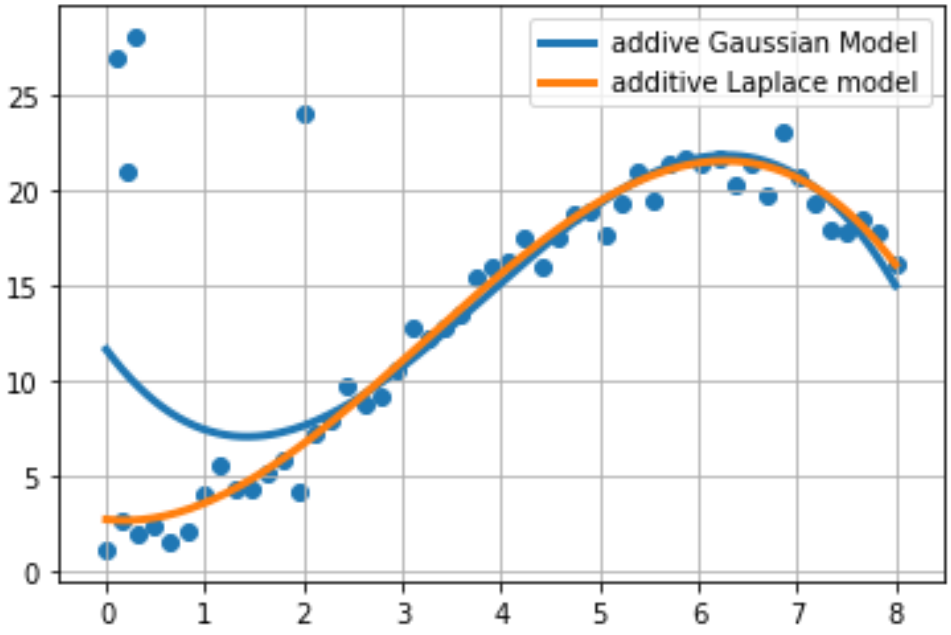
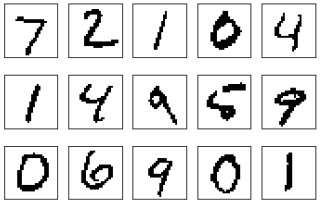
Probabilistic Graphical Models

- independence
- conditional independence
- graphs
- graphical model (DAGs)


addive Gaussian Model
additive Laplace model
25
20
15
10
5
0
0
1
2
3
4
5
6
7
8

Big Picture



$\widehat{y} = \arg\max_y p(\boldsymbol{x}|y)p(y)$ posterior $\propto$ likelihood $\times$ prior

$\widehat{y} = \arg\max_y p(\boldsymbol{x}|y)$ ML (maximum likelihood) estimate

- still need to estimate $p(\boldsymbol{x}|y)$ from data.
  - directly
    - 2^{784} possible values for $\boldsymbol{x}$
  - naïve Bayes: features $x_1, x_2 \ldots$ are independent.

$$p(\boldsymbol{x}|y) \approx \prod_{j=1}^{n} p(x_j|y)$$

    - 784 Bernoulli paramaters for each class

- *conditional independence* – compromise between Naive Bayes and full density estimate.

Factoring Distributions

- from the definition of conditional:

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

$$p(x_2, x_3) = p(x_2)p(x_3|x_2)$$

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

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

- in general:

$$p(\boldsymbol{x}) = p(x_1)p(x_2|x_1)p(x_3|x_1, x_2)p(x_4|x_1, x_2, x_3)\ldots p(x_n|x_1, \ldots, x_{n-1})$$

$$p(\boldsymbol{x}) = \prod_{j=1}^{n} p(x_j|x_1, x_2, \ldots x_{j-1})$$

- how many paramaters if $x_i \in \{0, 1\}$?

- is this decomposition unique?

Independence

- random variables X_1 and X_2 are independent if and only if

$$p(x_1, x_2) = p(x_1)p(x_2) \text{ for all } x_1, x_2$$

or equivalently

$$p(x_1|x_2) = p(x_1) \text{ for all } x_1, x_2$$

- definition of conditional:

$$p(x_1|x_2) = \frac{p(x_1, x_2)}{p(x_2)}$$

- notation

$$X_1 \perp X_2 \qquad X_1 \not\perp X_2$$

- $X_1, X_2, \ldots, X_n$ are independent if and only if

$$p(\boldsymbol{x}) = \prod_{i=1}^{n} p(x_i) \text{ for all } \boldsymbol{x} \in \mathcal{X}^n$$

Conditional Independence

- X_1, X_2 are *conditionally independent* given X_3 if and only if

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

which implies

$$p(x_1|x_2, x_3) = p(x_1|x_3) \text{ for all } x_1, x_2, x_3$$

$$p(x_2|x_1, x_3) = p(x_2|x_3) \text{ for all } x_1, x_2, x_3$$

- notation

$$X_1 \perp X_2|X_3 \qquad\qquad X_1 \not\perp X_2|X_3$$

- X_1, X_2 are *conditionally independent* given X_3, X_4, X_5 ...

$$p(x_1, x_2|x_3, x_4, x_5) = p(x_1|x_3, x_4, x_5)p(x_2|x_3, x_4, x_5)$$

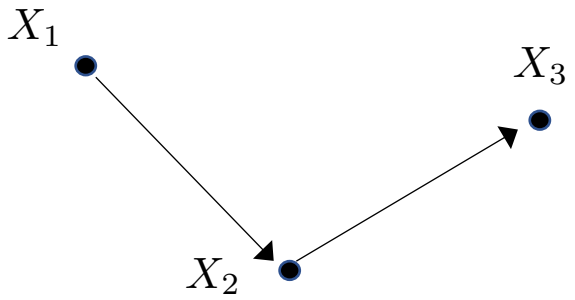
$$X_1 \perp X_2|X_3, X_4, X_5$$

Example 1

1) flip one coin X_1.
2) if $X_1 = 1$, flip coin X_2, else set $X_2 = 0$.
3) set $X_3 = X_2$

X_1	X_2	X_3	$p(x_1, x_2, x_3)$

- are X_1, X_2 independent?
- are X_1, X_3 independent?
- are X_1, X_2, X_3 independent?
- are X_1, X_3 independent given X_2?

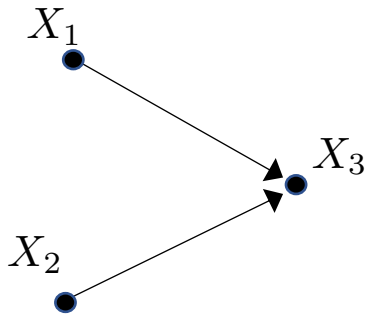


Example 2

X_1	X_2	X_3	$p(x_1, x_2, x_3)$

1) flip two fair coins, X_1 and X_2.
2) let $X_3 = \text{XOR}(X_1, X_2)$

- are X_1, X_2 independent?
- are X_1, X_3 independent?
- are X_2, X_3 independent?
- are X_1, X_2, X_3 independent?
- are X_1, X_2 independent given X_3?



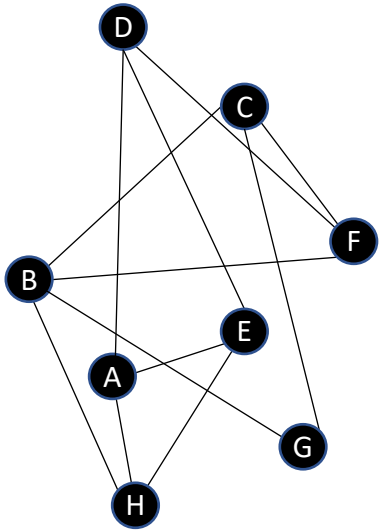
Segue - Graphs (in general)

- a graph (or network) is a set of nodes, and edges that connect some of those nodes

$$G := (V, E)$$

- V is the set of nodes (or vertices)
- an edge $e \in E$ is a two-element subset of V
- edge-list format:

```
A, E
B, C
H, E
B, F
B, G
B, H
C, F
C, G
D, E
D, F
E, H
```

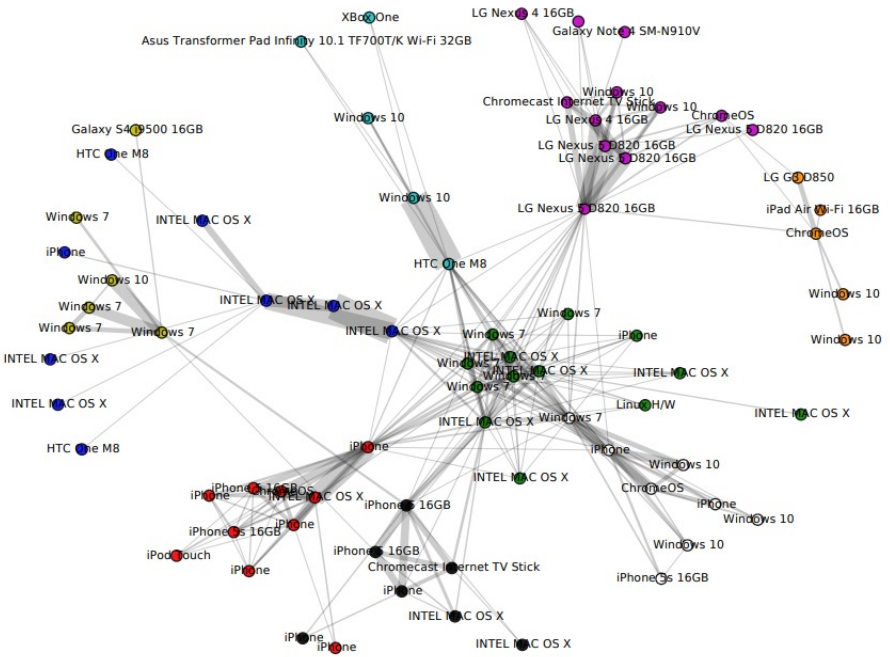


- edges might be weighted
- edges might be directed

Examples of graphs/networks

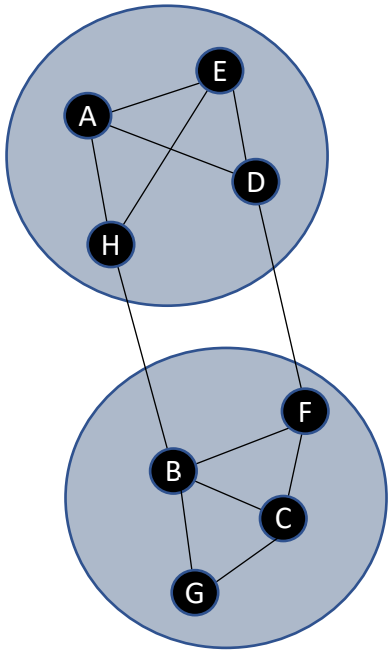
- internet device graphs
- biological networks
- epidemiological contact graphs
- social networks
- computation graphs
- visualizations of neural networks
- links between webpages
- *probabilistic graphical models*

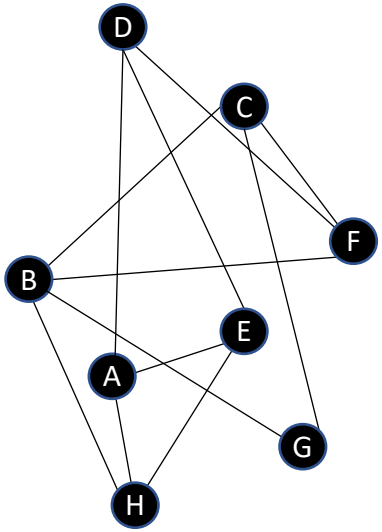
  (an illustration of the conditional independence structure between random variables)



What do people do with graphs in ML?

- community detection





Adjacency matrix:

	A	B	C	D	E	F	G	H
A	1	0	0	0	1	0		1
B	0	1	1	0	0	1	0	1
C	0	1	1	0	0	0	1	0
D	0	0	0	1	1	1	0	0
E	1	0	0	1	1	0	0	1
F	0	1	0	1	0	1	0	0
G	0	1	1	0	0	0	1	0
H	1	1	0	0	1	0	0	1

	A	E	H	D	B	F	G	C
A	1	1	1	1	0	0	0	0
E	1	1	1	0	0	0	0	0
H	1	1	1	0	1	0	0	0
D	1	0	0	1	0	1	0	0
B	0	0	1	0	1	1	1	1
F	0	0	0	1	1	1	0	1
G	0	0	0	0	1	0	1	1
C	0	0	0	0	1	1	1	1

- find graph embeddings (replace nodes with a vector in euclidean space)

- study structure, talk about how big they are

Probabilistic graphical models

- *probabilistic graphical models:*

  represent the conditional independence structure between joint random variables

  - Directed Acyclic Graphical Models (DAGs) (i.e, bayesian networks, bayes nets, belief nets, causal nets)
    - *directed* each edge has a direction
    - *acyclic* means no cycles (you can't get back to a node)

- relating $\boldsymbol{x} \sim p(\boldsymbol{x})$ to $G = (V, E)$

  $p(\boldsymbol{x})$ factorizes over G if ...

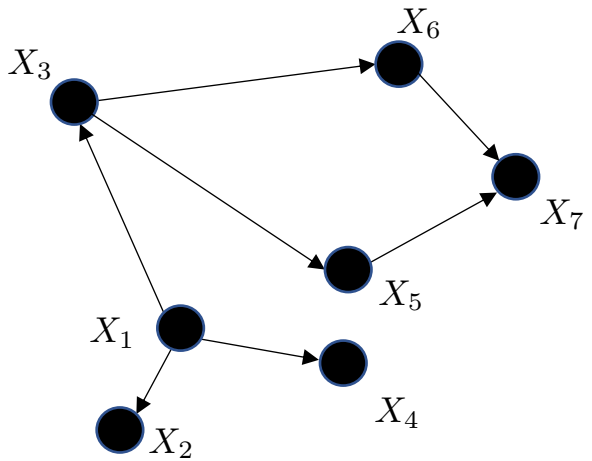
  - the nodes represent the random variables $X_1, X_2, \ldots$

    $$V = \{X_1, X_2, \ldots, X_n\}$$

  - edges represent dependencies between $X_1, X_2, \ldots$

    $$p(\boldsymbol{x}) = \prod_j p(x_j | \text{parents of } x_j)$$

  - key property: number nodes so that parents come before children



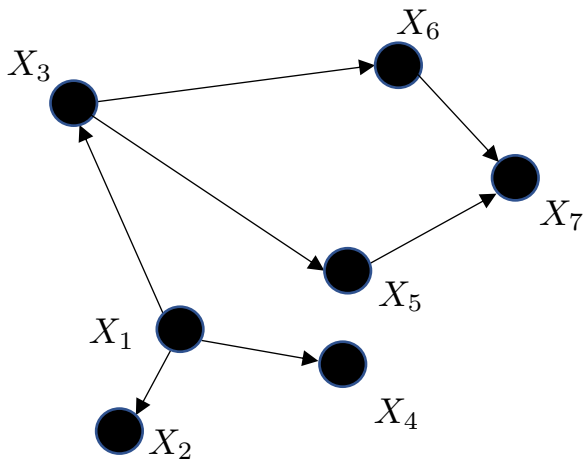
Probabilistic graphical models

- relating $\boldsymbol{x} \sim p(\boldsymbol{x})$ to $G = (V, E)$
  - edges represent dependencies between $X_1, X_2, \ldots$

$X_i \perp X_j | \{\text{parents of } X_i\}$ for all $j < i$ (that aren't parents)

X_i is independent of X_j for $j < i$ conditioned on parents of X_i

  - example: X_7.



$x_j \perp \boldsymbol{x}_{\text{predecessor}(j) \setminus \text{parents}(j)} | \boldsymbol{x}_{\text{parents}}$

$X_7 \perp X_3 | X_5, X_6$

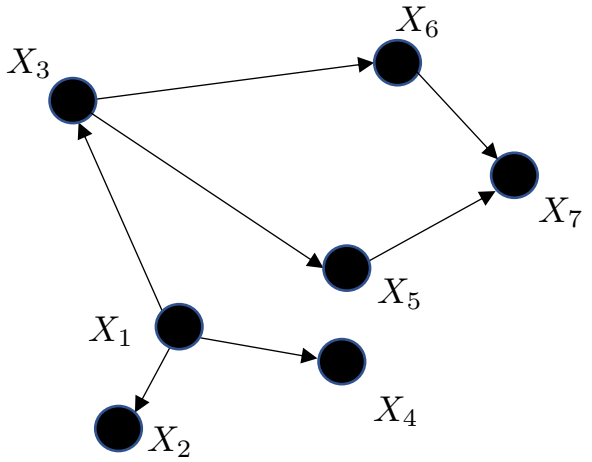
Probabilistic graphical models

- edges represent dependencies between $X_1, X_2, \ldots$

$X_i \perp X_j | \{\text{parents of } X_i\}$ for all $j < i$

$X_7 \perp X_3 | X_5, X_6$

$p(x_7|x_1, x_2, x_3, x_4, x_5, x_6) = p(x_7|x_5, x_6)$

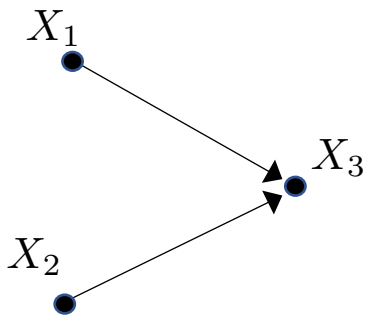


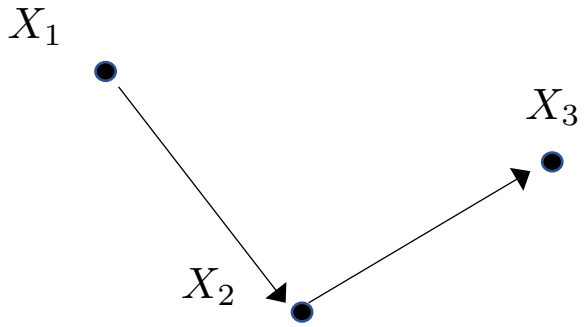
$$p(\boldsymbol{x}) = \prod_{j=1}^{n} p(x_j|x_1, x_2, \ldots x_{j-1})$$

$$p(\boldsymbol{x}) = \prod_{j} p(x_j|\text{parents of } x_j)$$

$$p(\boldsymbol{x}) = p(x_1)p(x_2|x_1)p(x_3|x_1, x_2)p(x_4|x_1, x_2, x_3)p(x_5|x_1, x_2, x_3, x_4)p(x_6|x_1, x_2, x_3, x_4, x_5)p(x_7|x_1, x_2, x_3, x_4, x_5, x_6)$$

Examples





$$p(\boldsymbol{x}) = \prod_j p(x_j|\text{parents of } x_j)$$

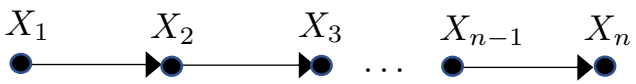
$$p(x_1, x_2, x_3) = p(x_1)p(x_2|x_1)p(x_3|x_1, x_2)$$

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

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

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

Example: Markov Chain



$$p(\boldsymbol{x}) = \prod_j p(x_j|\text{parents of } x_j)$$

$$p(\boldsymbol{x}) = p(x_1)p(x_2|x_1)p(x_3|x_1,x_2)p(x_4|x_1,x_2,x_3)\ldots p(x_n|x_1,\ldots,x_{n-1})$$

$$p(\boldsymbol{x}) = p(x_1)p(x_2|x_1)p(x_3|x_2)p(x_4|x_3)\ldots p(x_n|x_{n-1})$$

Example: Naïve Bayes

- naive assumption:

$$p(\boldsymbol{x}) = \prod_i p(x_i)$$

- DAG:

$$p(\boldsymbol{x}) = \prod_i p(x_i|\text{parents of } x_i)$$

$$p(\boldsymbol{x}) = p(x_1)p(x_2|x_1)p(x_3|x_1,x_2)p(x_4|x_1,x_2,x_3)\ldots p(x_n|x_1,\ldots,x_{n-1})$$