left(x^{(1)} \to x^{(2)}\right) \mathrm{d}x^{(1)}$$

- by applying K function again, to obtain $x^{(3)}$ with probability:

$$\begin{aligned}
\pi^{(3)}\left(x^{(3)}\right) &= \int_{x^{(1)}} \int_{x^{(2)}} \pi^{(1)}\left(x^{(1)}\right) K\left(x^{(1)} \to x^{(2)}\right) K\left(x^{(2)} \to x^{(3)}\right) \mathrm{d}x^{(1)} \mathrm{d}x^{(2)} \\
&= \int_{x^{(1)}} \pi^{(1)}\left(x^{(1)}\right) \underbrace{\int_{x^{(2)}} K\left(x^{(1)} \to x^{(2)}\right) K\left(x^{(2)} \to x^{(3)}\right) \mathrm{d}x^{(2)}} \mathrm{d}x^{(1)} \\
&= \int_{x^{(1)}} \pi^{(1)}\left(x^{(1)}\right) \underbrace{K^2\left(x^{(1)} \to x^{(3)}\right)} \mathrm{d}x^{(1)} \qquad \to \text{converge closer to } \pi\left(x^{(3)}\right)
\end{aligned}$$

- This says,

$$\lim_{t \to \infty} \pi^{(t)}\left(x^{(t)}\right) \to \pi\left(x^{(t)}\right)$$

Burn in samples

- We know,

$$\lim_{t \to \infty} \pi^{(t)}\left(x^{(t)}\right) \to \pi\left(x^{(t)}\right)$$

- But, in practice,

$$\lim_{t \to B} \pi^{(t)}\left(x^{(t)}\right) \to \pi\left(x^{(t)}\right)$$

- $\left\{x^{(1)}, \dots, x^{(B)}\right\}$ are the **burn-in** samples, which we discard.

What is MCMC research is all about

- **equilibrium equation**:

$$\pi(x^*) = \int_x \pi(x) K\left(x \to x^*\right) \mathrm{d}x$$

- In machine learning, we always know the expression of stationary distribution $\pi(x)$,
- Our task is therefore, **find an appropriate** $K\left(x \to x^*\right)$ to generate samples in a Markov fashion.

Detailed Balance

- At equilibrium, that stationary distribution satisfies:

$$\pi(x^*) = \int_x \pi(x) K\left(x \to x^*\right) \mathrm{d}x \qquad \textbf{equilibrium equation}$$

- Proving **equilibrium equation** may be difficult in some cases, therefore, we instead prove detail balance:
- **detailed balance** condition holds when:

$$\pi(x) K\left(x \to x^*\right) = \pi\left(x^*\right) K\left(x^* \to x\right)$$

- **detailed balance** implies **equilibrium equation**:

$$\begin{aligned}\int_x \pi(x) K\left(x \to x^*\right) \mathrm{d}x &= \int_x \pi(x^*) K\left(x^* \to x\right) \mathrm{d}x \\ &= \pi(x^*) \int_x K(x^* \to x) \mathrm{d}x \\ &= \pi(x^*) \qquad \textbf{equilibrium equation}\end{aligned}$$

- the reverse is not always true.

Extend target distribution with auxiliary variables

- At equilibrium, that stationary distribution satisfies:

$$\pi(x^*) = \int_x \pi(x) K\left(x \rightarrow x^*\right) \mathrm{d}x$$

- under many scenarios, we may have an extended joint density (x, u):

$$\pi(x|u)\pi(u) K\left(u, x \rightarrow u^*, x^*\right) = \pi\left(x^*|u^*\right)\pi(u^*) K\left(x^*, u^* \rightarrow x, u\right)$$

- u is auxiliary variables help samping
- one needs to ensure that:

$$\int_u \pi(x, u) \mathrm{d}u = \pi(x)$$

Alternative Use of Stochastic Matrix

- Before dive deep into MCMC algorithms, let's have a look at alternative use of stochastic matrix
- PageRank algorithm is different to MCMC, in PageRank algorithm: K is known
- PageRank algorithm then computes π which is the **invariant distribution**, tells the importance of each web page.

PageRank algorithm

- Imagine we have the following four web pages and their links
- we can then compute the probability of navigating from i^{th} page (discrete state) to j^{th} page (discrete state)

- Page 1 links to pages $\{2, 3\}$

$$\implies K_{1\to 1} = 0, K_{1\to 2} = \frac{1}{2}, K_{1\to 3} = \frac{1}{2}, K_{1\to 4} = 0$$

- Page 2 has links to pages $\{1, 3, 4\}$

$$\implies K_{2\to 1} = \frac{1}{3}, K_{2\to 2} = 0, K_{2\to 3} = \frac{1}{3}, K_{2\to 4} = \frac{1}{3}$$

- Page 3 has links to pages $\{1, 3\}$

$$\implies K_{3\to 1} = \frac{1}{2}, K_{3\to 2} = 0, K_{3\to 3} = \frac{1}{2}, K_{3\to 4} = 0$$

- Page 4 has links to pages $\{2, 3\}$

$$\implies K_{4\to 1} = 0, K_{4\to 2} = \frac{1}{2}, K_{4\to 3} = \frac{1}{2}, K_{4\to 4} = 0$$

Stochastic matrix K

- From the preceding example, **Left stochastic matrix** is:

$$\begin{bmatrix} K_{1\to 1} & K_{2\to 1} & K_{3\to 1} & K_{4\to 1} \\ K_{1\to 2} & K_{2\to 2} & K_{3\to 2} & K_{4\to 2} \\ K_{1\to 3} & K_{2\to 3} & K_{3\to 3} & K_{4\to 3} \\ K_{1\to 4} & K_{2\to 4} & K_{3\to 4} & K_{4\to 4} \end{bmatrix} = \begin{bmatrix} 0 & \frac{1}{3} & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{2} & \frac{1}{2} \\ 0 & \frac{1}{3} & 0 & 0 \end{bmatrix}$$

- From Power Method Convergence Theorem, we know:
  - sequence $\{\pi^{(1)}, K\pi^{(1)}, K^2\pi^{(1)}, \ldots, K^t\pi^{(1)} \ldots, K^\infty\pi^{(1)}\}$ converges to the vector π^*

$$\lim_{t\to\infty} K^t\pi^{(1)} = \pi^*$$

  where π^* is a **probabilistic eigenvector** of K corresponding to the eigenvalue 1.
- **Exercise** What is the usefulness of π^* in the setting of web pages?

Usefulness of π^* in the setting of web pages

The **answer** to usefulness of π^* in the setting of web pages is:

- Shows how **important** each webpage is
- i.e., regardless of the probabilities of the initial webpage visit: $\pi^{(1)}$,
- $\pi^{(1)} \rightarrow \pi^*$, where $\pi^*(i)$ is the target distribution i.e, the probability that the visit will end up at a web page i.
- Note that this is a **reverse problem** of MCMC

Dangling nodes

- What happens when you have the following K:

$$\begin{bmatrix} K_{1\to1} & K_{2\to1} & K_{3\to1} & K_{4\to1} \\ K_{1\to2} & K_{2\to2} & K_{3\to2} & K_{4\to2} \\ K_{1\to3} & K_{2\to3} & K_{3\to3} & K_{4\to3} \\ K_{1\to4} & K_{2\to4} & K_{3\to4} & K_{4\to4} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 1 & 0 \end{bmatrix}$$

- Note that 4^{th} has no out-going node
- **Exercise** check eigenvector correspond to eigenvalue of 1
- What is the eigenvector correspond to eigenvalue of 1, if we change K into:

$$\begin{bmatrix} K_{1\to1} & K_{2\to1} & K_{3\to1} & K_{4\to1} \\ K_{1\to2} & K_{2\to2} & K_{3\to2} & K_{4\to2} \\ K_{1\to3} & K_{2\to3} & K_{3\to3} & K_{4\to3} \\ K_{1\to4} & K_{2\to4} & K_{3\to4} & K_{4\to4} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 1 & \mathbf{1} \end{bmatrix}$$

  Exercise give reason to why this is so?
- **Exercise** How can we solve this?

Dangling nodes: what may be the solution?

$$\begin{bmatrix} K_{1\to 1} & K_{2\to 1} & K_{3\to 1} & K_{4\to 1} \\ K_{1\to 2} & K_{2\to 2} & K_{3\to 2} & K_{4\to 2} \\ K_{1\to 3} & K_{2\to 3} & K_{3\to 3} & K_{4\to 3} \\ K_{1\to 4} & K_{2\to 4} & K_{3\to 4} & K_{4\to 4} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 1 & 0 \end{bmatrix}$$

- One simply solution is:

$$\begin{bmatrix} K_{1\to 1} & K_{2\to 1} & K_{3\to 1} & K_{4\to 1} \\ K_{1\to 2} & K_{2\to 2} & K_{3\to 2} & K_{4\to 2} \\ K_{1\to 3} & K_{2\to 3} & K_{3\to 3} & K_{4\to 3} \\ K_{1\to 4} & K_{2\to 4} & K_{3\to 4} & K_{4\to 4} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & \frac{1}{3} \\ 0 & \frac{1}{2} & 0 & \frac{1}{3} \\ \frac{1}{2} & 0 & 0 & \frac{1}{3} \\ \frac{1}{2} & \frac{1}{2} & 1 & 0 \end{bmatrix}$$

- in words, it means any page doesn't have out-link, we assume it has equal probability of visiting entire web.
- Of course, **data mining** researchers may argue certain web page (having certain properties) may attract higher weights etc.

Disconnected sub-graphs

- What happens when you have the following K:

$$\begin{bmatrix} K_{1\to 1} & K_{2\to 1} & K_{3\to 1} & K_{4\to 1} \\ K_{1\to 2} & K_{2\to 2} & K_{3\to 2} & K_{4\to 2} \\ K_{1\to 3} & K_{2\to 3} & K_{3\to 3} & K_{4\to 3} \\ K_{1\to 4} & K_{2\to 4} & K_{3\to 4} & K_{4\to 4} \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{2} & 1 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{2} \\ 0 & 0 & 1 & \frac{1}{2} \end{bmatrix}$$

- node $\{1,2\}$ and $\{3,4\}$ each form a sub-graph.
- **Exercise** check eigenvector correspond to eigenvalue of 1, also multiplicity of eigenvalue 1
- **Exercise** How can we solve this?

Disconnected sub-graphs: what may be the solution?

$$\begin{bmatrix} K_{1\to1} & K_{2\to1} & K_{3\to1} & K_{4\to1} \\ K_{1\to2} & K_{2\to2} & K_{3\to2} & K_{4\to2} \\ K_{1\to3} & K_{2\to3} & K_{3\to3} & K_{4\to3} \\ K_{1\to4} & K_{2\to4} & K_{3\to4} & K_{4\to4} \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{2} & 1 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{2} \\ 0 & 0 & 1 & \frac{1}{2} \end{bmatrix}$$

- One solution is to use a convex combination between K and a square matrix having identical elements $\frac{1}{n}$:

$$K' = (1-p)K + p\left(\frac{1}{n}\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}\right)$$

- in words, it means most of the time $1-p$, a surfer will follow links to navigate a page
- but with probability p, it will arbitrarily close the current page and go to the new one
- **Exercise** Prove K remains a left stochastic matrix

How to compute the **one hundred billion** dimension eigenvector?

- starting from the vector (not probabilistic eigenvector), x:

$$x = \begin{bmatrix} 1 & 1 & \dots & 1 \end{bmatrix}^\top$$

- generate the sequence: $\{x, Kx, K^2x \dots K^tx\}$ until convergence then its is the eigenvectors of K correspond to eigenvalue of 1, up to a normalisation constant c
- This is solved using **power method**

Power method

- **power method** is used to finding an eigenvector of a square matrix corresponding to the **largest** eigenvalue (in terms of absolute value)
- for stochastic matrix K has eigenvalues:

$$1 = \lambda_1 > |\lambda_2| \geq |\lambda_3| \geq \ldots \geq |\lambda_n|$$

- the initial vector: x as a linear combination of eigenvectors of K:

$$x = c_1 v_1 + c_2 v_2 + \ldots c_n v_n$$

Then,

$$\begin{aligned} Kx &= K(c_1 v_1 + c_2 v_2 + \ldots c_n v_n) \\ &= c_1 \underbrace{\lambda_1}_{=1} v_1 + c_2 \lambda_2 v_2 + \ldots c_n \lambda_n v_n \qquad \text{definition of eigven value/vector} \\ &= c_1 v_1 + c_2 \lambda_2 v_2 + \ldots c_n \lambda_n v_n \\ \implies K^2 x &= c_1 v_1 + c_2 \lambda_2^2 v_2 + \ldots c_n \lambda_n^2 v_n \\ \implies K^t x &= c_1 v_1 + c_t \lambda_2^t v_2 + \ldots c_n \lambda_n^t v_n \end{aligned}$$

$$\lambda_j^k \to 0 \text{ when } j \geq 2 \implies K^t x \to c_1 v_1$$

Few notes

- The second largest eigen value determines the convergence
- **Homework** Perform the following simulations:
  - generate lots of K and choose one which has **large** second eigen values in absolute value
  - also generate a K which has **small** second eigen values in absolute value
  - in both cases, try to compute the sequence $\{x, Kx, K^2x \dots K^tx\}$, using an arbitrary vector x
- **Homework** Generate K from known eigen value/vectors are called Inverse Eigenvalue Problems. Use IEP to generate stochastic matrices above
- Try something like, "Doubly Stochastic Matrices with Prescribed Positive Spectrum"

Reversible Jump MCMC

- the problem setting is:

$$\Pr(\mathbb{M}_k) = \pi_k \qquad \sum_k \pi_k = 1$$

$$\theta|\mathbb{M}_k \sim p_k(\theta) \qquad \mathcal{L}_k(y_1, \dots, y_n|\theta) = \prod_{i=1}^{n} l_k(y_i|\theta)$$

- $p_k(\theta|\mathcal{D}) =$

Reversible Jump MCMC (2)

- the problem setting is:

$$\Pr(\mathbb{M}_k) = \pi_k \qquad \sum_k \pi_k = 1$$

$$\theta|\mathbb{M}_k \sim p_k(\theta) \qquad \mathcal{L}_k(y_1, \dots, y_n|\theta) = \prod_{i=1}^{n} l_k(y_i|\theta)$$

- $p_k(\theta|\mathcal{D}) =$