Machine Learning for Data Science (CS4786)
Lecture 18

Graphical Models and Hidden Markov Models

Course Webpage :
http://www.cs.cornell.edu/Courses/cs4786/2016fa/

Bayesian Networks

- Bayes net: directed acyclic graph + P(node|parents)
- Directed acyclic graph G = (V,E)
  - Edges going from parent nodes to child nodes
  - Direction indicates parent “generates” child
- Provide conditional probability table/distribution P(node|parents)

$$P(X_1, \ldots, X_N) = \prod_{i=1}^{N} P(X_i | \text{Parents}(X_i))$$

Representational Power

- Not all joint distributions can be represented by Bayesian Networks
- Eg. $X_1 \perp X_4 \mid X_3, X_2$ and $X_3 \perp X_2 \mid X_1, X_4$
  This dependence can never be captured by a bayesian network, Why?

Representational Power

- Not all joint distributions can be represented by Bayesian Networks
- Eg. $X_1 \perp X_4 \mid X_3, X_2$ and $X_3 \perp X_2 \mid X_1, X_4$
  This dependence can never be captured by a bayesian network, Why?

Which distributions can be represented by Bayesian networks?

LOCAL MARKOV PROPERTY

- Each variable is conditionally independent of its non-descendants given its parents
- Any joint distribution satisfying the local markov property w.r.t. graph factorizes over the graph

Why?

Factorizing Joint Probability

- Fact about DAG: we obtain an ordering of nodes (called topological sort) such that for every directed edge between X_i to X_j, X_i appears before X_j in sorted order.
- Assume nodes are arranged according to some topological sort
- For any distribution we have:

$$
\begin{aligned}
P_\theta(X_1, \ldots, X_N) &= \prod_{i=1}^{N} P_\theta(X_i | X_1, \ldots, X_{i-1}) \\
&= \prod_{i=1}^{N} P_\theta(X_i | \text{Parents}(X_i))
\end{aligned}
$$

Two main questions

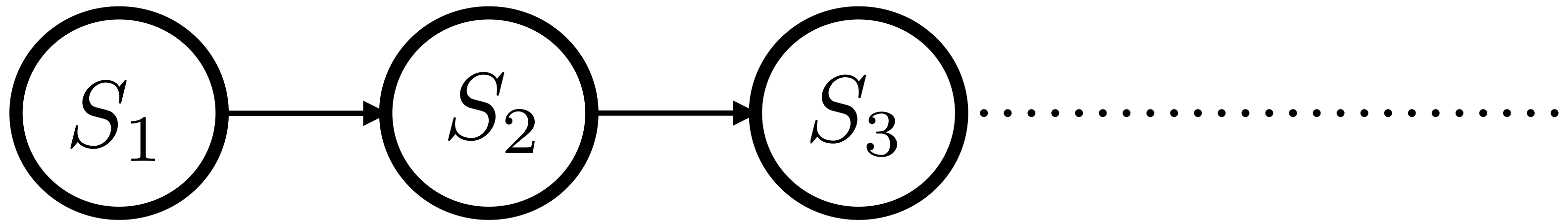
- Learning/estimation: Given observations, can we learn the parameters for the graphical model ?
- Inference: Given model parameters, can we answer queries about variables in the model
    - Eg. what is the most likely value of a latent variable given observations
    - Eg. What is the distribution of a particular variable conditioned on others

Hidden Markov Model (HMM)

- Speech recognition
- Natural language processing models
- Robot localization
- User attention modeling
- Medical monitoring

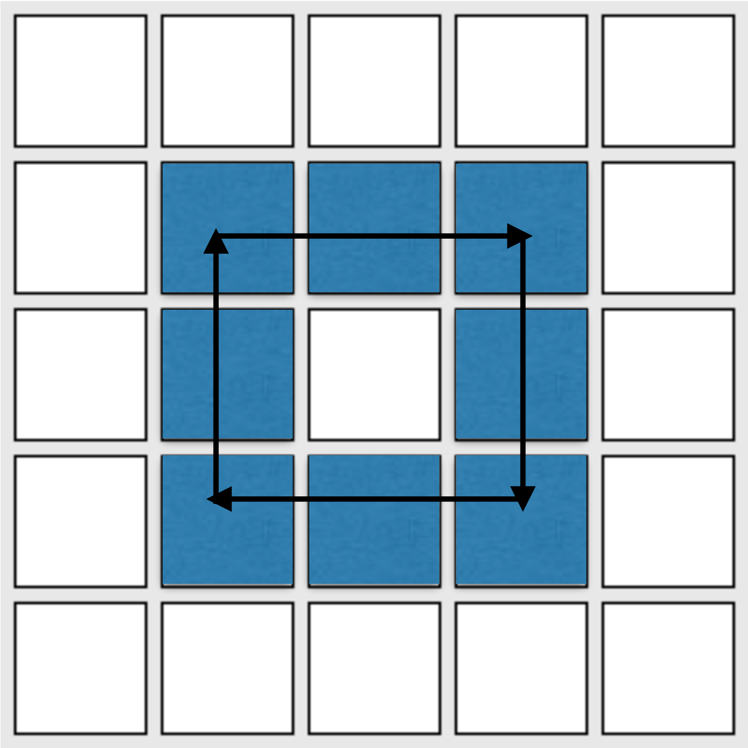
Time! … sequence of observations

Markov Model



- Each node is identically distributed given its predecessor (stationary)
- The values the nodes take are called states
- Parameters?
  - $P(S_1)$ the initial probability table
  - $P(S_t|S_{t-1})$ the transition probabilities

Markov Model



Bot tends to follow outlined path, but with some probability jumps to arbitrary neighbor

- Number of states: 25 (one for each location)
- For white boxes probability of jumping to any of the 4 neighbors is same 1/4
- For Blue boxes, probability of following path is 0.9 and jumping to some other neighbor is 0.0333333

Markov Model

- If we observe the bot long enough, we get an estimate of its behavior (the transition table of jumping from state to state)
- If we observe enough number of times, we can also estimate initial distribution over states

Markov Model

- Inference question: what is probability that we will be in state k at time t? $P(S_t = k)$?

Answer:

$$P(S_t = k) = \sum_{s_1=1}^{K} \cdots \sum_{s_{t-1}=1}^{K} P(S_1 = s_1, \ldots, S_{t-1} = s_{t-1}, S_t = k)$$

$$= \sum_{s_1=1}^{K} \cdots \sum_{s_{t-1}=1}^{K} \prod_{i=1}^{t-1} (P(S_i = s_i | S_{i-1} = s_{i-1}) \times P(S_t = k | S_{t-1} = s_{t-1}))$$

For every t we can repeat the above or…

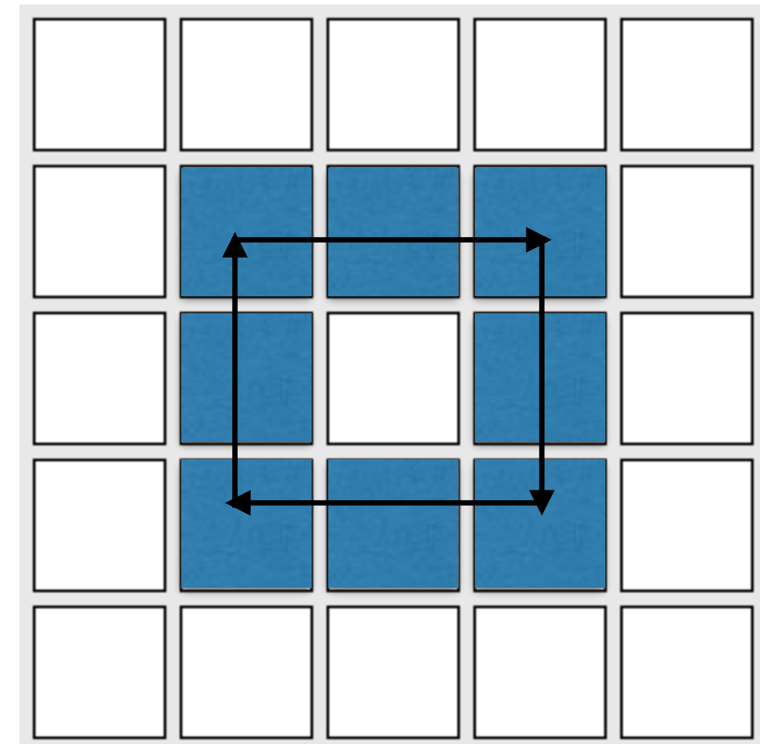
$$P(S_t = k) = \sum_{s_{t-1}=1}^{K} P(S_t = k | S_{t-1} = s_{t-1}) P(S_{t-1} = s_{t-1})$$

recursively compute probability of previous state

Markov Model

- As time goes by, $P(S_t = k)$ approaches a fixed distribution called stationary distribution
- Without any further observations, you are unlikely to find the bot on a new run (only by luck)

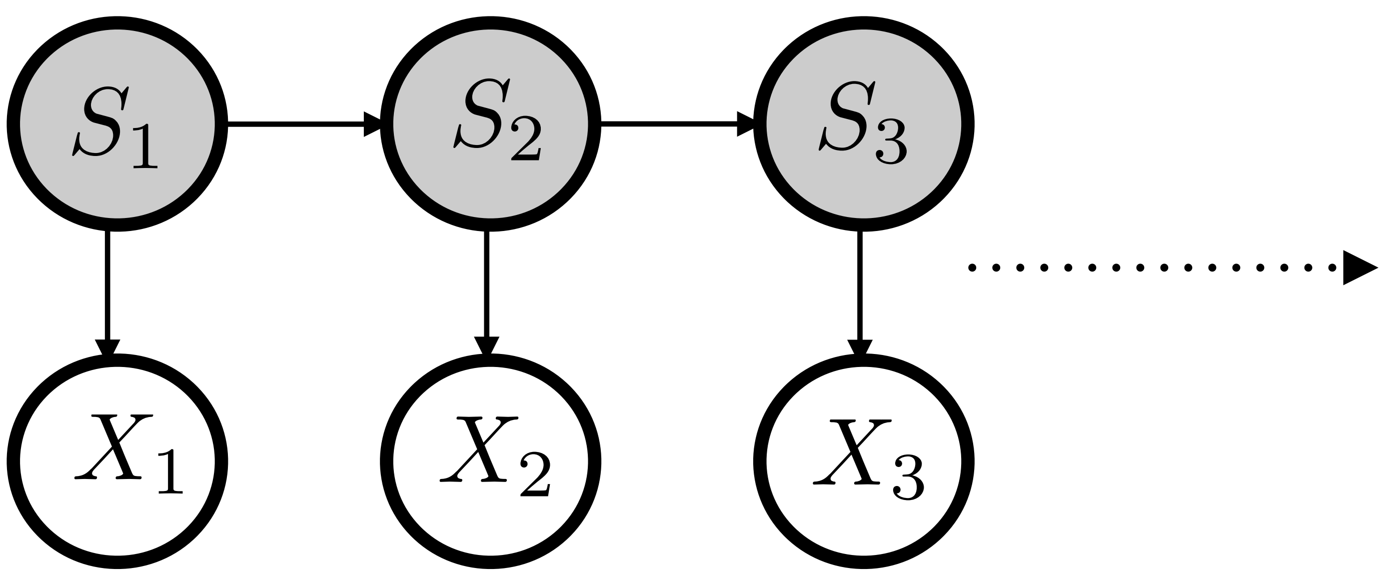
Hidden Markov Model (HMM)



Same example:

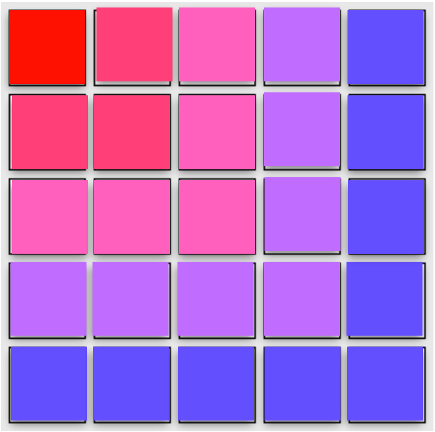
But you don't observe location (dark room)

You hear how close the bot is!



X_t's are loudness of what you hear

HIDDEN MARKOV MODEL (HMM)



- Both during the initial training/estimation phase, you never see the bot you only hear it
- But you hear it at any point in time
- We will come back to learning next class.
- What is probability that bot will be in state k at time t given the entire sequence of observations?

$$P(S_t = k | X_1, \ldots, X_N)?$$

Inference in HMM

$$
\begin{aligned}
P(S_t = k|X_1,\ldots,X_N) \\
&\propto P(X_{t+1},\ldots,X_N|S_t = k, X_1,\ldots,X_t)P(S_t = k|X_1,\ldots,X_t) \\
&\propto P(X_{t+1},\ldots,X_N|S_t = k, X_1,\ldots,X_t)P(S_t = k, X_1,\ldots,X_t) \\
&\propto P(X_{t+1},\ldots,X_N|S_t = k, X_1,\ldots,X_t)P(X_t|S_t = k, X_1,\ldots,X_{t-1})P(S_t = k, X_1,\ldots,X_{t-1}) \\
&= P(X_{t+1},\ldots,X_N|S_t = k)P(X_t|S_t = k)P(S_t = k, X_1,\ldots,X_{t-1})
\end{aligned}
$$

We know $P(X_t|S_t = k)$'s and $P(S_t|S_{t-1})$

Compute $P(X_{t+1},\ldots,X_N)$ and $P(S_t = k, X_1,\ldots,X_{t-1})$ recursively.

Real World Applications

- Speech recognition HMMs:
  - Observations are wave forms (continuous valued)
  - States are specific positions in specific words (so, tens of thousands)
- Robot tracking:
  - Observations are range readings (continuous)
  - States are positions on a map (continuous)
- Machine translation HMMs:
  - Observations are words (tens of thousands)
  - States are translation options