Lecture 23: Sequential data and Markov models

Reading: Section 14.10

GU4241/GR5241 Statistical Machine Learning

Linxi Liu
April 13, 2017

Motivation: PageRank

Simple random walk

Start with a graph G. Define a random sequence of vertices as follows:

- Choose a vertex X_1 uniformly at random.
- Choose a vertex X_2 uniformly at random from the neighbors of X_1. Move to X_2.
- Iterate: At step n, uniformly sample a neighbor X_n of X_{n-1}, and move to X_n.

This is called *simple random walk* on G.

Motivation: PageRank

Google's PageRank Algorithm

To sort the web pages matching a search query by importance, PageRank:

1. Defines a graph G whose vertices are web pages and whose edges are web links.
2. Computes the probability distribution on vertices x in G given by

   $P_n(x) = \mathsf{Pr}\{X_n = x\}$, where $X_1, \ldots X_n$ is a simple random walk on G

   and n is very large.

We will try to understand (a) why and (b) how P_n can be computed.

Sequential Data

So far: I.i.d. sequences

We have assumed that samples are of the form

$$X_1 = x_1, X_2 = x_2, \ldots \qquad \text{where} \qquad X_1, X_2, \ldots \overset{\text{i.i.d.}}{\sim} P$$

for some distribution P. In particular, the order of observations does not matter.

Now: Dependence on the past

We now consider sequences in which the value X_n can be stochastically dependent on $X_1, \ldots, X_{n-1}$, so we have to consider conditional probabilities of the form

$$P(X_n | X_1, \ldots, X_{n-1}) \; .$$

Sequential Data

Application examples

- Speech and handwriting recognition.
- Time series, e.g. in finance. (These often assume a *continuous* index. Our index n is discrete.)
- Simulation and estimation algorithms (Markov chain Monte Carlo).
- Random walk models (e.g. web search).

Markov Models

Markov models

The sequence $(X_n)_n$ is called a **Markov chain of order** r if X_n depends only on a fixed number r of previous samples, i.e. if

$$P(X_n|X_{n-1},\ldots,X_1) = P(X_n|X_{n-1},\ldots,X_{n-r}) \ .$$

If we simply call $(X_n)_n$ a **Markov chain**, we imply $r = 1$.

Initial state

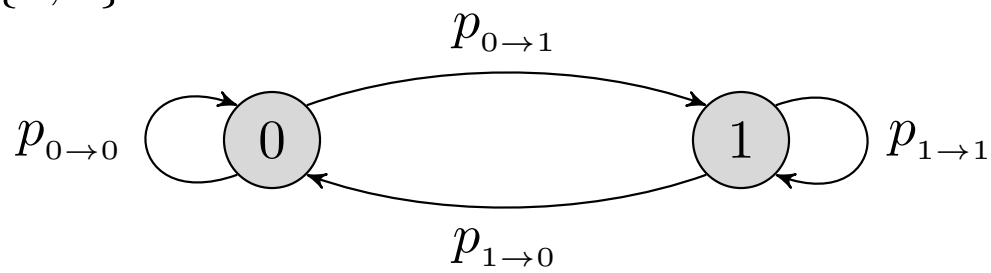
The first state in the sequence is special because it does not have a "past", and is usually denoted X_0.

Example: $r = 2$

$$\underbrace{X_0 = x_0,\ X_1 = x_1}_{\substack{X_4 \text{ is independent} \\ \text{of these } \textit{given } X_2, X_3}},\ \underbrace{X_2 = x_2,\ X_3 = x_3}_{X_4 \text{ may depend on these}},\ X_4 = ?$$

Graphical Representation

Suppose $\mathbf{X} = \{0, 1\}$.



- We regard 0 and 1 as possible "states" of X, represented as vertices.
- Each pair $X_{n-1} = s, X_n = t$ in the sequence is regarded as a "transition" from s to t and represented as an edge in the graph.
- Each edge $s \to t$ is weighted by the probability

$$p_{s\to t} := \mathsf{Pr}\{X_n = t | X_{n-1} = s\} \ .$$

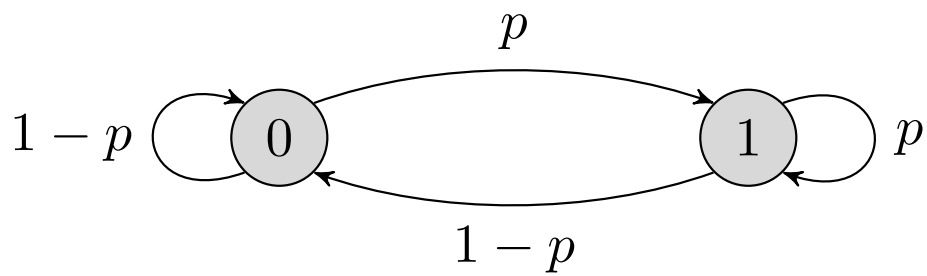
State space

The elements of the sample space $\mathbf{X}$ are called the **states** of the chain. $\mathbf{X}$ is often called the **state space**. We generally assume that $\mathbf{X}$ is finite, but Markov chains can be generalized to infinite and even uncountable state spaces.

Graphical Representation

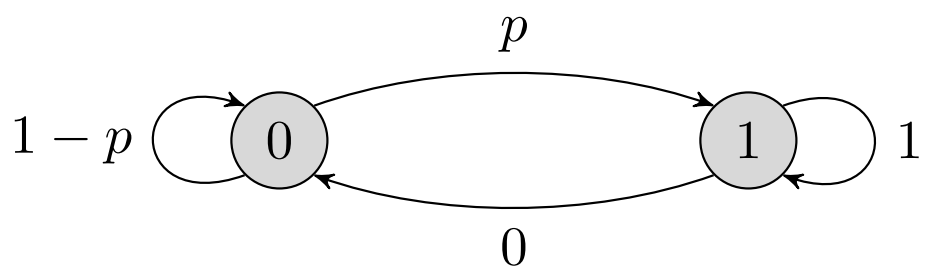
First example: Independent coin flips

Suppose X is a biased coin with $\Pr\{X_n = 1\} = p$ independently of X_{n-1}. I.e., the sequence (X_n) is iid Bernoulli with parameter p.



Breaking independence

Here is a simple modification to the chain above; only $p_{1\to 0}$ and $p_{1\to 1}$ have changed:



This is still a valid Markov chain, but the elements of the sequence are no longer independent.

Graphical Representation

Observation

The graph representation is only possible if $p_{s\to t}$ is independent of n. Otherwise we would have to draw a different graph for each n.

If $p_{s\to t}$ does not depend on n, the Markov chain is called **stationary**.

Transition matrix

The probabilities $p_{s\to t}$ are called the **transition probabilities** of the Markov chain. If $|\mathbf{X}| = d$, the $d \times d$-matrix

$$\mathbf{p} := (p_{i\to j})_{j,i\leq d} = \begin{pmatrix} p_{1\to 1} & \cdots & p_{d\to 1} \\ \vdots & & \vdots \\ p_{1\to d} & \cdots & p_{d\to d} \end{pmatrix}$$

is called the **transition matrix** of the chain. This is precisely the adjacency matrix of the graph representing the chain. Each column is a probability distribution on d events.

Graphical Representation

Complete description of a Markov chain

The transition matrix does not completely determine the chain: It determines the probability of a state given a previous state, but not the probability of the starting state. We have to additionally specify the distribution of the first state.

Inital distribution

The distribution of the first state, i.e. the vector

$$P_{\mathsf{init}} := (\mathsf{Pr}\{X_0 = 1\}, \ldots, \mathsf{Pr}\{X_0 = d\}) \ ,$$

is called the **initial distribution** of the Markov chain.

Representing stationary Markov chains

Any stationary Markov chain with finite state space can be completely described by a transition matrix $\mathbf{p}$ and an initial distribution P_{init}. That is, the pair $(\mathbf{p}, P_{\mathsf{init}})$ completely determines the joint distribution of the sequence $(X_0, X_1, \ldots)$.

Random walks on graphs

Simple random walk

Suppose we are given a directed graph G (with unweighted edges). We had already mentioned that the **simple random walk** on G is the vertex-valued random sequence $X_0, X_1, \ldots$ defined as:

- We select a vertex X_0 in G uniformly at random.
- For $n = 1, 2, \ldots$, select X_n uniformly at random from the children of X_{n-1} in the graph.

Markov chain representation

Clearly, the simple random walk on a graph with d vertices is a Markov chain with

$$P_{\text{init}} = \left(\frac{1}{d}, \ldots, \frac{1}{d}\right) \qquad \text{and} \qquad p_{i \to j} = \frac{1}{\#\text{ edges out of } i}$$

Random Walks and Markov Chains

Generalizing simple random walk

We can generalize the idea of simple random walk by substituting the uniform distributions by other distributions. To this end, we can weight each edge in the graph by a probability of following that edge.

Adjacency matrix

If the edge weights are proper probabilities, each column of the adjacency matrix must sum to one. In other words, the matrix is the transition matrix of a Markov chain.

Random walks and Markov chains

If we also choose a general distribution for the initial state of the random walk, we obtain a completely determined Markov chain. Hence:

Any Markov chain on a finite state space is a random walk on a weighted graph and vice versa.

Internet Search

Queries

The first step in internet search is query matching:

1. The user enters a search query (a string of words).
2. The search engine determines all web pages indexed in its database which match the query.

This is typically a large set. For example, Google reports ca 83 million matches for the query "random walk".

The ranking problem

- For the search result to be useful, the most useful link should with high probability be among the first few matches shown to the user.
- That requires the matching results to be *ranked*, i.e. sorted in order of decreasing "usefulness".

Popularity Scoring

Available data

Using a web crawler, we can (approximately) determine the link structure of the internet. That is, we can determine:

- Which pages there are.
- Which page links which.

A web crawler cannot determine:

- How often a link is followed.
- How often a page is visited.

Web graph

The link structure can be represented as a graph with

$$\text{vertices} = \text{web pages} \qquad \text{and} \qquad \text{edges} = \text{links}.$$

Random Walk Network Models

Key idea

The popularity of a page x is proportional to the probability that a "random web surfer" ends up on page x after a n steps.

Probabilistic model

The path of the surfer is modeled by a random walk on the web graph.

Modeling assumptions

Two assumptions are implicit in this model:

1. Better pages are linked more often.
2. A link from a high-quality page is worth more than one from a low-quality page.

Remarks

- We will find later that the choice of n does not matter.
- To compute the popularity score, we first have to understand Markov chains a bit better.

State Probabilities

we can make a change:
P2 = P*P1=P^2*Pinit
if we want to know the distribution of the xn,
then n step of transition
then
Pn = p^n*Pinit
one is the initial distribution and the other is trasition matrix,then we can know the Markocv chain process

Probability after $n = 1$ steps

If we *know* the initial state, then

$$\Pr\{X_1 = s_1 \,|\, X_0 = s_0\} = p_{s_0 \to s_1}$$

P_1 describes the probability of X_1 if we do *not* know the starting state (i.e. the probability before we start the chain):

$$P_1(s_1) = \Pr\{X_1 = s_1\} = \sum_{s_0 \in \mathbf{X}} \Pr\{X_1 = s_1 \,|\, X_0 = s_0\} P_{\text{init}}(s_0)$$

$$= \sum_{s_0 \in \mathbf{X}} p_{s_0 \to s_1} P_{\text{init}}(s_0) \ .$$

p(x1 = 1) = Σp(x1 = s1, x0 = s0)
= Σp(x1=s1|x0 = s0)*Pinit(s0)
p1 = p*Pinit
this is the distribution of the random walk after 1 click

Pn will be the distribution after n clicks

Matrix representation

Recall that $\mathbf{p}$ is a $d \times d$-matrix and P_{init} a vector of length d. The equation for P_1 above is a matrix-vector product, so

$$P_1 = \mathbf{p} \cdot P_{\text{init}} \ .$$

State Probabilities

Probability after $n = 2$ steps

The same argument shows that P_2 is given by

$$P_2(s_2) = \sum_{s_1 \in \mathbf{X}} p_{s_1 \to s_2} P_1(s_1) \ ,$$

hence

$$P_2 = \mathbf{p} \cdot P_1 = \mathbf{p} \cdot \mathbf{p} \cdot P_{\text{init}} \ .$$

For arbitary n

$$P_n = \mathbf{p}^n P_{\text{init}}$$

Limits and Equilibria

Limiting distribution

Instead of considering P_n for a specific, large n, we take the limit

$$P_\infty := \lim_{n\to\infty} P_n = \lim_{n\to\infty} \mathbf{p}^n P_{\text{init}} \ ,$$

provided that the limit exists.

Observation

if xn is the limit distriburion, then xn+1 is still limit distribution ?????

If the limit P_∞ exists, then

$$\mathbf{p} \cdot P_\infty = \mathbf{p} \cdot \lim_{n\to\infty} \mathbf{p}^n P_{\text{init}} = \lim_{n\to\infty} \mathbf{p}^n P_{\text{init}} = P_\infty \ ,$$

which motivates the next definition.

Equilibrium distribution

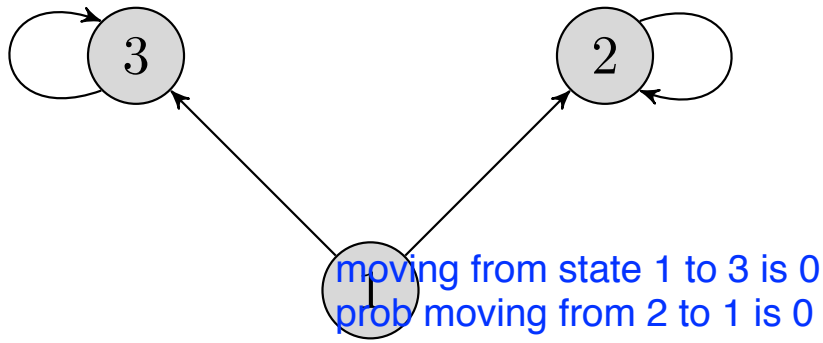
If $\mathbf{p}$ is the tansition matrix of a Markov chain, a distribution P on $\mathbf{X}$ which is invariant under $\mathbf{p}$ in the sense that

$$\mathbf{p} \cdot P = P$$

is called an **equilibrium distribution** or **invariant distribution** of the Markov chain.

What Can Go Wrong?

Problem 1: The equilibrium distribution may not be unique



For this chain, both $P = (0, 1, 0)$ and $P' = (0, 0, 1)$ are valid equilibria. Which one emerges depends on the initial state and (if we start in state 1) on the first transition.

Remedy

Require that there is a path in the graph (with non-zero probability) from each state to every other state. A Markov chain satisfying this condition is called **irreducible**. positive prob when moving from one stage to another

What Can Go Wrong?

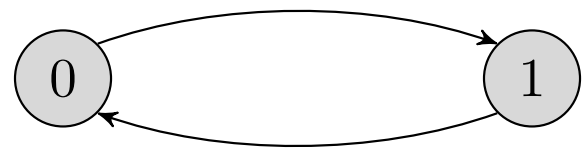
Recall that a sequence in $\mathbb{R}$ does not have a limit if it "oscillates". For example,

$$\lim_n 1^n = 1 \qquad \text{but} \qquad \lim_n (-1)^n \text{ does not exist}$$

Problem 2: The limit may not exist

there are only two statge, the limited distribution will confined in these two statge;
markov chain will between 0 1 0 1
periodic 0 1 0 1
i1. we

- The chain on the right has no limit distribution.
- If we start e.g. in state 0, then:
  - 0 can only be reached in even steps.
  - 1 only in odd steps.
- The distribution P_n oscillates between

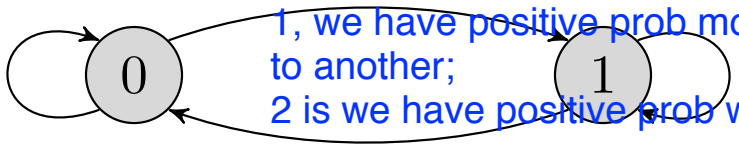


$$P_{\text{even}} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \qquad \text{and} \qquad P_{\text{odd}} = \begin{pmatrix} 0 \\ 1 \end{pmatrix} .$$

What Can Go Wrong?

Remedy

To prevent this (particular) problem, we can add two edges:



1, we have positive prob moving from one state to another;
2 is we have positive prob when we stay the same place

if a markov chain satisfy irreuceble and aperiodic then we can find invariance distribution to take positive liimtation and other 3 qualities

Now each state is reachable in every step.

The problem (at least this example) is that we have to leave the state before we can return to it. We prevent this, we introduce the following definition.

Aperiodic chains

We call a stationary Markov chain **aperiodic** if, for every state s,

$$\Pr\{X_n = s \,|\, X_{n-1} = s\} = p_{s \to s} > 0 \,.$$

In short, a stationary chain is aperiodic if the transition matrix has non-zero diagonal.

Equilibrium Distributions

We have introduced two definitions which prevent two rather obvious problems.

Surprisingly, these definitions are all we need to guarantee limits.

Theorem

Suppose a Markov chain $(\mathbf{p}, P_{\text{init}})$ is stationary, and for each state $s \in \mathbf{X}$:

1. There is a path (with non-zero probability) from s to every other state (i.e. the chain is irreducible).
2. $p_{s \to s} > 0$ (i.e. the chain is aperiodic).

Then:

- The limit distribution P_∞ exists. invariant distribution?
- The limit distribution is also the equlibrium distribution.
- The equilibrium distribution is unique.

Computing the Equilibrium

Power method

If the the transition matrix $\mathbf{p}$ makes the chain irreducible and aperiodic, we know that

$$\text{equilibrium distribution} = \text{limit distribution} \quad .$$

This means we can compute the approximate the equilibrium P_∞ by P_n. In other words, we start with any distribution P_{init} (e.g. uniform) and repeatedly multiply by $\mathbf{p}$:

$$P_{n+1} = \mathbf{p} \cdot P_n$$

We can threshold the change between steps, e.g. by checking $\|P_{n+1} - P_n\| < \tau$ for some small τ.

Computing the Equilibrium

Remark: Eigenstructure

The power method can be regarded as an eigenvector computation. The definition

$$P = \mathbf{p} \cdot P$$

of the equilibrium means that $P = P_\infty$ is an eigenvector of $\mathbf{p}$ with eigenvalue 1. If $\mathbf{p}$ is irreducible and aperiodic, it can be shown that 1 is the largest eigenvalue.

PageRank

this markov chain is good;
we can use power method to find limit distribution

Constructing the transition matrix

We start with the web graph and construct the transition matrix of simple random walk, i.e.

$$a_{ji} := \begin{cases} \frac{1}{\#\text{ edges out of } i} & \text{if } i \text{ links to } j \\ 0 & \text{otherwise} \end{cases}$$

A chain defined by $A := (a_{ij})$ will almost certainly not be irreducible (think of web pages which do not link anywhere). We therefore regularize A by defining

$$\mathbf{p} := (1-\alpha)A + \frac{\alpha}{d}\begin{pmatrix} 1 & \cdots & 1 \\ \vdots & & \vdots \\ 1 & \cdots & 1 \end{pmatrix}$$

for some small $\alpha > 0$.
Clearly, this makes $\mathbf{p}$ both irreducible and aperiodic.

PageRank

Computing the equilibrium

Given $\mathbf{p}$, the equilibrium distribution is computed using the power method. Since the web changes, the power method can be re-run every few days with the previous equilibrium as initial distribution.

we can apply the page rank algotithm
and return the high quality links

The Random Surfer Again

We can now take a more informed look at the idea of a random web surfer:

- Suppose the surfer is more likely to start on a popular page than on an unpopular one.
- In terms of the popularity model, this means

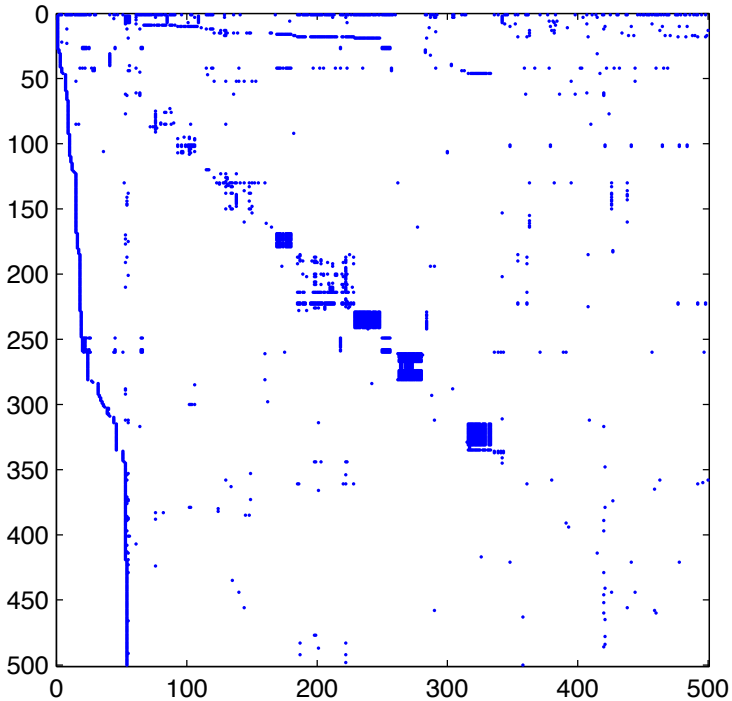
$$X_0 \sim P_{\mathsf{equ}} \ ,$$

  where P_{equ} is the equilibrium distribution of the chain.
- After following any number of links n (with probabilities given by the transition matrix $\mathbf{p}$),

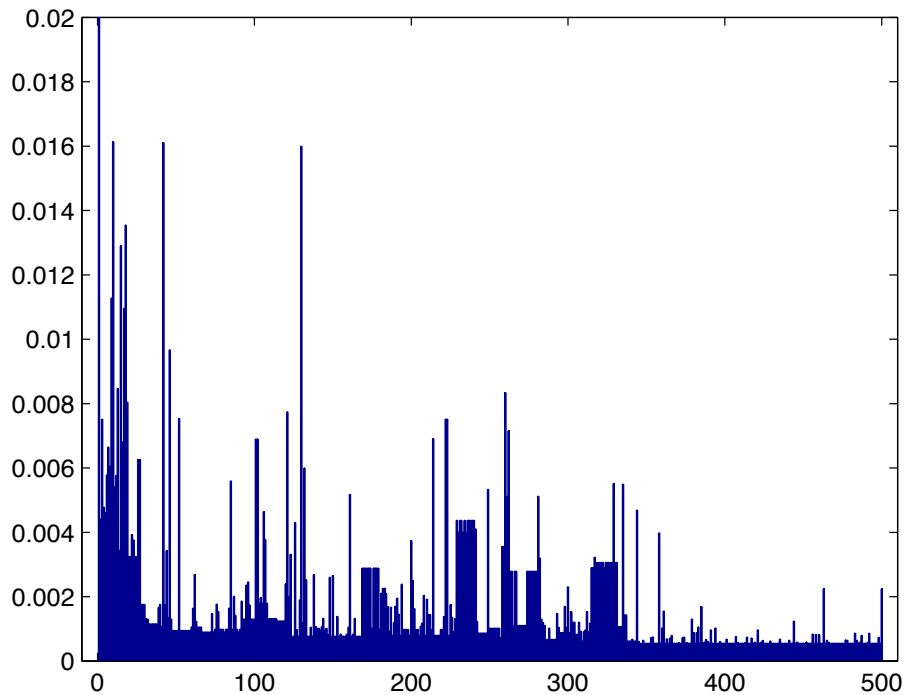
$$P_n = \mathbf{p}^n P_{\mathsf{equ}} = P_{\mathsf{equ}} \ .$$

- In this sense, P_{equ} is really the consistent solution to our problem, even if we *compute* it by starting the random walk from e.g. a uniform initial distribution instead.
- In particular, it does not matter how we choose n in the model.

Example



Adjacence matrix of the web graph of 500 web pages. The root (index 0) is www.harvard.edu.



Equilibrium distribution computed by PageRank.

See K. Murphy, "Machine Learning", MIT Press 2012.