POLITECNICO DI MILANO

Markov Chains

Information Retrieval and Data Mining

Prof. Matteo Matteucci

Probabilistic Reasoning for Time Series

To describe an ever changing world we can use a series of random variables describing the world state at any time instant!

$$X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4 \rightarrow \ldots$$

- It represents a sequence of states: $X_1, X_2, X_3, \ldots$
- The transition from X_{t-1} to X_t depends only on X_{t-1}

$$P(X_t|X_{t-1}, X_{t-2}, \ldots, X_1, X_0) = P(X_t|X_{t-1}) \qquad \text{(Markov Property)}$$

- When transition probabilities are the same a any *t*, we are facing a stationary process.
- A *Bayesian Network* that forms a chain!

(Let's skip basic stuff and go to hidden models)

Stochastic Processes and Markov Chains

- Given X_t the value of a system characteristic at time t described as a (state) random variable, we have:
  - *Discrete Stochastic Process:* describes the a relationship between the stochastic description of a system (X_0, X_1, X_2, …) at some discrete time steps.
  - A *Continuous Stochastic Process* is a stochastic process where the state can be observed at any time.

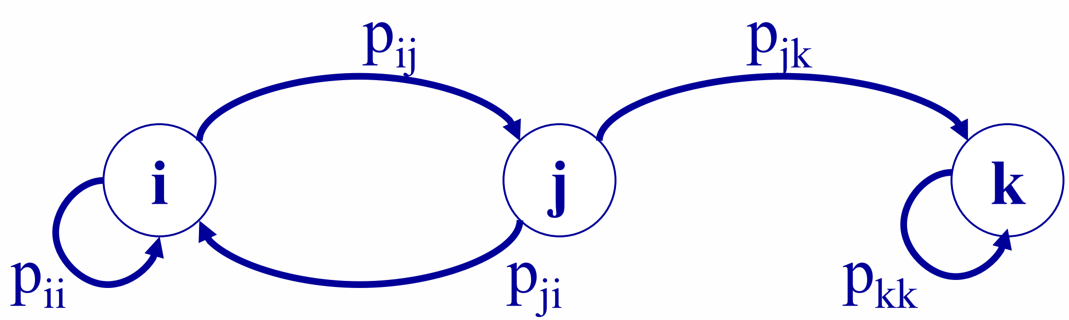
- A Discrete Stochastic Process is a (first order) ***Markov Chain*** when we have thet $\forall t = 1, 2, 3, \ldots$ and for all n states it holds:
  - $P(X_{t+1}=i_{t+1}|X_t=i_t, X_{t-1}=i_{t-1},\ldots,X_1=i_1,X_0=i_0)=P(X_{t+1}=i_{t+1}|X_t=i_t)$
- Whenever the probability of an event is independent from time the Markov Chain is *Stationary:* $P(X_{t+1}=j|X_t=i)=p_{ij}$

Markov Chain Description

- A Markov Chain can be described using a *Transition Matrix* where p_{ij} describes the probability of getting into state *j* starting from state *i*:

$$P = \begin{pmatrix} p_{11} p_{12} p_{13} \cdots p_{1n} \\ p_{21} p_{22} p_{23} \cdots p_{2n} \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ p_{n1} p_{n2} p_{n3} \cdots p_{nn} \end{pmatrix} \qquad \sum_{j=1}^{n} p_{ij} = 1$$

- This transition matrix can be described also using a directed graph



Computing Probabilities

- Given a Markov Chain in state *i* at time *m* we can compute states probability after *n* time steps:

$$P(X_{m+n}=j|X_m=i)=P(X_n=j|X_0=i)=P_{ij}(n)$$

- If we take n=2 we have

$P_{ij}(2) = \sum_k p_{ik} \cdot p_{kj}$ Scalar product of row *i* and column *j*

- In general $P_{ij}(n)$ = *ij-th* element of P^{n}.

- The probability of being in a given state *j* at time *n* without knowing the exact state of Markov Chain at time *0* is thus:

$$\sum_i q_i \cdot P_{ij}(n) = q \cdot (\text{column } j \text{ of } P^n)$$

- where:

$$q_i = \text{state } i \text{ probability at time } 0$$

The Cola Example (I)

- Suppose our company produces two brands of Cola (i.e., Cola1, and Cola2) and there are no other Colas on the market. A person buying Cola1 will buy Cola1 again with probability 0.9. A persona buying Cola2 will buy Cola2 again with probability 0.8.

$$P = \begin{array}{c|cc} & \text{Cola1} & \text{Cola2} \\ \hline \text{Cola1} & 0.90 & 0.10 \\ \text{Cola2} & 0.20 & 0.80 \end{array}$$

  - Someone has bought Cola2, what's the probability he/she will buy Cola1 after 2 times?
  - Someone has bought Cola1, what's the probability he/she will buy Cola1 again after 3 times?
  - Suppose at some time 60% of clients bought Cola1 and 40% Cola2. After three purchases what's the percentage of people buying Cola1?

The Cola Example (II)

- Someone has bought Cola2, what's the probability he/she will buy Cola1 after 2 times?
  - $P(X_2=1|X_0=2)=P_{21}(2)$

$$P2=\begin{pmatrix} 0.90 & 0.10 \\ 0.20 & 0.80 \end{pmatrix}\begin{pmatrix} 0.90 & 0.10 \\ 0.20 & 0.80 \end{pmatrix}=\begin{pmatrix} 0.83 & 0.17 \\ & 0.66 \end{pmatrix}$$

- Someone has bought Cola1, what's the probability he/she will buy Cola1 again after 3 times?
  - $P(X_3=1|X_0=1)=P_{11}(3)$

$$P3=\begin{pmatrix} 0.83 & 0.17 \\ 0.34 & 0.66 \end{pmatrix}\begin{pmatrix} 0.90 & 0.10 \\ 0.20 & 0.80 \end{pmatrix}=\begin{pmatrix} & 0.219 \\ 0.438 & 0.562 \end{pmatrix}$$

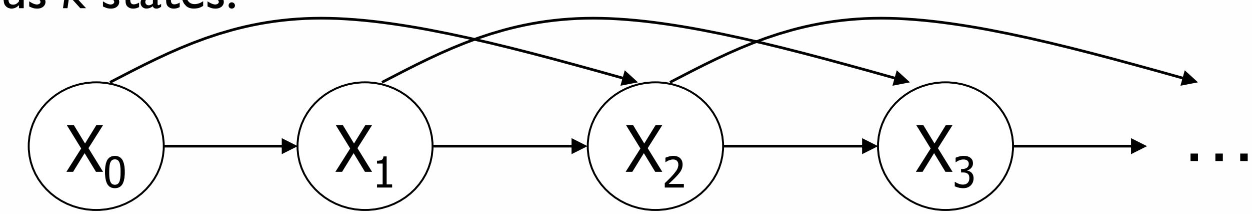
The Cola Example (III)

- Suppose at some time 60% of clients bought Cola1 and 40% Cola2. After three purchases what's the percentage of people buying Cola1?

$$p=\sum_i q_i \cdot P_{ij}(3) = q \cdot (\text{column } 1 \text{ of } P^3)$$

$$p= \begin{bmatrix} 0.60 & 0.40 \end{bmatrix} \begin{bmatrix} 0.781 \\ 0.438 \end{bmatrix} = 0.6438$$

- **<u>Note</u>:** What we have discussed so far is the first-order Markov Chain. More generally, in k^{th}-order Markov Chain, each state transition depends on previous k states.



<u>What's the size of transition probability matrix?</u>

A Bunch of Definitions

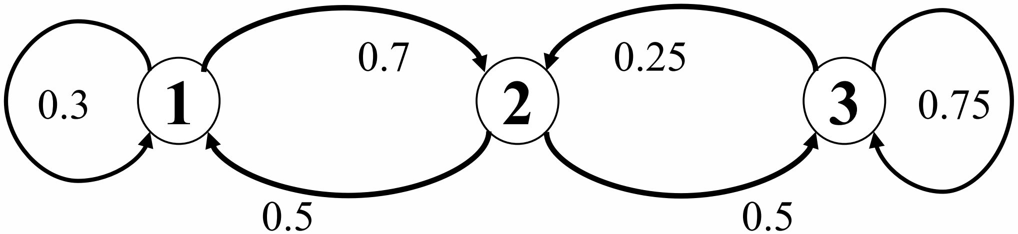
- Given a Markov Chain we define:
  - State *j* is *reachable* from *i* if it exist a path from *i* to *j*
  - States *i* and *j* *communicate* if *i* is reachable from *j* and viceversa
  - A set of states *S* in a Markov Chain is *closed* if no state outside S is reachable from a state in *S*
  - A state *i* is an *absorbing* state if $p_{ii}=1$
  - A state *i* is *transient* if exists *j* reachable from *i*, but *i* is not reachable from *j*
  - A state that is not transient is defined as *recurrent*
  - A state *i* is *periodic* with period $k>1$ if *k* is the biggest number that divides the length of all path from *i* to *i*
  - A state that is not periodic is said *a-periodic*

- If all states in a Markov Chain are *recurrent*, *a-periodic*, and *communicate* with each other, it is said to be **Ergothic**

Examples of Ergothic Markov Chains

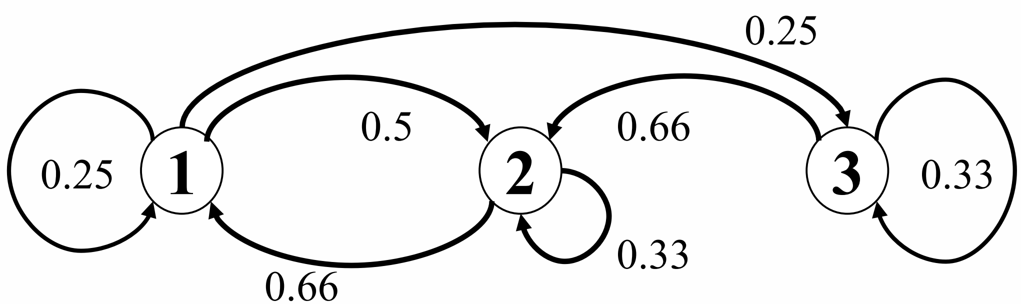
- A simple example of Ergothic Markov Chain is the following:

$$P = \begin{pmatrix} 0.3 & 0.7 & 0 \\ 0.5 & 0 & 0.5 \\ 0 & 0.25 & 0.75 \end{pmatrix}$$

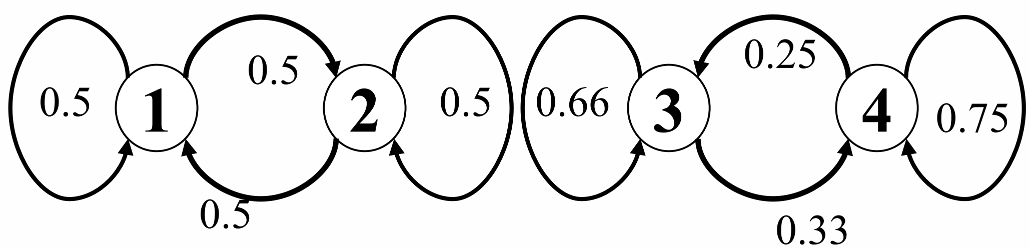


- Do the following transitions represent Ergothic Markov Chains?

$$P = \begin{pmatrix} 1/4 & 1/2 & 1/4 \\ 2/3 & 1/3 & 0 \\ 0 & 2/3 & 1/3 \end{pmatrix}$$



$$P = \begin{pmatrix} 1/2 & 1/2 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 \\ 0 & 0 & 2/3 & 1/3 \\ 0 & 0 & 1/4 & 3/4 \end{pmatrix}$$



Steady State Distribution

- Being *P* the transition matrix of an Ergothic Markov Chain with *n* states we have that

$$\lim_{n \to +\infty} \boldsymbol{P}_{ij}(n) = \pi_j$$

- With $\pi = [\pi_1\ \pi_2\ \pi_3\ \ldots\ \pi_n] = \pi \cdot P$ being the *Steady State Distribution*

- The Cola Example:

$$P = \begin{pmatrix} 0.9 & 0.1 \\ 0.2 & 0.8 \end{pmatrix}$$

$$\pi = \begin{pmatrix} 0.67 & 0.33 \\ 0.67 & 0.33 \end{pmatrix}$$

STEADY STATE

n	$P_{11}(n)$	$P_{12}(n)$	$P_{21}(n)$	$P_{22}(n)$
1	.90	.10	.20	.80
2	.83	.17	.34	.66
3	.78	.22	.44	.56
5	.72	.28	.56	.44
10	.68	.32	.65	.35
20	.67	.33	.67	.33
30	.67	.33	.67	.33
40	.67	.33	.67	.33

Transitory Behavior

- The behavior of a Markov Chain before getting to the Steady State id defined *transitory*

- We can compute the expected *number of transition* to reach state *j* being in state *i* for an Ergothic Markov Chain:

$$m_{ij} = p_{ij}(1) + \sum_{k\neq j} p_{ik} \cdot (1+m_{kj}) = 1+\sum_{k\neq j} p_{ik} \cdot m_{kj}$$

- The Cola Example:
  - How many bottle on average a Cola1 buyer will have before switching to Cola2?

$$m_{12}=1+\sum_{k\neq j} p_{1k} \cdot m_{k2}=1+0.9 \cdot m_{12} \qquad \longrightarrow \qquad m_{12}=10$$

  - What about viceversa?

$$m_{21}=1+\sum_{k\neq j} p_{2k} \cdot m_{k1}=1+0.8 \cdot m_{21} \qquad \longrightarrow \qquad m_{21}=5$$

Dealing with Absorbing States

- We have and *absorbing Markov Chain* if there exist one or more absorbing states and all the other are transient.

- For an absorbing Markov Chain we can write the transition matrix as:

$$P = \left(\begin{array}{c|c} Q & R \\ \hline 0 & I \end{array}\right)$$

- where:
  - *Q* is the transition matrix for transient states
  - *R* is the trantion matrix from transient to absorbing states

- What kind of inference we could make with this model?
  - How long it will take to get in an absorbing state given that we start from a transient one?
  - Starting from a transient state, how long does it takes to get to an absorbing one?

Inference in Absorbing Markov Chains

- How long I remain in a transient state given that we start from a transient one?
  - Being in a transient state *i* the average time spent in a transient state *j* is the *ij*-th element of $(I-Q)^{-1}$
- Starting from a transient state, how long does it takes to get to an absorbing one?
  - Being in transient state *i* the probability to get into an absorbing state *j* is the *ij*-th element of $(I-Q)^{-1} \cdot R$
- <u>Example</u>: in a company there are 3 levels: junior, senior, partner. You can leave the company as partner or not
  - How long does a junior remains in the company?
  - What's the probability for a junior to leave the company as partner?

$$
P = \begin{array}{c} \begin{array}{ccccc} J & S & P & LN & LP \end{array} \\ \left(\begin{array}{ccc|cc} 0.80 & 0.15 & 0 & 0.05 & 0 \\ 0 & 0.70 & 0.20 & 0.10 & 0 \\ 0 & 0 & 0.95 & 0 & 0.05 \\ \hline 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{array}\right) \end{array}
$$

The Company Example

- How long does a junior remains in the company?

$$(I\text{-}Q)^{-1} = \begin{pmatrix} 5 & 2.5 & 10 \\ 0 & 3.3 & 13.3 \\ 0 & 0 & 20 \end{pmatrix}$$

  - He/she will stay as Junior: $m_{11} = 5$
  - He/she will stay as Senior: $m_{12} = 2.5$
  - He/She will stay as Partner: $m_{13} = 10$

<u>17.5 years!</u>

- What's the probability for a junior to leave the company as partner?

$$(I\text{-}Q)^{-1} \cdot R = \begin{pmatrix} 0.5 & 0.5 \\ 0.3 & 0.7 \\ 0 & 1 \end{pmatrix}$$

  - He/She will end up in state LP: $m_{12} = 0.5$

Exercise: Gambler's Ruin

- Suppose we are a gambler and we start from a 3$ capital, with probability $p=1/3$ we can win *1$* and with probability $1\text{-}p=2/3$ we loose *1$*. We fail if out capital get to 0 and we win if our capital becomes 5.

$$X_0=3 \rightarrow X_1=? \rightarrow X_2=? \rightarrow X_3=? \rightarrow \ldots$$

- We can describe our capital as a Markov Chain being X_t our capital:
  - Possible states: *0, 1, 2, 3, 4, 5*
  - Transition probability: $p(X_{t+1}=X_t+1)=1/3$, $p(X_{t+1}=X_t-1)=2/3$
- What kind of reasoning can we apply to this model?
  - What's the probability of sequence 3, 4, 3, 2, 3, 2, 1, 0?
  - What's the probability of success for the gambler?
  - What's the average number of bets the gambler will make?

Why Should I Care All This Crazy Math?

"Nice, but unless I want to gamble why should I care? I'm a computer engineer what this has to do with practical intelligent systems?"

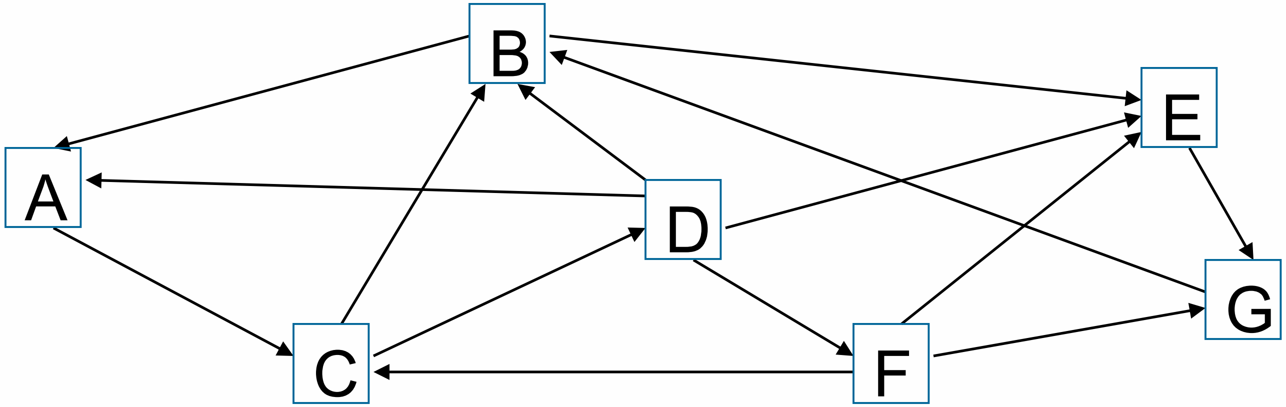
- What do you this is the greatest revolution (or revolutionary company) on the web in the last decade?
- Assume a link from page A to page B is a recommendation of page B by the author of A (we say B is successor of A).
  - Quality of a page is related to its *in-degree*.
  - The of a page is related to the quality of pages *linking to it*
- This recursively defines the ***PageRank*** of a page [Brin & Page '98]

For a (better) detailed description feel free to read:
http://www-db.stanford.edu/~backrub/google.html
http://www.iprcom.com/papers/pagerank/

Google's PageRank

- Suppose the web is an Ergothic Markov Chain (I know this is a big assumption). Consider browsing as an infinite random walk (surfing):
  - Initially the surfer is at a random page
  - At each step, the surfer proceeds
    - to a randomly chosen web page with probability d
    - to a randomly chosen successor of the current page with probability 1-d

- The PageRank of a page is the fraction of steps the surfer spends on it in the limit.

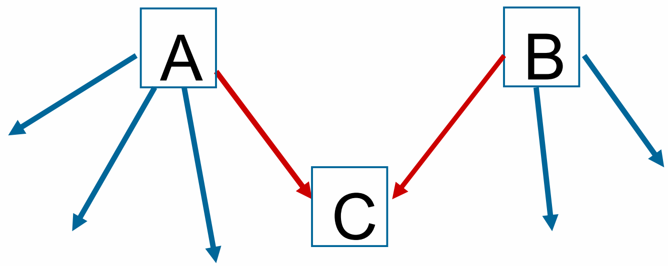


Definition of PageRank

- PageRank = the steady state probability for this Markov Chain

$$PageRank(u) = d + (1-d) \sum_{(v,u)\in E} PageRank(v) / outdegree(v)$$

  - *n* is the total number of nodes in the graph
  - *d* is the probability of a random jump



PageRank(C) = d/n+(1-d)(1/4 PageRan(A) +1/3 PageRank(B))

- Summarizes the "web opinion" about the page importance
  - Query-independent
  - It can be faked … read the provided links if you are curious!

POLITECNICO DI MILANO

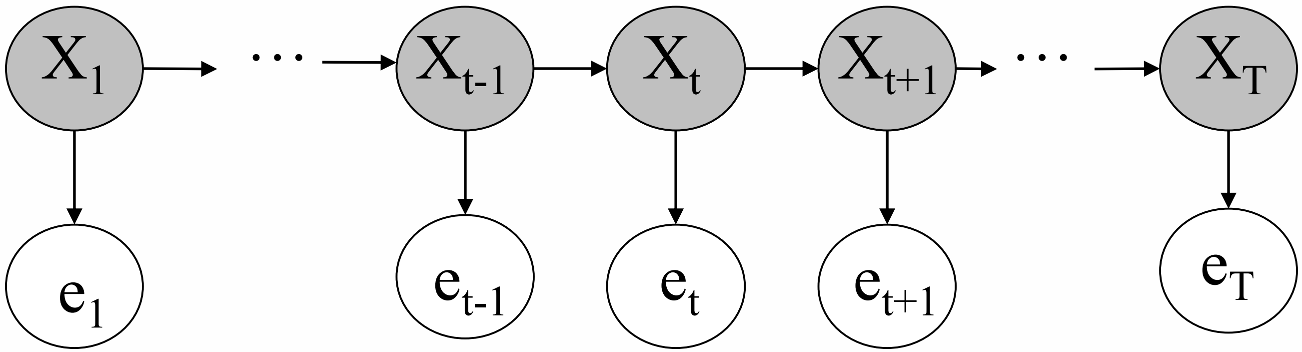
Hidden Markov Models

Information Retrieval and Data Mining

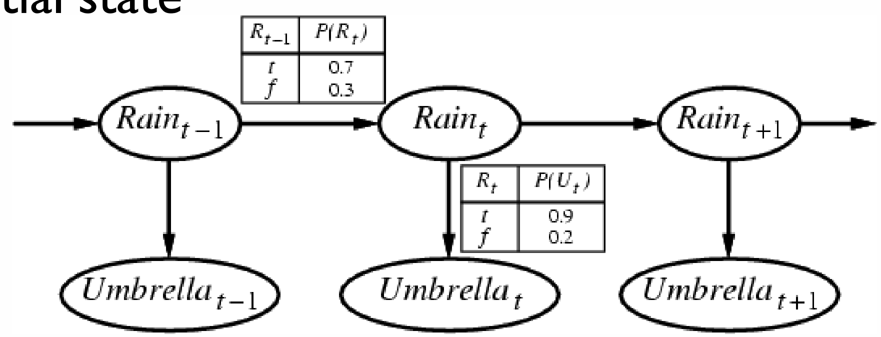
Prof. Matteo Matteucci

Hidden Markov Models (HMM)

- We may not observe directly the states. The we get another Bayesian Network named as <u>*Hidden Markov Model*</u> *(HMM)*.



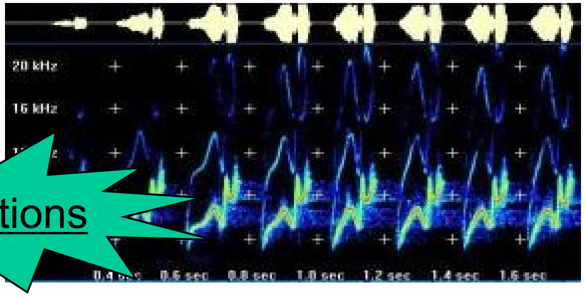
- An HMM is described by a quintuple <*S,E,P,A,B*>
  - *S* : $\{s_1, \dots, s_N\}$ are the values for the hidden states
  - *E* : $\{e_1, \dots, e_T\}$ are the values for the observations
  - *P*: probability distribution of the initial state
  - *A*: transition probability matrix
  - *B*: emission probability matrix



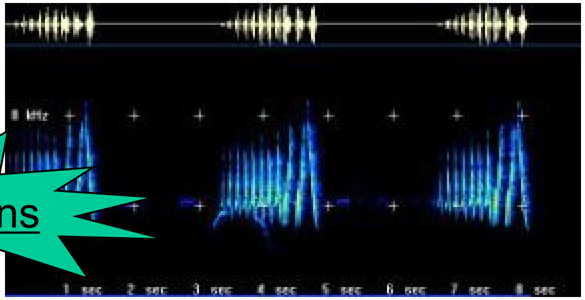
For a deeper description feel free to read:
http://www.cs.ubc.ca/~murphyk/Bayes/rabiner.pdf

An Example: The Audio Spectrum

- Audio Spectrum of the song for the Prothonotary Warbler



- Audio Spectrum of the song for the Chestnut-sided Warbler



- What can we ask to an HMM?
  - What bird is this? → Time Series Classification
  - How will the song continue? → Time Series Prediction
  - Is this bird sick? → Outlier Detection
  - What phases does this song have? → Time Series Segmentation

Another Time Series Problem

- What can we ask to this HMM?
  - Will the stock go up or down? → Time Series Prediction
  - What type stock is this (eg, risky)? → Time Series Classification
  - Is the behavior abnormal (eg, BF)? → Outlier Detection

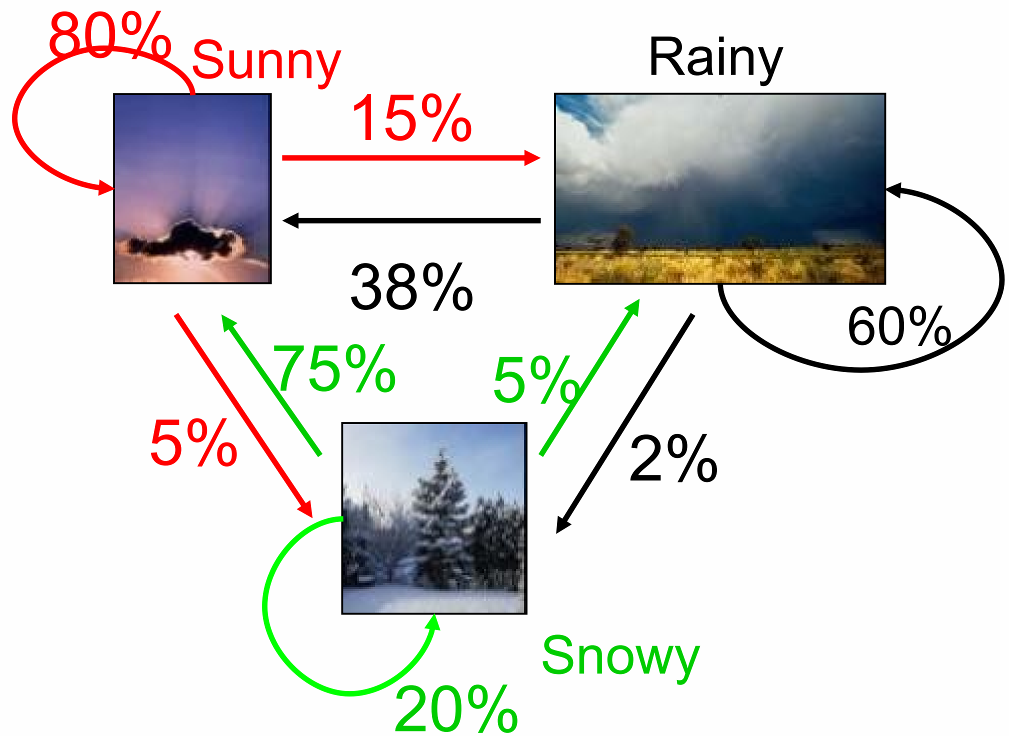
Music Analysis

- What can we ask to this HMM?
  - Can we compose more of that? → Time Series Prediction
  - Is this Beethoven or Bach? → Time Series Classification
  - Can we segment it into themes? → Time Series Segmentation

Weather: A Markov Chain Model

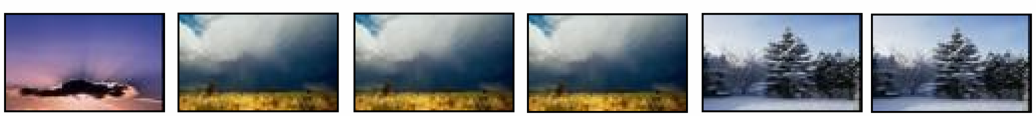
- States: $\{S_{sunny}, S_{rainy}, S_{snowy}\}$
- State transition probabilities:

$$P = \begin{pmatrix} 0.80 & 0.15 & 0.05 \\ 0.38 & 0.60 & 0.02 \\ 0.75 & 0.05 & 0.20 \end{pmatrix}$$



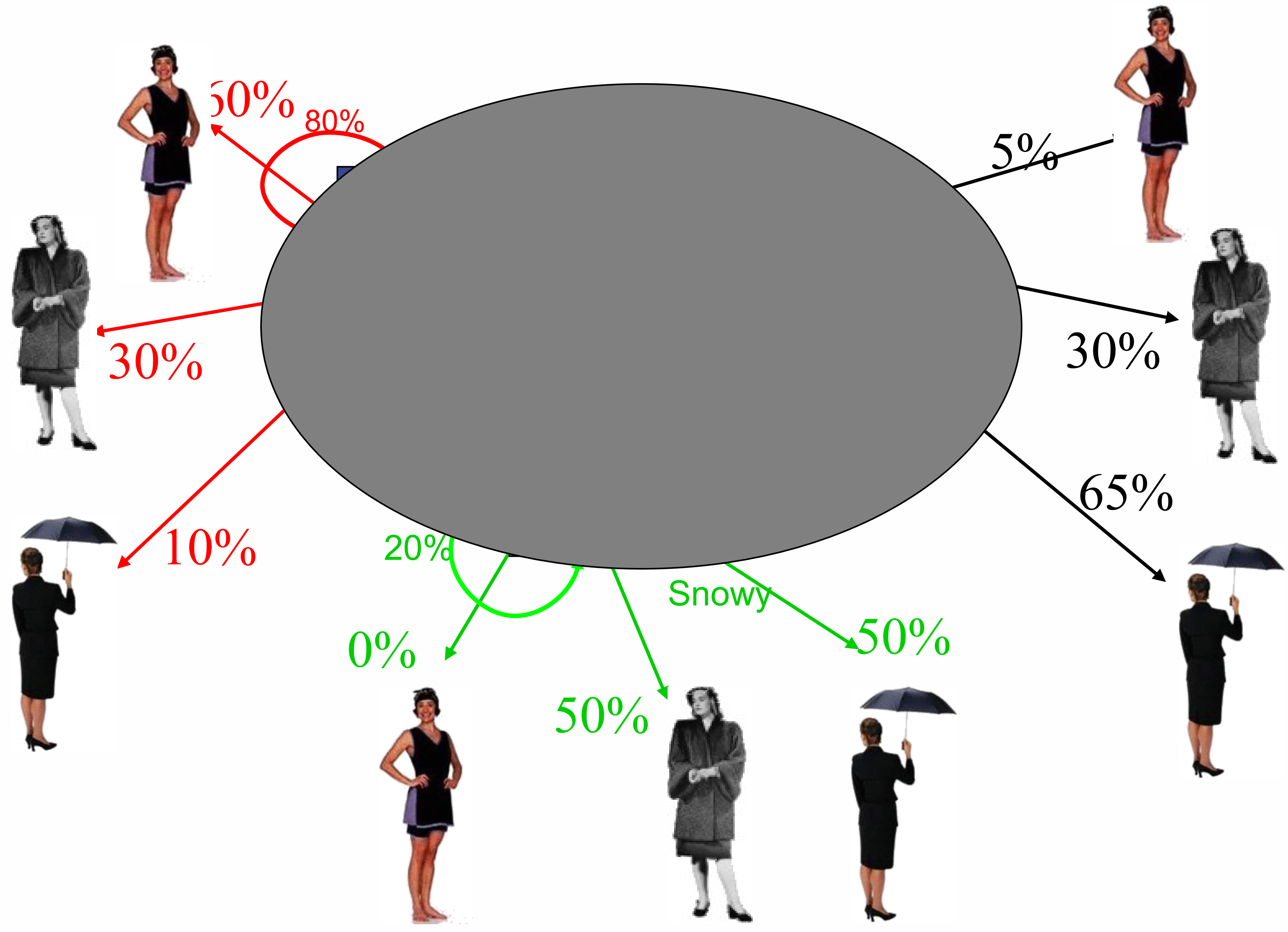
- Initial state distribution:

$$q = (0.7 \;\; 0.25 \;\; 0.05)$$

- Given: 
- What is the probability of this series?

$$P(s) = P(S_{sunny})P(S_{rainy}|S_{sunny})P(S_{rainy}|S_{rainy})P(S_{rainy}|S_{rainy})P(S_{snowy}|S_{rainy})P(S_{snowy}|S_{snowy}) = 0.7 \cdot 0.15 \cdot 0.6 \cdot 0.6 \cdot 0.02 \cdot 0.2 = 0.0001512$$

Weather: An Hidden Markov Models



Ingredients and Fundamental Questions

- States: $\{S_{sunny}, S_{rainy}, S_{snowy}\}$
- Observations: $\{O_{shorts}, O_{coat}, O_{umbrella}\}$
- State transition probabilities: $A = \begin{pmatrix} 0.80 & 0.15 & 0.05 \\ 0.38 & 0.60 & 0.02 \\ 0.75 & 0.05 & 0.20 \end{pmatrix}$
- Observation probabilities: $B = \begin{pmatrix} 0.60 & 0.30 & 0.10 \\ 0.05 & 0.30 & 0.65 \\ 0.00 & 0.50 & 0.50 \end{pmatrix}$
- Initial state distribution: $P = (0.7 \;\; 0.25 \;\; 0.05)$
- Given:  ..
  - What is the probability of this series?
  - What is the underlying sequence of state?
  - How can I learn my HMM parameters?

Computing Forward Probability

- We define the Forward Probability as the probability of actual state and observations

$$P(X_t=s_i,\ e_{1:t})$$

- Why compute forward probability?
  - Probability of observations: $P(e_{1:t})$.
  - Prediction: $P(X_{t+1}=s_i \mid e_{1:t})=?$

Same form, use recursion

$$
\begin{aligned}
P(X_t=s_i,\ e_{1:t}) &= P(X_t=s_i, e_{1:t-1}, e_t)\\
&= \Sigma_j\, P(X_{t-1}=s_j, X_t=s_i, e_{1:t-1}, e_t)\\
&= \Sigma_j\, P(e_t|X_t=s_i, X_{t-1}=s_j, e_{1:t-1}) P(X_t=s_i, X_{t-1}=s_j, e_{1:t-1})\\
&= \Sigma_j\, P(e_t|X_t=s_i) P(X_t=s_i|X_{t-1}=s_j, e_{1:t-1}) P(X_{t-1}=s_j, e_{1:t-1})\\
&= \Sigma_j\, P(e_t|X_t=s_i) P(X_t=s_i|X_{t-1}=s_j) P(X_{t-1}=s_j,\ e_{1:t-1})
\end{aligned}
$$

$$
\begin{aligned}
\alpha_i(t) &= P(X_t=s_i,\ e_{1:t})\\
&= \Sigma_j P(X_t=s_i|X_{t-1}=s_j) P(e_t|X_t=s_i)\, \alpha_j(t-1)\\
&= \Sigma_j\, A_{ij}\, B_{iet}\, \alpha_j(t-1)
\end{aligned}
$$

The Viterbi Algorithm

- From observations, compute the most likely hidden state sequence:

$$\text{argmax } P(x_{1:t}|e_{1:t}) = \text{argmax } P(x_{1:t}, e_{1:t})/P(e_{1:t})$$
$$= \text{argmax } P(x_{1:t}, e_{1:t})$$

- By applying the Bayesian Network factorization

$$P(x_{1:t}, e_{1:t}) = P(X_0) \Pi_{i=1,t} P(X_i|X_{i-1}) P(e_i|X_i)$$

- The solution we are looking for is the one that minimizes

$$-\log P(x_{1:t}, e_{1:t}) = -\log P(X_0) + \Sigma_{i=1,t}(-\log P(X_i|X_{i-1}) - \log P(e_i|X_i))$$

- Given a HMM construct a graph that consists 1+t*N nodes:
  - One initial node and *N* node at time *i* where j^{th} represents $X_i=s_j$.
  - The link between the nodes $X_{i-1}=s_j$ and $X_i=s_k$ is associated with the length $-\log(P(X_i=s_k| X_{i-1}=s_j)P(e_i|X_i=s_k))$
- The problem becomes that of finding the shortest path from $X_0=s_0$ to one of the nodes $X_t=s_t$.

Baum-Welch Algorithm

- The previous two kinds of computation needs parameters $\mu=(P, A, B)$. Where do the probabilities come from?
- Solution: Baum-Welch Algorithm (special case of EM)
  - Unsupervised learning from observations
  - Find $\text{argmax}_{\mu}\ P\mu(e_{1:t})$
- Given an observation sequence, find out which transition probability and emission probability table assigns the highest probability to the observations:
  1. Start with an initial set of parameters μ_0 (possibly arbitrary)
  2. Compute pseudo counts: how many times the transition from $X_{i-i}=s_j$ to $X_i=s_k$ occurred?
  3. Use the pseudo counts to obtain a better set of parameters μ_1
  4. Iterate until $P_{\mu 1}(e_{1:t})$ is not bigger than $P_{\mu}(e_{1:t})$

Pseudo Counts and Backward Probability

- Given the observation sequence $e_{1:T}$,
  - pseudo count of state s_i at time t is the probability $P(X_t=s_i|e_{1:T})$

$$
\begin{aligned}
P(X_t=s_i|e_{1:T}) &= P(X_t=s_i, e_{1:t}, e_{t+1:T})/P(e_{1:T}) \\
&= P(e_{t+1:T}| X_t=s_i, e_{1:t})P(X_t=s_i, e_{1:t})/P(e_{1:T}) \\
&= P(e_{t+1:T}| X_t=s_i)P(X_t=s_i|e_{1:t})P(e_{1:t})/P(e_{1:T}) \\
&= \alpha_i(t)\ \beta_i(t)/P(e_{t+1:T}|e_{1:t})
\end{aligned}
$$

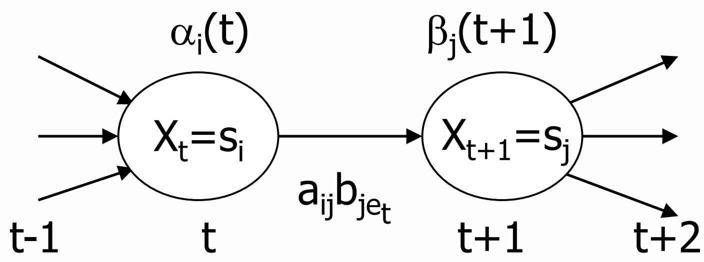
  - pseudo counts of the link from $X_t=s_i$ to $X_{t+1}=s_j$ is the probability

$$
\begin{aligned}
P(X_t=s_i,X_{t+1}=s_j|e_{1:T})&=P(X_t=s_i,X_{t+1}=s_j,e_{1:t},e_{t+1},e_{t+2:T})/P(e_{1:T}) \\
&= P(X_t=s_i,e_{1:t})P(X_{t+1}=s_j|X_t=s_i)P(e_{t+1}|X_{t+1}=s_j) \\
&\quad P(e_{t+2:T}|X_{t+1}=s_j)/P(e_{1:T}) \\
&=P(X_t=s_i,e_{1:t})A_{ij}B_{jet+1}P(e_{t+2:T}|X_{t+1}=s_j)/P(e_{1:T}) \\
&= \alpha_i(t)\ A_{ij}\ B_{jet}\ \beta_j(t+1)/P(e_{1:T})
\end{aligned}
$$

- Being $\beta_j(t)=P(e_{t+1},\dots,e_T|X_t=s_j)$ *we can compute it backward*
  - $\beta_j(T)=1;$
  - $\beta_j(t)= \Sigma_j\, A_{ij}\ B_{jet}\ \beta_j(t+1).$

HMM Parameters Update

- We can efficiently compute forward and backward probability for all the states in the Hidden Markov Model



- To update our estimate of HMM parameters
  - *count(i)*: the total pseudo count of state s_i.
  - *count(i,j)*: the total pseudo count of transition from s_i to s_j.
  - Add $P(X_t=s_i,X_{t+1}=s_j|e_{1:T})$ to *count(i,j)*
  - Add $P(X_t=s_i|e_{1:T})$ to *count(i)*
  - Add $P(X_t=s_i|e_{1:T})$ to *count(i,et)*
  - Updated A_{ij}= *count(i,j)/count(i)*
  - Updated B_{jet}=*count(j,e_t)/count(j)*

Summary on HMM

- HMMs are generative probabilistic models for time series with hidden information (state).

- There a few issues remaining:
  - Zero probability problem
    - Training sequence: AAABBBAAA
    - Test sequence: AAABBBCAAA
  - Finding "right" number of states, right structure
  - Numerical instabilities

- Beside these problems they are extremely practical, best known methods in speech recognition, computer vision, robotics, …

You'd be surprised by the relationships between HMM and Kalman Filtering or Kalman Smoothing!