Recitation - Markov Chains

1) Used as a modeling tool
2) Used as a tool for inference (MCMC)

Modeling Tool (Sequential Data)

We can write any joint distribution btw. a set of variables as: (example will be for 4 R.V, but can apply to arbitrary no.)

$$p(x_1, x_2, x_3, x_4) = p(x_1 | x_2, x_3, x_4)\, p(x_2 | x_3, x_4)\, p(x_3 | x_4)\, p(x_4) \ldots$$

$$\text{or } = p(x_1)\, p(x_2 | x_1)\, p(x_3 | x_2, x_1)\, p(x_4 | x_3, x_2, x_1)$$

Making a Markov assumption is a simplification of the joint distribution: A first-order Markov assumption means:

$$p(x_1, x_2, x_3, x_4) = p(x_4 | x_3)\, p(x_3 | x_2)\, p(x_2 | x_1)\, p(x_1)$$

depends only the previous variable in the sequence.

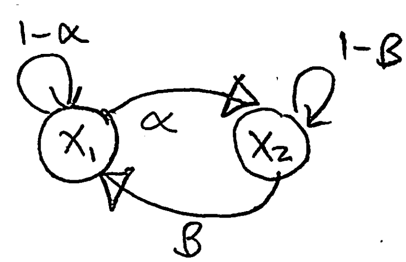
A second order Markov assumption:

$$p(x_1, x_2, x_3, x_4) = p(x_4 | x_3, x_2)\, p(x_3 | x_2, x_1)\, p(x_2 | x_1)\, p(x_1)$$

Assume now that $X_1, X_2, \ldots X_N$ is discrete s.t $X_n \in \{1, \ldots, K\}$ then we can write our conditional dist. as a $K \times K$ matrix.

Transition Matrix for $p(X_2 | X_1)$ (completely defines the Markov behavior for a first-order assumption)

$$\begin{array}{cc} & X_2 \\ A = X_1 & \underbrace{\begin{bmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{bmatrix}}_{\text{rows sum to } 1} \end{array}$$



Examples where Markov models are used:

N-gram language models

1) State space (i.e $X_n \in \{1 \dots K\}$) is represented as all the words in a dictionary.

2) If using a first-order assumption, then the model is called a bigram model: $p(X_3 = \text{cat} \mid X_2 = \text{furry})$

Second-order: trigram model

zero-order: unigram model (bag of words)

Purposes

1) sentence completion

2) data compression (assign short codewords to more probable strings)

3) text classification (spam vs. real email) what would a unigram model be equivalent to? naive bayes

4) generating artificial text

---

2) Interpretation as a stochastic dynamical system. (mcmc)

Define A to be our one-step transition matrix

Let $A_{ij} = p(X_t = j \mid X_{t-1} = i)$ and $\pi_t(j) = p(X_t = j)$

conditional probability — marginal prob. at time point t. for state j.

Let π_0 be the initial state.

Then:

$$\pi_1(j) = \sum_i \pi_0(i) A_{ij} \quad \text{or} \quad \pi_1 = \pi_0 A$$

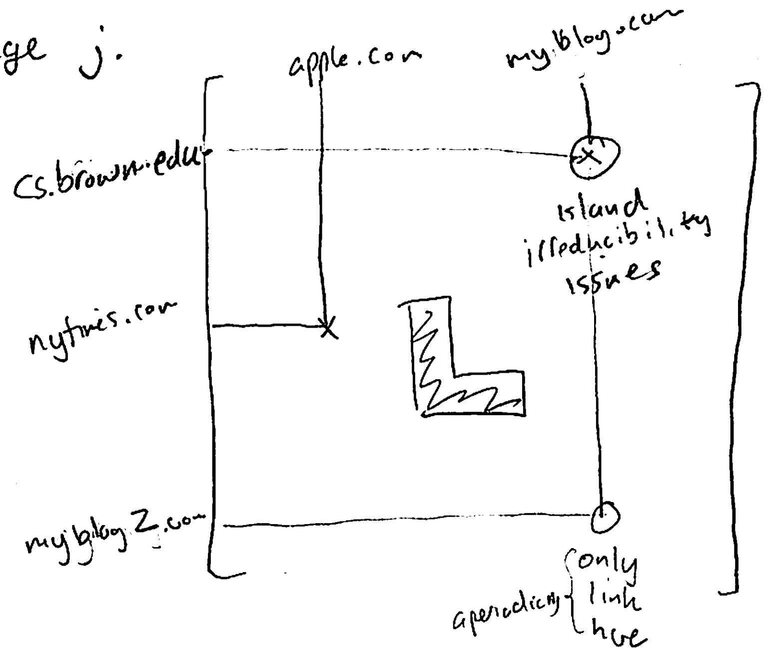
After many, many iterations, we may come to a point where $\pi_N = \pi_N A$. If this ever happens then π has reached what we call as a stationary distribution.

Thought experiment: Searching for data on the WWW.

Pretend that nodes are webpages and links between them representing some probability of going to another page (e.g. normalized #links, page with more incoming links have higher pr

We can specify a transition matrix as: $L + E$

where L_{ij} is the probability of going from some webpage i to webpage j.



$+ E$

uniform noise

allows the new transition matrix to be no longer aperiodic and irreducible

same dimension as L

Finding the unique distribution also becomes an eigenvector problem....

$$\underbrace{p(x_{i+1})\,[L+E]}_{\text{Left eigenvector representation}} = p(x)$$

where we assume that $p(x_{i+1})$ and $p(x)$ is the stationary distribution.

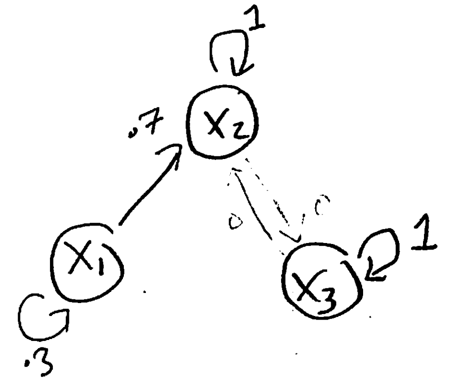
What's this technique used for? $\Rightarrow$ Google page rank!!

Uses the stationary dist as a way of ranking pages

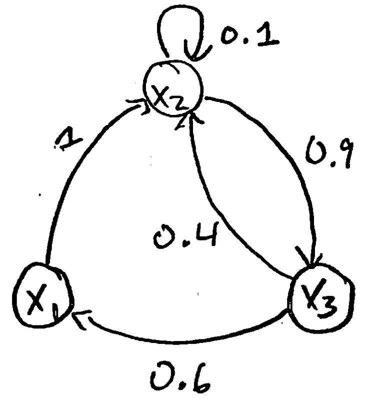
The one they're running is probably a lot lot more fancier, but this is their billion dollar idea.

* Show MATLAB example of stationary distribution *
(unique and not unique)

This idea of a stationary distribution existing is a fundamental cornerstone to MCMC technique. (The defacto workhorse for stochastic learning / inference)



Transition 1
seen in Matlab
code (Invalid mcmc)



Transition 2
Valid mcmc

For a unique stationary dist to exist, it must obey these 2 properties: Irreducibility
Aperiodicity

Irreducibility

For any state of the Markov chain, there is a positive probability of visiting all other states. (transition graph is connected)

Aperiodicity

The graph should not get trapped in cycles.

A sufficient condition to show that $p(x)$ is the desired invariant dist is the following reversibility (detailed balance) condition:

transition matrix

$$p(x_i)\,T(x_{i-1} \mid x_i) = p(x_{i-1})\,T(x_i \mid x_{i-1})$$

Summing over x_{i-1}, we get the marginal $p(x_i)$

$$\sum_{x_{i-1}} p(x_i) = \sum_{x_{i-1}} p(x_{i-1})\,T(x_i \mid x_{i-1})$$