CS 112: Computer System Modeling Fundamentals

Prof. Jenn Wortman Vaughan

May 31, 2011

Lecture 18

Reminders & Announcements

- Homework 5 is due in section on Friday, which will be a review session for the final exam
- Check the website and catch up on your reading now!
- The best way to prepare for the final is to practice working through the problems in the book

Markov Chains

A Markov chain is specified by:

- A set of states $S = \{1, 2, \ldots, m\}$
- A set of transition probabilities $p_{i,j}$ where

$$p_{i,j} = P(X_{t+1} = j \mid X_t = i)$$

The key independence assumption is the Markov property:

$$P(X_{t+1} = j \mid X_t = i, X_{t-1} = x_{t-1}, X_{t-2} = x_{t-2}, \ldots, X_0 = x_0)$$
$$= P(X_{t+1} = j \mid X_t = i)$$

Classification of States

Accessibility: State j is accessible from state i if for some n, the n-step transition probability from i to j is positive.

Recurrence: State i is recurrent if for every state j that is accessible from i, i is also accessible from j.

If i is not recurrent, then it is transient.

The set of all states accessible from a recurrent state form a recurrent class. Note that all of the states in a recurrent class are accessible from each other.

Periodicity

- A recurrent class is periodic if it can be broken into $d > 1$ disjoint subsets $S_1, S_2, \ldots, S_d$ in such a way that
  - All transitions from states in S_i lead to states in S_{i+1} for $i \in \{1, \ldots, d-1\}$
  - All transitions from states in S_d lead to states in S_1

- The period of the class is the number d of subsets

n-Step Transitions

- We can efficiently compute the n-step transition probability, $P(X_n = j \mid X_0 = i)$, using the recursive formula

$$P(X_n = j \mid X_0 = i) = \sum_{k=1}^{m} p_{k,j} P(X_{n-1} = k \mid X_0 = i)$$

Today...

- Long-term behavior of Markov chains
  - Steady state probabilities & the balance equations

  (Break for course evaluations)

- Application: Google's PageRank algorithm

Back to the Faulty Router

My faulty router can be either online or offline. If it is online one day, it will be online the next day with probability 0.8. If it is offline one day, it will remain offline the next day with probability 0.4.

- What fraction of the time will my router be online in the long run?

Convergence of Markov Chains

- Under what circumstances does $P(X_n = i)$ converge to a unique value for each i as n grows large?

Convergence of Markov Chains

- Under what circumstances does $P(X_n = i)$ converge to a unique value for each i as n grows large?

- Under what circumstances might this *not* happen?

Convergence of Markov Chains

- Under what circumstances does $P(X_n = i)$ converge to a unique value for each i as n grows large?

- Under what circumstances might this *not* happen?
    - Multiple recurrent classes

Convergence of Markov Chains

- Under what circumstances does $P(X_n = i)$ converge to a unique value for each i as n grows large?

- Under what circumstances might this *not* happen?
  - Multiple recurrent classes
  - Periodic recurrent class

Convergence of Markov Chains

- Under what circumstances does $P(X_n = i)$ converge to a unique value for each i as n grows large?

- Under what circumstances might this *not* happen?
  - Multiple recurrent classes
  - Periodic recurrent class

- If these probabilities *do* converge, what do we know about the values they converge to?

Steady-State Convergence Theorem

Theorem: Consider any Markov chain with a single recurrent class, which is not periodic. Then the steady state probabilities of the chain are the unique values π_1, ..., π_m that satisfy the system of equations:

$$\pi_j = \sum_{k=1}^{m} \pi_k p_{k,j} \qquad \text{for } j = 1,\ldots,m$$

$$\sum_{k=1}^{m} \pi_k = 1$$

These are referred to as the balance equations.

Steady-State Convergence Theorem

Theorem: Consider any Markov chain with a single recurrent class, which is not periodic. Then the steady state probabilities of the chain are the unique values $\pi_1, \ldots, \pi_m$ that satisfy the system of equations:

$$\pi_j = \sum_{k=1}^{m} \pi_k p_{k,j} \qquad \text{for } j = 1, \ldots, m$$

$$\sum_{k=1}^{m} \pi_k = 1$$

These are referred to as the balance equations.

When these conditions hold, X_0 doesn't matter.

Back to the Faulty Router

My faulty router can be either online or offline. If it is online one day, it will be online the next day with probability 0.8. If it is offline one day, it will remain offline the next day with probability 0.4.

- What fraction of the time will my router be online in the long run?

Back to the Faulty Router

My faulty router can be either online or offline. If it is online one day, it will be online the next day with probability 0.8. If it is offline one day, it will remain offline the next day with probability 0.4.

- What fraction of the time will my router be online in the long run?
- Suppose that if the router remains offline for 3 straight days, I (temporarily) repair it, resetting it to the online state. What fraction of the time will it be online?

Google's PageRank

- Google determines which search results to return based on a mix of relevance and quality ("rank")

- How should the rank of a webpage be determined?

Google's PageRank

- Google determines which search results to return based on a mix of relevance and quality ("rank")

- How should the rank of a webpage be determined?
  - High quality webpages link to other high quality webpages. The rank of a webpage should be a function of the rank of pages that link to it.

Google's PageRank

- Google determines which search results to return based on a mix of relevance and quality ("rank")

- How should the rank of a webpage be determined?
  - High quality webpages link to other high quality webpages. The rank of a webpage should be a function of the rank of pages that link to it.
  - If a webpage links to n other pages, each should inherit a $1/n$ share of its rank.

Google's PageRank

- Let S_i be the set of pages that link to page *i*, and let n_j>0 be the number of pages that *j* links to. Then we want

$$R_i = \sum_{j \in S_i} R_j \frac{1}{n_j} \quad \text{for all pages } i$$

Google's PageRank

- Let S_i be the set of pages that link to page *i*, and let $n_j > 0$ be the number of pages that *j* links to. Then we want

$$R_i = \sum_{j \in S_i} R_j \frac{1}{n_j} \quad \text{for all pages } i$$

- These equations can be interpreted as the balance equations of a random surfer Markov chain!

Google's PageRank

- Let S_i be the set of pages that link to page *i*, and let n_j>0 be the number of pages that *j* links to. Then we want

$$R_i = \sum_{j \in S_i} R_j \frac{1}{n_j} \quad \text{for all pages } i$$

- These equations can be interpreted as the balance equations of a random surfer Markov chain!

- Unfortunately, there might not be a unique solution...

Google's PageRank

- We can get around this problem by making the random surfer a little more random...

Google's PageRank

- We can get around this problem by making the random surfer a little more random...
  - At each time step, with probability α, a random link on the current page is followed (all equally likely)

Google's PageRank

- We can get around this problem by making the random surfer a little more random...
  - At each time step, with probability α, a random link on the current page is followed (all equally likely)
  - With probability 1-α, a new page is chosen uniformly at random from *all n* webpages

Google's PageRank

- We can get around this problem by making the random surfer a little more random...
  - At each time step, with probability α, a random link on the current page is followed (all equally likely)
  - With probability 1-α, a new page is chosen uniformly at random from *all n* webpages

$$R_i = \alpha \sum_{j \in S_i} R_j \frac{1}{n_j} + (1-\alpha)\frac{1}{n} \quad \text{for all pages } i$$

Google's PageRank

- We can get around this problem by making the random surfer a little more random...
  - At each time step, with probability α, a random link on the current page is followed (all equally likely)
  - With probability 1-α, a new page is chosen uniformly at random from *all n* webpages

$$R_i = \alpha \sum_{j \in S_i} R_j \frac{1}{n_j} + (1-\alpha)\frac{1}{n} \quad \text{for all pages } i$$

- This new MC has one recurrent class, and it is not periodic.

Other Applications

Similar ideas have been used in a variety of applications:

- Measuring the impact of bloggers in the blogosphere
- Measuring the impact of scientific journals based on citations
- Measuring how trustworthy buyers and sellers are on eBay or other e-commerce sites
- Determining the importance of species in the food chain