Graphical Models

Course Webpage:
http://www.cs.cornell.edu/Courses/cs4786/2017fa/
Data
DATA

\[ \theta \in \Theta \]

\[ P_\theta \text{ explains data} \]
Probabilistic Model

Latent Variables

Observed variables

$\theta \in \Theta$

$P_\theta$ explains data
Probabilistic Model

$\theta \in \Theta$

$P_{\theta}$ explains data

$x_1, \ldots, x_N$

Latent Variables

Observed variables
Abstract away the parameterization specifics

Focus on relationship between random variables
Let $X = (X_1, \ldots, X_N)$ be the random variables of our model (both latent and observed)

- Joint probability distribution over variable can be complex esp. if we have many complexly related variables

- Can we represent relation between variables in conceptually simpler fashion?

- We often have prior knowledge about the dependencies (or conditional (in)dependencies) between variables
A graph whose nodes are variables $X_1, \ldots, X_N$

Graphs are an intuitive way of representing relationships between large number of variables

Allows us to abstract out the parametric form that depends on $\theta$ and the basic relationship between the random variables.
Graphical Models

- A graph whose nodes are variables $X_1, \ldots, X_N$
- Graphs are an intuitive way of representing relationships between large number of variables
- Allows us to abstract out the parametric form that depends on $\theta$ and the basic relationship between the random variables.

Draw a picture for the generative story that explains what generates what.
Graphical Models

- Variables $X_i$ is written as $X_i$ if $X_i$ is observed
- Variables $X_i$ is written as $X_i$ if $X_i$ is latent
- Parameters are often left out (its understood and not explicitly written out). If present they don’t have bounding objects
- An directed edge $\rightarrow$ is drawn connecting every parent to its child (from parent to child)

$Y \rightarrow X_i \quad \quad X_1 \ldots X_N$ drawn repeatedly from $P(Y|X)$
Example: Sum of Coin Flips

\[ S_1 \rightarrow S_2 \rightarrow S_3 \]

\[ X_1 \rightarrow S_2 \rightarrow X_2 \rightarrow X_3 \]
Eg. Spam classification
Example: Mixture Models

Eg. Clustering
Mixture of Multinomials

\[ \mathbf{x}_n \rightarrow c_n \rightarrow \pi \]

\[ \mathbf{x}_n \rightarrow c_n \rightarrow N \]

\[ \pi \rightarrow c_n \]

\[ \mathbf{x}_n \rightarrow N \]
Example: Latent Dirichlet Allocation

\[ \alpha \rightarrow \pi_m \]

\[ \beta \rightarrow \varphi_k \]

\[ c_n \rightarrow x_n \]

Eg. Topic modelling
Example: Hidden Markov Model

Eg. Speech recognition
• Directed acyclic graph $G = (V,E)$ (graph with no directed cycle)
Bayesian Networks

- Directed acyclic graph $G = (V,E)$ (graph with no directed cycle)
  - Edges going from parent nodes to child nodes
Bayesian Networks

- Directed acyclic graph $G = (V,E)$ (graph with no directed cycle)
  - Edges going from parent nodes to child nodes
  - Direction indicates parent “generates” child
Example: CI and MI
**Example: CI and MI**

Marginally independent:

- Genetic Info of Mother
- Genetic Info of Father
Example: CI and MI

Marginally independent but Conditionally dependent given child

Genetic Info of Mother

Genetic Info of Father

Genetic Info of Child
EXAMPLE: CI AND MI

Marginally independent but Conditionally dependent given child

Genetic Info of Mother

Genetic Info of Father

Genetic Info of Child

Marginally dependent

Genetic Info of Sister

Genetic Info of Brother
Example: CI and MI

Marginally independent but Conditionally dependent given child

- Genetic Info of Mother
- Genetic Info of Father
- Genetic Info of Child

Marginally dependent but Conditionally independent given Parent

- Genetic Info of Parents
- Genetic Info of Sister
- Genetic Info of Brother
Conditional and Marginal Independence

- **Conditional independence**
  - $X_i$ is conditionally independent of $X_j$ given $A \subset \{X_1, \ldots, X_N\}$:
    
    $$X_i \perp X_j \mid A \iff P_\theta(X_i, X_j \mid A) = P_\theta(X_i \mid A) \times P_\theta(X_j \mid A)$$
    
    $$\iff P_\theta(X_i \mid X_j, A) = P_\theta(X_i \mid A)$$

- **Marginal independence**:
  
  $$X_i \perp X_j \mid \emptyset \iff P_\theta(X_i, X_j) = P_\theta(X_i)P_\theta(X_j)$$
BAYESIAN NETWORKS
Bayesian Networks

• Directed acyclic graph $G = (V,E)$ (graph with no directed cycle)

  • Edges going from parent nodes to child nodes

  • Direction indicates parent “generates” child
Bayesian Networks

- Directed acyclic graph $G = (V,E)$ (graph with no directed cycle)
- Edges going from parent nodes to child nodes
- Direction indicates parent “generates” child
- Local Markov Property: Each node conditionally independent of its non-descendants given its parents
Bayesian Networks

- Directed acyclic graph $G = (V, E)$ (graph with no directed cycle)
  - Edges going from parent nodes to child nodes
  - Direction indicates parent “generates” child
- Local Markov Property: Each node conditionally independent of its non-descendants given its parents

Joint probability factorizes as:

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

$$P_\theta(X_1, \ldots, X_N) = \prod_{i=1}^{N} P_\theta(X_i|X_1, \ldots, X_{i-1})$$
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:

$$P_\theta(X_1, \ldots, X_N) = \prod_{i=1}^{N} P_\theta(X_i | X_1, \ldots, X_{i-1})$$

$$= \prod_{i=1}^{N} P_\theta(X_i | \text{Parents}(X_i))$$

(Local Markov Property)
Bayesian Networks

- Bayes net: directed acyclic graph + P(node|parents)
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
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 \( P(node|parents) \)
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)

\[
P(X_1, \ldots, X_N) = \prod_{i=1}^{N} P(X_i | \text{Parents}(X_i))
\]
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?
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?
Two main questions
Two main questions

• Learning/estimation: Given observations, can we learn the parameters for the graphical model?
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
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
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
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

E-step in EM is inference of Latent given observed
Given parameters of a graphical model, we can answer any questions about distributions of variables in the model.

Example queries:

1. What is the probability of a given assignment for a subset of variables (marginal)?

2. What is the probability of a particular assignment of a subset of variables given observed values (evidence) of some subset of the variables (conditional)?

3. Given observed values (evidence) of some subset of variables what is the most likely assignment for a given subset of variables?
Inference in Graphical Models

Given parameters of a graphical model, we can answer any questions about distributions of variables in the model.

Example queries:

1. What is the probability of a given assignment for a subset of variables (marginal)?

2. What is the probability of a particular assignment of a subset of variables given observed values (evidence) of some subset of the variables (conditional)?

3. Given observed values (evidence) of some subset of variables what is the most likely assignment for a given subset of variables?

Suffices to be able to compute joint probability
Given parameters of a graphical model, we can answer any questions about distributions of variables in the model.

Example queries:

1. What is the probability of a given assignment for a subset of variables (marginal)?

2. What is the probability of a particular assignment of a subset of variables given observed values (evidence) of some subset of the variables (conditional)?

3. Given observed values (evidence) of some subset of variables what is the most likely assignment for a given subset of variables?

Suffices to be able to compute joint probability

Why?
Can compute any marginal from joint:

\[ P(A = a, B = b, C = c) = \sum_{d} P(A = a, B = b, C = c, D = d) \]
Can compute any marginal from joint:

\[ P(A = a, B = b, C = c) = \sum_d P(A = a, B = b, C = c, D = d) \]

Can compute any conditional from marginal:

\[ P(A = a | B = b, C = c) = \frac{P(A = a, B = b, C = c)}{P(B = b, C = c)} \]
Can compute any marginal from joint:

\[ P(A = a, B = b, C = c) = \sum_d P(A = a, B = b, C = c, D = d) \]

Can compute any conditional from marginal:

\[ P(A = a | B = b, C = c) = \frac{P(A = a, B = b, C = c)}{P(B = b, C = c)} \]

For Bayesian Networks \( P(\text{node}|\text{Parents}) \) completely defines joint.
Next class

- Start with example of Hidden Markov Model (HMM)