Introduction to Probabilistic Graphical Models

Christoph Lampert

IST Austria (Institute of Science and Technology Austria)

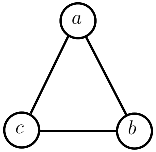
Factor Graphs

Relationship Factorizations to Graphs

- Consider $p(a, b, c) = \phi(a, b)\phi(b, c)\phi(c, a)$

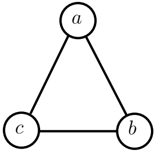
Relationship Factorizations to Graphs

- Consider $p(a, b, c) = \phi(a, b)\phi(b, c)\phi(c, a)$
- What is the graph of the corresponding Markov network?



Relationship Factorizations to Graphs

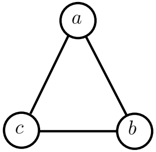
- Consider $p(a, b, c) = \phi(a, b)\phi(b, c)\phi(c, a)$
- What is the graph of the corresponding Markov network?



- How about this one? $p(a, b, c) = \phi(a, b, c)$

Relationship Factorizations to Graphs

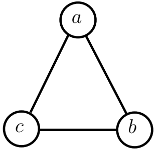
- Consider $p(a, b, c) = \phi(a, b)\phi(b, c)\phi(c, a)$
- What is the graph of the corresponding Markov network?



- How about this one? $p(a, b, c) = \phi(a, b, c)$
- The same!

Relationship Factorizations to Graphs

- Consider $p(a, b, c) = \phi(a, b)\phi(b, c)\phi(c, a)$
- What is the graph of the corresponding Markov network?



- How about this one? $p(a, b, c) = \phi(a, b, c)$
- The same!

- no one-to-one relation between the graph and the factorization of the potential functions!

Relationship Factorizations to Graphs

Why is this a problem?

- Many problems have only small (*e.g*. pairwise) interactions, *e.g*. "friendship" in Facebook

Relationship Factorizations to Graphs

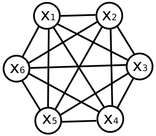
Why is this a problem?

- Many problems have only small (*e.g.* pairwise) interactions, *e.g.* "friendship" in Facebook
- $p(x_1, \ldots, x_6) = \frac{1}{Z} \prod_{i \neq j} \phi_{ij}(x_i, x_j)$ with $x_i \in \{1, \ldots, L\}$
- $\binom{6}{2} = 15$ factors of size 2 $\rightarrow$ distribution specified by $15L^2$ values

Relationship Factorizations to Graphs

Why is this a problem?

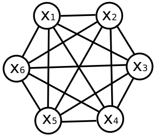
- Many problems have only small (*e.g.* pairwise) interactions, *e.g.* "friendship" in Facebook
- $p(x_1, \ldots, x_6) = \frac{1}{Z} \prod_{i \neq j} \phi_{ij}(x_i, x_j)$ with $x_i \in \{1, \ldots, L\}$
- $\binom{6}{2} = 15$ factors of size 2 $\rightarrow$ distribution specified by $15L^2$ values
- corresponding graph: fully connected



Relationship Factorizations to Graphs

Why is this a problem?

- Many problems have only small (*e.g.* pairwise) interactions, *e.g.* "friendship" in Facebook
- $p(x_1, \dots, x_6) = \frac{1}{Z} \prod_{i \neq j} \phi_{ij}(x_i, x_j)$ with $x_i \in \{1, \dots, L\}$
- $\binom{6}{2} = 15$ factors of size 2 $\rightarrow$ distribution specified by $15L^2$ values
- corresponding graph: fully connected



- also compatible with, *e.g.*,

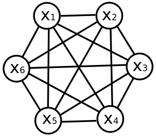
$$p(x_1, \dots, x_6) = \frac{1}{Z}\phi(x_1, x_2, x_3, x_4)\phi(x_1, x_2, x_5, x_6)\phi(x_3, x_4, x_5, x_6) \quad \rightarrow \quad 3L^4 \text{ values!}$$

- or even $p(x_1, \dots, x_6) = \frac{1}{Z}\phi(x_1, \dots, x_6) \quad \rightarrow \quad L^6$ values!

Relationship Factorizations to Graphs

Why is this a problem?

- Many problems have only small (*e.g.* pairwise) interactions, *e.g.* "friendship" in Facebook
- $p(x_1, \dots, x_6) = \frac{1}{Z} \prod_{i \neq j} \phi_{ij}(x_i, x_j)$ with $x_i \in \{1, \dots, L\}$
- $\binom{6}{2} = 15$ factors of size 2 $\rightarrow$ distribution specified by $15L^2$ values
- corresponding graph: fully connected



- also compatible with, *e.g.*,

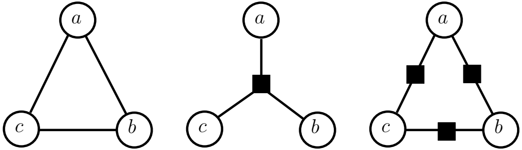
$$p(x_1, \dots, x_6) = \frac{1}{Z} \phi(x_1, x_2, x_3, x_4) \phi(x_1, x_2, x_5, x_6) \phi(x_3, x_4, x_5, x_6) \qquad \rightarrow \qquad 3L^4 \text{ values!}$$

- or even $p(x_1, \dots, x_6) = \frac{1}{Z} \phi(x_1, \dots, x_6)$ $\rightarrow$ L^6 values!

The graph alone does not tell us if the model is tractable or not. So why bother with it???

Relationship Potentials to Graphs

- We overcome his by augmenting the notation.
- We introduce an extra node (a square) for each factor in the factorization
  The square is connected to all nodes contributing to the factor.

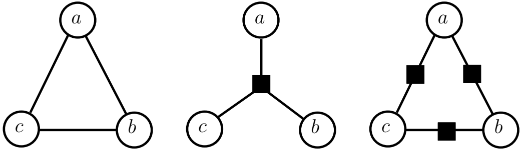


(a) (b) (c)

- (a): Markov Network graph

Relationship Potentials to Graphs

- We overcome his by augmenting the notation.
- We introduce an extra node (a square) for each factor in the factorization The square is connected to all nodes contributing to the factor.

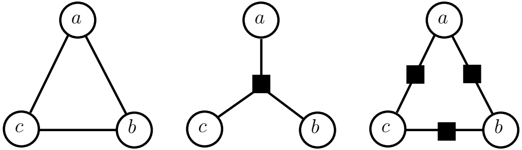


(a) (b) (c)

- (a): Markov Network graph
- (b): Factor graph representation of $p(a, b, c) \propto \phi(a, b, c)$

Relationship Potentials to Graphs

- We overcome his by augmenting the notation.
- We introduce an extra node (a square) for each factor in the factorization
  The square is connected to all nodes contributing to the factor.

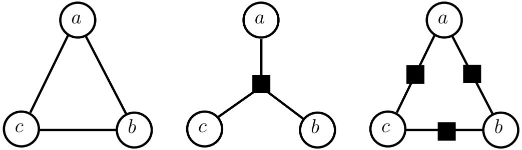


(a) (b) (c)

- (a): Markov Network graph
- (b): Factor graph representation of $p(a, b, c) \propto \phi(a, b, c)$
- (c): Factor graph representation of $p(a, b, c) \propto \phi(a, b)\phi(b, c)\phi(c, a)$

Relationship Potentials to Graphs

- We overcome his by augmenting the notation.
- We introduce an extra node (a square) for each factor in the factorization
  The square is connected to all nodes contributing to the factor.

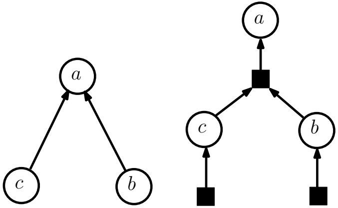


(a) (b) (c)

- (a): Markov Network graph
- (b): Factor graph representation of $p(a, b, c) \propto \phi(a, b, c)$
- (c): Factor graph representation of $p(a, b, c) \propto \phi(a, b)\phi(b, c)\phi(c, a)$
- Different factor graphs can have the same Markov network (b,c)⇒(a)

Directed Factor Graphs

- This also works for directed graph / belief network.
- The structure of the factorization is retained:



- But doesn't add much information, so typically not used.

Factor Graph Definition

Factor Graph

Given a function

$$f(x_1, \ldots, x_n) = \prod_i \psi_i(\mathcal{X}_i),$$

the factor graph (FG) has a node (represented by a square) for each factor $\psi_i(\mathcal{X}_i)$ and a variable node (represented by a circle) for each variable x_j.

Factor Graph Definition

Factor Graph

Given a function

$$f(x_1, \ldots, x_n) = \prod_i \psi_i(\mathcal{X}_i),$$

the factor graph (FG) has a node (represented by a square) for each factor $\psi_i(\mathcal{X}_i)$ and a variable node (represented by a circle) for each variable x_j. When used to represent a distribution

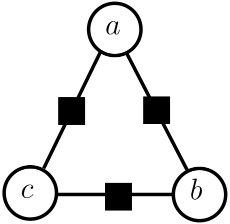
$$p(x_1, \ldots, x_n) = \frac{1}{Z} \prod_i \psi_i(\mathcal{X}_i),$$

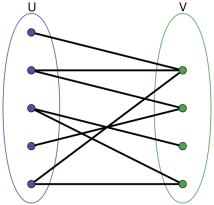
a normalization constant is assumed.

Bipartite graph

Bipartite

A bipartite graph is a graph whose vertices can be divided into two disjoint sets U and V such that every edge connects a vertex in U to one in V

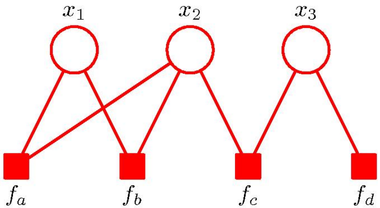




- Factor graphs are bipartite graphs. Edge are always between a *variables node* (circle) and a *factor node* (square).

Factor graph: example 1

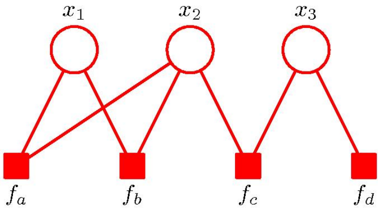
- Question: which factorization ?



- Answer:

Factor graph: example 1

- Question: which factorization ?



- Answer:

$$p(x) = \frac{1}{Z} f_a(x_1, x_2) f_b(x_1, x_2) f_c(x_2, x_3) f_d(x_3)$$

Factor graph: example 2

- Question: Which factor graph ?

$$p(x_1, x_2, x_3) = p(x_1)p(x_2)p(x_3 \mid x_1, x_2)$$

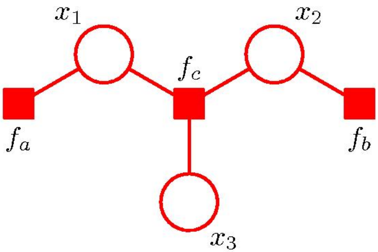
- Answer:

Factor graph: example 2

- Question: Which factor graph ?

$$p(x_1, x_2, x_3) = p(x_1)p(x_2)p(x_3 \mid x_1, x_2)$$

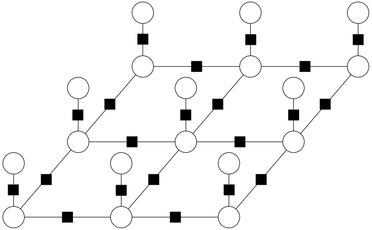
- Answer:



Example: A Factor Graph and Energy Function for Image Denoising

X: 

Y: 

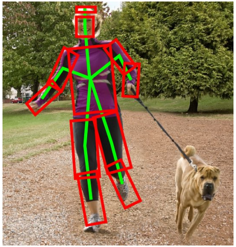


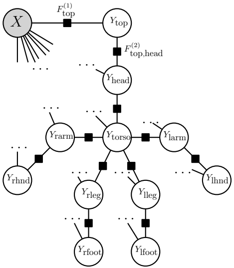
$$p(x,y) = \frac{1}{Z} e^{-E(x,y)} \qquad E(x,y) = \sum_{i \in \{\text{pixels}\}} E_i(x_i, y_i) + \sum_{(i,j) \in \{\text{edges}\}} E_{ij}(y_i, y_j)$$

Pairwise Markov Random Field (MRF):

- $E_i(x_i, y_i) = \alpha(x_i - y_i)^2$ outputs are likely similar to inputs
- $E_{ij}(y_i, y_j) = \beta|y_i - y_j|$ neighboring outputs are likely similar to each other $\rightarrow$ smooth output
- $\alpha \in \mathbb{R}$ and $\beta \in \mathbb{R}$ can be adjusted per image

Example: A Factor Graph and Energy Function for Human Pose Estimation





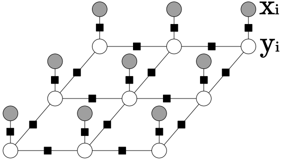
$$p(y|x) = \frac{1}{Z} e^{-E(y;x)} \qquad E(y;x) = \sum_{i \in \{\text{head,torso},\dots\}} E_i(y_i;x_i) + \sum_{(i,j)} E_{ij}(y_i, y_j)$$

- unary factors (depend on one label): appearance
  - *e.g.* $E_{\text{head}}(y;x)$ *"Does location y in image x look like a head?"*
- pairwise factors (depend on two labels): geometry
  - *e.g.* $E_{\text{head-torso}}(y_{\text{head}}, y_{\text{torso}})$ *"Is location y_{head} above location y_{torso}?"*

Example: A Factor Graph and Energy Function for Image Segmentation

X: 

Y: 



$$p(y|x) = \frac{1}{Z} e^{-E(y;x)} \qquad E(y;x) = \sum_{i \in \{\text{pixels}\}} E_i(y_i; x_i) + \sum_{(i,j) \in \{\text{edges}\}} E_{ij}(y_i, y_j)$$

Energy function components ("Ising" model):

- $E_i(y_i = 1, x_i) = \begin{cases} \text{low} & \text{if } x_i \text{ is the right color, e.g. } \texttt{brown} \\ \text{high} & \text{otherwise} \end{cases}$ $\qquad E_i(y_i = 0, x_i) = -E_i(y_i = 1, x_i)$
- $E_i(y_i, y_j) = \begin{cases} \text{low} & \text{if } y_i = y_j \\ \text{high} & \text{otherwise} \end{cases}$ $\qquad$ higher probability if neighbors have same labels $\rightarrow$ smooth labelings

Example: A Factor Graph and Energy Function for Graph Matching

$$G = (V, \mathcal{E}) \qquad X : \begin{pmatrix} x_{1a} & x_{1b} & x_{1c} & x_{1d} \\ x_{2a} & x_{2b} & x_{2c} & x_{2d} \\ x_{3a} & x_{3b} & x_{3c} & x_{3d} \\ x_{4a} & x_{4b} & x_{4c} & x_{4d} \end{pmatrix} \in \{0,1\}^{|V|\times|V'|} \qquad G' = (V', \mathcal{E}')$$

(which left node matches to which right node)

Example: A Factor Graph and Energy Function for Graph Matching

$$G = (V, \mathcal{E}) \qquad X : \begin{pmatrix} x_{1a} & x_{1b} & x_{1c} & x_{1d} \\ x_{2a} & x_{2b} & x_{2c} & x_{2d} \\ x_{3a} & x_{3b} & x_{3c} & x_{3d} \\ x_{4a} & x_{4b} & x_{4c} & x_{4d} \end{pmatrix} \in \{0,1\}^{|V| \times |V'|} \qquad G' = (V', \mathcal{E}')$$

(which left node matches to which right node)

$$p_1(x) = \frac{1}{Z} e^{-E_1(x)} \qquad E_1(x) = \begin{cases} \sum_{(i,j) \in V \times V'} x_{ij} \, |\deg(v_i) - \deg(v_j)|) & \text{if } x \text{ is a valid assignment,} \\ \infty & \text{otherwise.} \end{cases}$$

Example: A Factor Graph and Energy Function for Graph Matching

$$G = (V, \mathcal{E}) \qquad X : \begin{pmatrix} x_{1a} & x_{1b} & x_{1c} & x_{1d} \\ x_{2a} & x_{2b} & x_{2c} & x_{2d} \\ x_{3a} & x_{3b} & x_{3c} & x_{3d} \\ x_{4a} & x_{4b} & x_{4c} & x_{4d} \end{pmatrix} \in \{0,1\}^{|V|\times|V'|} \qquad G' = (V', \mathcal{E}')$$

(which left node matches to which right node)

$$p_1(x) = \frac{1}{Z} e^{-E_1(x)} \qquad E_1(x) = \begin{cases} \sum_{(i,j)\in V\times V'} x_{ij} \,|\mathrm{deg}(v_i) - \mathrm{deg}(v_j)|) & \text{if } x \text{ is a valid assignment,} \\ \infty & \text{otherwise.} \end{cases}$$

$$p_2(x) = \frac{1}{Z} e^{-\alpha E_1(x) - \beta E_2(x)} \qquad E_2(x) = \begin{cases} \sum_{(i,j,k,l)\in V\times V\times V'\times V'} x_{ik} x_{jl} \,|\mathrm{dist}(v_i, v_j) - \mathrm{dist}(v'_k, v'_l)| & \text{if } x \text{ is valid,} \\ \infty & \text{otherwise.} \end{cases}$$

Example: A Factor Graph and Energy Function for Graph Matching

$$G = (V, \mathcal{E}) \qquad X : \begin{pmatrix} x_{1a} & x_{1b} & x_{1c} & x_{1d} \\ x_{2a} & x_{2b} & x_{2c} & x_{2d} \\ x_{3a} & x_{3b} & x_{3c} & x_{3d} \\ x_{4a} & x_{4b} & x_{4c} & x_{4d} \end{pmatrix} \in \{0,1\}^{|V|\times|V'|} \qquad G' = (V', \mathcal{E}')$$

(which left node matches to which right node)

$$p_1(x) = \frac{1}{Z} e^{-E_1(x)} \qquad E_1(x) = \begin{cases} \sum_{(i,j)\in V\times V'} x_{ij} \left|\deg(v_i) - \deg(v_j)\right|) & \text{if } x \text{ is a valid assignment,} \\ \infty & \text{otherwise.} \end{cases}$$

$$p_2(x) = \frac{1}{Z} e^{-\alpha E_1(x) - \beta E_2(x)} \qquad E_2(x) = \begin{cases} \sum_{(i,j,k,l)\in V\times V\times V'\times V'} x_{ik} x_{jl} \left|\mathrm{dist}(v_i, v_j) - \mathrm{dist}(v'_k, v'_l)\right| & \text{if } x \text{ is valid,} \\ \infty & \text{otherwise.} \end{cases}$$

$$p_3(x) = \frac{1}{Z} e^{-\alpha E_1(x) - \beta E_2(x) - \gamma E_3(x)} \qquad E_3(x) = \begin{cases} \sum_{(i,j,k,r,s,t)\in V\times V\times V\times V'\times V'\times V'} x_{ir} x_{js} x_{kt} \left|\angle(v_i, v_j, v_k) - \angle(v'_r, v'_s, v'_t)\right| & \text{if } x \text{ is valid,} \\ \infty & \text{otherwise.} \end{cases}$$

Example: A Factor Graph and Energy Function for Graph Matching

$$G = (V, \mathcal{E}) \qquad X : \begin{pmatrix} x_{1a} & x_{1b} & x_{1c} & x_{1d} \\ x_{2a} & x_{2b} & x_{2c} & x_{2d} \\ x_{3a} & x_{3b} & x_{3c} & x_{3d} \\ x_{4a} & x_{4b} & x_{4c} & x_{4d} \end{pmatrix} \in \{0,1\}^{|V|\times|V'|} \qquad G' = (V', \mathcal{E}')$$

(which left node matches to which right node)

$$p_1(x) = \frac{1}{Z} e^{-E_1(x)} \qquad E_1(x) = \begin{cases} \sum\limits_{(i,j)\in V\times V'} x_{ij} \, |\mathrm{deg}(v_i) - \mathrm{deg}(v_j)|) & \text{if } x \text{ is a valid assignment,} \\ \infty & \text{otherwise.} \end{cases}$$

$$p_2(x) = \frac{1}{Z} e^{-\alpha E_1(x) - \beta E_2(x)} \qquad E_2(x) = \begin{cases} \sum\limits_{(i,j,k,l)\in V\times V\times V'\times V'} x_{ik}x_{jl} \, |\mathrm{dist}(v_i, v_j) - \mathrm{dist}(v'_k, v'_l)| & \text{if } x \text{ is valid,} \\ \infty & \text{otherwise.} \end{cases}$$

$$p_3(x) = \frac{1}{Z} e^{-\alpha E_1(x) - \beta E_2(x) - \gamma E_3(x)} \qquad E_3(x) = \begin{cases} \sum\limits_{(i,j,k,r,s,t)\in V\times V\times V\times V'\times V'\times V'} x_{ir}x_{js}x_{kt} \, |\angle(v_i, v_j, v_k) - \angle(v'_r, v'_s, v'_t)| & \text{if } x \text{ is valid,} \\ \infty & \text{otherwise.} \end{cases}$$

Assign higher probability if similarity or geometry matches well.

Summary (so far)

The graphs of graphical models represent families of probability distributions graphically:

- Bayesian networks: directed acyclic graphs, product of conditional distribution
  - by default, arrows have no causal interpretation
  - but: causal Bayesian networks also exist

Summary (so far)

The graphs of graphical models represent families of probability distributions graphically:

- Bayesian networks: directed acyclic graphs, product of conditional distribution
  - by default, arrows have no causal interpretation
  - but: causal Bayesian networks also exist
- Markov networks: undirected, local cliques of dependent variables

Summary (so far)

The graphs of graphical models represent families of probability distributions graphically:

- Bayesian networks: directed acyclic graphs, product of conditional distribution
  - by default, arrows have no causal interpretation
  - but: causal Bayesian networks also exist
- Markov networks: undirected, local cliques of dependent variables
- Factor graphs
  - makes the factorization explicit
  - not a larger class of distributions, "just" a different way of drawing the graph

Summary (so far)

The graphs of graphical models represent families of probability distributions graphically:

- Bayesian networks: directed acyclic graphs, product of conditional distribution
  - by default, arrows have no causal interpretation
  - but: causal Bayesian networks also exist
- Markov networks: undirected, local cliques of dependent variables
- Factor graphs
  - makes the factorization explicit
  - not a larger class of distributions, “just” a different way of drawing the graph
- for modeling undirected models, thinking in terms of factor graphs is very useful
  - very often only a few factor 'types', evaluated on different subsets of variables

Summary (so far)

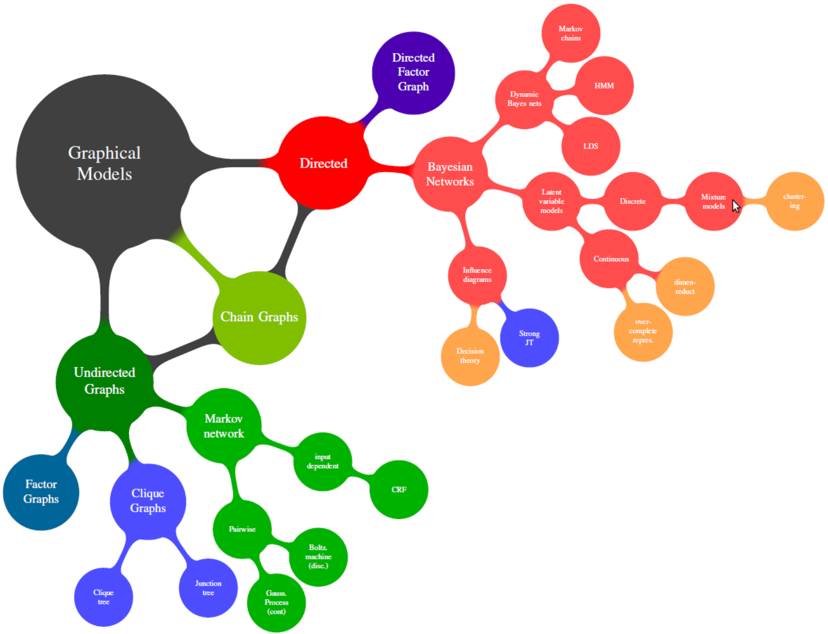
The graphs of graphical models represent families of probability distributions graphically:

- Bayesian networks: directed acyclic graphs, product of conditional distribution
  - by default, arrows have no causal interpretation
  - but: causal Bayesian networks also exist
- Markov networks: undirected, local cliques of dependent variables
- Factor graphs
  - makes the factorization explicit
  - not a larger class of distributions, "just" a different way of drawing the graph
- for modeling undirected models, thinking in terms of factor graphs is very useful
  - very often only a few factor 'types', evaluated on different subsets of variables

To specify an actual distribution, we also have to provide:

- for directed models: the conditional tables
- for undirected models: the potentials

Often, these are learned from training data (while the graph structure is fixed manually).


Graphical Models
Directed
Directed Factor Graph
Bayesian Networks
Dynamic Bayes nets
Markov chains
HMM
LDS
Latent variable models
Discrete
Mixture models
cluster-ing
Continuous
dimen-reduct
over-complete repres.
Influence diagrams
Decision theory
Strong JT
Chain Graphs
Undirected Graphs
Markov network
input dependent
CRF
Pairwise
Boltz. machine (disc.)
Gauss. Process (cont)
Factor Graphs
Clique Graphs
Clique tree
Junction tree

Learning from data

For many processes, we are not given the probabilities/factor values, but we observe data: $\mathcal{D} = \{x_1, \ldots, x_n\}$.

Probability Estimation

For a given model class, probability estimation is the task of identifying the probability distribution from observed data.

General assumption:

- training data is sampled independently from a distribution of interest (i.i.d.)

Learning from data

Example: coin toss

You repeatedly flip a coin. $x_i \in \{\texttt{head}, \texttt{tail}\}$ is the output of the i-th repeat. What are the coin's probabilities $p(\texttt{head})$ and $p(\texttt{tail})$?

Standard method:

- we write $\theta_{\texttt{head}} = p(\texttt{head})$ and $\theta_{\texttt{tail}} = p(\texttt{tail})$ (one is enough: $\theta_{\texttt{tail}} = 1 - \theta_{\texttt{head}}$)

Learning from data

Example: coin toss

You repeatedly flip a coin. $x_i \in \{\mathtt{head}, \mathtt{tail}\}$ is the output of the i-th repeat. What are the coin's probabilities $p(\mathtt{head})$ and $p(\mathtt{tail})$?

Standard method:

- we write $\theta_{\mathtt{head}} = p(\mathtt{head})$ and $\theta_{\mathtt{tail}} = p(\mathtt{tail})$ (one is enough: $\theta_{\mathtt{tail}} = 1 - \theta_{\mathtt{head}}$)
- we estimate a value for $\theta_{\mathtt{head}}$ from the data as

$$\hat{\theta}_{\mathtt{head}} = \frac{\text{number of } \mathtt{head} \text{ in the observations}}{\text{total number of observations}} = \frac{1}{n}\sum_{i=1}^{n} [\![x_i = \mathtt{head}]\!]$$

where $[\![\cdot]\!]$ are Iverson brackets: $[\![\mathrm{P}]\!] = \begin{cases} 1 & \text{if condition } P \text{ is true,} \\ 0 & \text{otherwise.} \end{cases}$

Learning from data

Example: coin toss

You repeatedly flip a coin. $x_i \in \{\texttt{head}, \texttt{tail}\}$ is the output of the i-th repeat. What are the coin's probabilities $p(\texttt{head})$ and $p(\texttt{tail})$?

Standard method:

- we write $\theta_{\texttt{head}} = p(\texttt{head})$ and $\theta_{\texttt{tail}} = p(\texttt{tail})$ (one is enough: $\theta_{\texttt{tail}} = 1 - \theta_{\texttt{head}}$)
- we estimate a value for $\theta_{\texttt{head}}$ from the data as

$$\hat{\theta}_{\texttt{head}} = \frac{\text{number of } \texttt{head} \text{ in the observations}}{\text{total number of observations}} = \frac{1}{n}\sum_{i=1}^{n} [\![x_i = \texttt{head}]\!]$$

where $[\![\cdot]\!]$ are Iverson brackets: $[\![\mathrm{P}]\!] = \begin{cases} 1 & \text{if condition } P \text{ is true,} \\ 0 & \text{otherwise.} \end{cases}$

- note: the $\hat{}$ of $\hat{\theta}_{\texttt{head}}$ indicates that this is an estimate based on data

Learning from data

Example: Gaussians

You know that a random variable has Gaussian distribution,

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

What are μ and σ?

Standard method: given i.i.d. samples: $x_1, \ldots, x_n$

- $\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i$
- $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})^2$

Learning from data

Example: Gaussians

You know that a set of d random variables $x = (x^1, \ldots, x^d)$ have a jointly Gaussian distribution,

$$p(x) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} e^{-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)}.$$

What are μ and Σ?

Standard method: given i.i.d. samples: $x_1, \ldots, x_n$

- $\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i$
- $\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})(x_i - \hat{\mu})^\top$

Maximum Likelihood Estimation

Assume a parametric model: $p(x) = p(x;\theta)$, try to find value of θ:

- Gaussian: $p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, parametrized by $\theta = (\mu, \sigma^2)$

Maximum Likelihood Estimation

Assume a parametric model: $p(x) = p(x; \theta)$, try to find value of θ:

- Gaussian: $p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, parametrized by $\theta = (\mu, \sigma^2)$

What about discrete probability tables? We make each entry a parameter:

- coin toss: $p(\texttt{head}) = \theta_{\texttt{head}}$, $p(\texttt{tail}) = \theta_{\texttt{tail}}$ has parameters $\theta = (\theta_{\texttt{head}}, \theta_{\texttt{tail}})$.

Maximum Likelihood Estimation

Assume a parametric model: $p(x) = p(x;\theta)$, try to find value of θ:

- Gaussian: $p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, parametrized by $\theta = (\mu, \sigma^2)$

What about discrete probability tables? We make each entry a parameter:

- coin toss: $p(\texttt{head}) = \theta_{\texttt{head}}$, $p(\texttt{tail}) = \theta_{\texttt{tail}}$ has parameters $\theta = (\theta_{\texttt{head}}, \theta_{\texttt{tail}})$.

Maximum Likelihood Estimation

Given a parametric model $p(x;\theta)$ and data, $\mathcal{D} = \{x_1, \ldots, x_n\}$, estimate parameters as

$$\hat{\theta}_{ML} = \operatorname*{argmax}_{\theta} \mathcal{L}(\theta) \quad \text{for} \quad \mathcal{L}(\theta) = p(\mathcal{D};\theta) = \prod_{i=1}^{n} p(x_i;\theta) \quad \text{data likelihood}$$

i.e. the parameter value that makes the observed data most likely.

Maximum Likelihood Estimation

The Maximum Likelihood Estimator

$$\hat{\theta}_{ML} = \underset{\theta}{\operatorname{argmax}} \, \mathcal{L}(\theta) \quad \text{for} \quad \mathcal{L}(\theta) = \prod_{i=1}^{n} p(x_i; \theta)$$

is equivalent to the maximizer of the log-likelihood,

$$L(\theta) := \log \mathcal{L}(\theta) = \sum_{i=1}^{n} \log p(x_i; \theta)$$

or the minimizer of its negative.

$$\underset{\theta}{\operatorname{argmax}} \, \mathcal{L}(\theta) = \underset{\theta}{\operatorname{argmax}} \log \mathcal{L}(\theta) = \underset{\theta}{\operatorname{argmax}} \, L(\theta) = \underset{\theta}{\operatorname{argmin}} [-L(\theta)]$$

(mathematically equivalent, but often easier expressions and numerically more stable)

Maximum Likelihood Estimation

Maximum likelihood estimator for Gaussian data

- Assume: Gaussian distribution with $\theta = (\mu, \sigma^2)$

$$p(x_i; \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \qquad L(\theta) = -\sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2} - n \log \sqrt{2\pi\sigma^2}$$

- smooth convex function of θ. Find minimum, $\hat{\theta}$, by setting derivative to 0:

$$0 = \frac{dL}{d\mu}(\hat{\theta}) = \frac{1}{\hat{\sigma}^2}\sum_{i=1}^{n}(x_i - \hat{\mu}) \quad \Rightarrow \quad n\hat{\mu} = \sum_{i=1}^{n} x_i \quad \Rightarrow \quad \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$0 = \frac{dL}{d\sigma^2} = \frac{1}{2\hat{\sigma}^4}\sum_{i=1}^{n}(x_i - \hat{\mu})^2 - \frac{n}{2\hat{\sigma}^2} \quad \Rightarrow \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2$$

Maximum likelihood estimate is standard solution. Also works for vectors (a bit more work)

Maximum likelihood estimator for coin toss

- Coin toss: use $x = 1$ for `head` and $x = 0$ for `tail`
- Assume: true distribution $p(x; \theta_0, \theta_1)$ with $p(X = 1) = \theta_1$, $p(X = 0) = \theta_0$.

$$p(x_i; \theta_0, \theta_1) = \theta_1^{x_i} (\theta_0)^{1-x_i} \qquad \text{with convention } 0^0 = 1$$

$$\begin{aligned} L(\theta_0, \theta_1) &= \sum_{i=1}^{n} \log[\theta_1^{x_i} \theta_0^{1-x_i}] \; = \; \sum_{i=1}^{n} [x_i \log \theta_1 + (1 - x_i) \log \theta_0] \\ &= \log \theta_1 \sum_{i=1}^{n} x_i + \log \theta_0 \sum_{i=1}^{n} (1 - x_i) \end{aligned}$$

Maximum likelihood estimator for coin toss

- Coin toss: use $x = 1$ for `head` and $x = 0$ for `tail`
- Assume: true distribution $p(x; \theta_0, \theta_1)$ with $p(X = 1) = \theta_1$, $p(X = 0) = \theta_0$.

$$p(x_i; \theta_0, \theta_1) = \theta_1^{x_i} (\theta_0)^{1-x_i} \qquad \text{with convention } 0^0 = 1$$

$$L(\theta_0, \theta_1) = \sum_{i=1}^{n} \log[\theta_1^{x_i} \theta_0^{1-x_i}] = \sum_{i=1}^{n} [x_i \log \theta_1 + (1 - x_i) \log \theta_0]$$

$$= \log \theta_1 \sum_{i=1}^{n} x_i + \log \theta_0 \sum_{i=1}^{n} (1 - x_i)$$

- monotonically increasing function of $\theta_0 \in [0, 1]$ and $\theta_1 \in [0, 1]$
- maximum at $\theta_0 = 1$ and $\theta_1 = 1$ $\quad \rightarrow \quad$ what went wrong?

Maximum likelihood estimator for coin toss

Minimize with side condition $\theta_0 + \theta_1 = 1 \to$ use a Lagrangian multiplier!

$$\max_{\theta_0,\theta_1 \in [0,1]} L(\theta_0, \theta_1), \qquad \text{subject to } \theta_0 + \theta_1 = 1$$

Lagrange conditions: solution $(\hat\theta_0, \hat\theta_1, \hat\lambda)$ is critical point of the Lagrangian:

$$\mathfrak{L}(\theta_0, \theta_1, \lambda) = L(\theta_0, \theta_1) - \lambda(\theta_0 + \theta_1 - 1)$$

$$0 = \frac{d\mathfrak{L}}{d\theta_1}(\hat\theta_0, \hat\theta_1, \hat\lambda) = \frac{1}{\hat\theta_1}\sum_{i=1}^n x_i - \lambda \qquad \to \qquad \hat\theta_1 = \frac{1}{\hat\lambda}\sum_{i=1}^n x_i$$

$$0 = \frac{d\mathfrak{L}}{d\theta_0}(\hat\theta_0, \hat\theta_1, \hat\lambda) = \frac{1}{\hat\theta_0}\sum_{i=1}^n (1 - x_i) - \lambda \qquad \to \qquad \hat\theta_0 = \frac{1}{\hat\lambda}\sum_{i=1}^n (1 - x_i)$$

$$0 = \frac{d\mathfrak{L}}{d\lambda}(\hat\theta_0, \hat\theta_1, \hat\lambda) = 1 - \frac{1}{\hat\lambda}\sum_{i=1}^n x_i - \frac{1}{\hat\lambda}\sum_{i=1}^n (1 - x_i) \qquad \to \qquad \hat\lambda = n$$

MLE for coin toss is the same as the usual (textbook) estimates: $\theta_k = \frac{1}{n}\sum_{i=1}^n [\![x_i = k]\!]$

Maximum Likelihood Estimation: Alternative Explanation

Reminder:

Kullback-Leibler divergence

Measure of (dis)similarity between probability distributions

- discrete:

$$D_{KL}(q\|p) = \sum_{x \in \mathcal{X}} q(x) \log \frac{q(x)}{p(x)}$$

- continuous:

$$D_{KL}(q\|p) = \int_{-\infty}^{\infty} q(x) \log \frac{q(x)}{p(x)}$$

Not a "distance": not symmetric, no triangular inequality

Let q be the empirical data distribution: $q(x) = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}(x)$ for $\delta_{x_i}(x) = [\![x = x_i]\!]$.

$$\underset{\theta}{\operatorname{argmin}}\, D_{KL}(q(x) \| p(x;\theta))$$

Let q be the empirical data distribution: $q(x) = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}(x)$ for $\delta_{x_i}(x) = [\![x = x_i]\!]$.

$$\operatorname*{argmin}_{\theta} D_{KL}(q(x) \| p(x;\theta)) = \operatorname*{argmin}_{\theta} \sum_{x} q(x) \log \frac{q(x)}{p(x;\theta)}$$

Let q be the empirical data distribution: $q(x) = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}(x)$ for $\delta_{x_i}(x) = [\![x = x_i]\!]$.

$$\begin{aligned}
\operatorname*{argmin}_{\theta} D_{KL}(q(x) \| p(x;\theta)) &= \operatorname*{argmin}_{\theta} \sum_{x} q(x) \log \frac{q(x)}{p(x;\theta)} \\
&= \operatorname*{argmin}_{\theta} \sum_{x} q(x) \log q(x) - \sum_{x} q(x) \log p(x;\theta)
\end{aligned}$$

Let q be the empirical data distribution: $q(x) = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}(x)$ for $\delta_{x_i}(x) = [\![x = x_i]\!]$.

$$
\begin{aligned}
\operatorname*{argmin}_{\theta} D_{KL}(q(x) \| p(x;\theta)) &= \operatorname*{argmin}_{\theta} \sum_{x} q(x) \log \frac{q(x)}{p(x;\theta)} \\
&= \operatorname*{argmin}_{\theta} \sum_{x} q(x) \log q(x) - \sum_{x} q(x) \log p(x;\theta) \\
&= \operatorname*{argmax}_{\theta} \sum_{x} q(x) \log p(x;\theta)
\end{aligned}
$$

Let q be the empirical data distribution: $q(x) = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}(x)$ for $\delta_{x_i}(x) = [\![x = x_i]\!]$.

$$\begin{aligned}
\operatorname*{argmin}_{\theta} D_{KL}(q(x) \| p(x;\theta)) &= \operatorname*{argmin}_{\theta} \sum_{x} q(x) \log \frac{q(x)}{p(x;\theta)} \\
&= \operatorname*{argmin}_{\theta} \sum_{x} q(x) \log q(x) - \sum_{x} q(x) \log p(x;\theta) \\
&= \operatorname*{argmax}_{\theta} \sum_{x} q(x) \log p(x;\theta) \\
&= \operatorname*{argmax}_{\theta} \frac{1}{n} \sum_{i=1}^{n} \log p(x_i;\theta)
\end{aligned}$$

Let q be the empirical data distribution: $q(x) = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}(x)$ for $\delta_{x_i}(x) = [\![x = x_i]\!]$.

$$
\begin{aligned}
\operatorname*{argmin}_\theta D_{KL}(q(x) \| p(x;\theta)) &= \operatorname*{argmin}_\theta \sum_x q(x) \log \frac{q(x)}{p(x;\theta)} \\
&= \operatorname*{argmin}_\theta \sum_x q(x) \log q(x) - \sum_x q(x) \log p(x;\theta) \\
&= \operatorname*{argmax}_\theta \sum_x q(x) \log p(x;\theta) \\
&= \operatorname*{argmax}_\theta \frac{1}{n} \sum_{i=1}^{n} \log p(x_i;\theta) \\
&= \operatorname*{argmax}_\theta L(\theta)
\end{aligned}
$$

Let q be the empirical data distribution: $q(x) = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}(x)$ for $\delta_{x_i}(x) = [\![x = x_i]\!]$.

$$
\begin{aligned}
\underset{\theta}{\operatorname{argmin}}\, D_{KL}(q(x) \| p(x;\theta)) &= \underset{\theta}{\operatorname{argmin}} \sum_x q(x) \log \frac{q(x)}{p(x;\theta)} \\
&= \underset{\theta}{\operatorname{argmin}} \sum_x q(x) \log q(x) - \sum_x q(x) \log p(x;\theta) \\
&= \underset{\theta}{\operatorname{argmax}} \sum_x q(x) \log p(x;\theta) \\
&= \underset{\theta}{\operatorname{argmax}} \frac{1}{n} \sum_{i=1}^{n} \log p(x_i;\theta) \\
&= \underset{\theta}{\operatorname{argmax}}\, L(\theta) \\
&= \underset{\theta}{\operatorname{argmax}}\, \mathcal{L}(\theta)
\end{aligned}
$$

Maximum likelihood is equivalent to finding the parameter that minimizes the KL-divergence between the model distribution and the empirical data distribution.

Estimating an unknown value from data

Maximum likelihood is one example of how to estimate an unknown value from data.
We'll see other estimators (MAP, Pseudolikelihood, ...) later.

Estimators

An estimator is a rule for calculating an estimate, $\hat{E}(S)$, of a quantity E based on observed data, S. If S is random, then $\hat{E}(S)$ is also random.

Properties of estimators: unbiasedness

We can compute the expected value of the estimate, $\mathbb{E}_S[\hat{E}(S)]$.

- if $\mathbb{E}_S[\hat{E}(S)] = E$, we call the estimator unbiased. Think of $\hat{E}$ as a noisy version of E.
- $\text{bias}(\hat{E}) = \mathbb{E}_S[\hat{E}(S)] - E$

Estimating an unknown value from data

Maximum likelihood is one example of how to estimate an unknown value from data.
We'll see other estimators (Bayesian, Pseudolikelihood, ...) later.

Estimators

An estimator is a rule for calculating an estimate, $\hat{E}(S)$, of a quantity E based on observed data, S. If S is random, then $\hat{E}(S)$ is also random.

Properties of estimators: variance

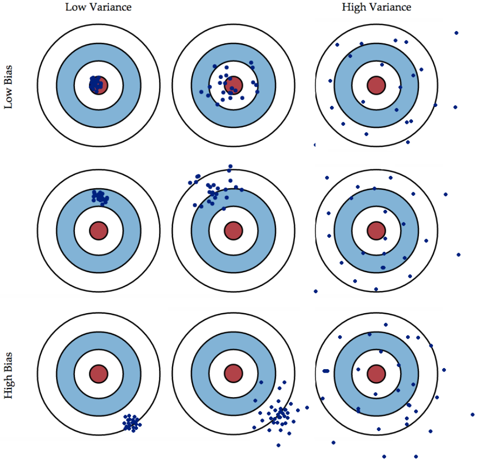
How far is one estimate from the expected value? $(\hat{E}(S) - \mathbb{E}_S[\hat{E}(S))^2$

- $\mathrm{Var}(\hat{E}) = \mathbb{E}_S[\big(\hat{E}(S) - \mathbb{E}_S[\hat{E}(S)]\big)^2]$

If $\mathrm{Var}(\hat{E})$ is large, then the estimate fluctuates a lot for different S.

Bias-Variance Trade-Off

It's good to have small or no bias, and it's good to have small variance.



If you can't have both at the same time, look for a reasonable trade-off.

Image: adapted from http://scott.fortmann-roe.com/docs/BiasVariance.html

Estimating an unknown value from data

For data sets of increasing size, $S_1, S_2, \ldots$, we can look a the behavior of the estimates $\hat{E}(S_1), \hat{E}(S_2), \ldots$. It would be nice if they converged to the true value, E.

Properties of estimators: consistency

We call an estimator $\hat{E}$ a consistent estimator of a value E if

$$\Pr\{\lim_{n\to\infty} \|E(S_n) - E| > \epsilon\} = 0$$

("$E(S_n)$ converges to E in probability")

Any unbiased estimator is consistent if its variance that converges to 0 as the size of S grows to infinity. For example: MLE of coin toss, MLE of Gaussian.

Consistency of Maximum Likelihood

Assume that the observed data comes from a distribution that is in the model class (and some weak technical conditions are fulfilled).

Consistency of Maximum Likelihood

Assume that the observed data comes from a distribution that is in the model class (and some weak technical conditions are fulfilled).

- Maximum likelihood is a consistent estimator.
  $\rightarrow$ in the limit of infinite data, the parameter estimate will converge to the true value.

Consistency of Maximum Likelihood

Assume that the observed data comes from a distribution that is in the model class (and some weak technical conditions are fulfilled).

- Maximum likelihood is a consistent estimator.
  $\rightarrow$ in the limit of infinite data, the parameter estimate will converge to the true value.

What if the observed data does not come from a distribution in the model class?

- Maximum likelihood is not consistent (there might not even be a 'correct' parameter).
- It might not converge to the 'best possible' parameter, either.