Probabilistic Graphical Models

Carlos Carvalho, Mladen Kolar and Robert McCulloch

11/19/2015

Classification revisited

The goal of classification is to learn a mapping from features to the target class.

- Classifier: $f : X \mapsto Y$
- X are features (Booth_Student, Taken_ML_Class, ...)
- Y is the target class (Big_Salary: yes, no)

Suppose that you know $P(Y \mid X)$ exactly, how should you classify?

- $X = (\text{Booth\_Student} = 1, \text{Taken\_ML\_Class} = 1, \ldots)$
- How do we predict $\hat{y} = \text{Big\_Salary} \in \{\text{yes}, \text{no}\}$?

Bayes optimal classifier

$$\hat{y} = \arg\max_y P(Y = y \mid X = x)$$

In practice, we do not know $P(Y = y \mid X = x)$.

We model it directly.

Logistic regression: $P(Y \mid X) = \frac{\exp(x^T b)}{1+\exp(x^T b)}$

Similarly, we can use tree based models or neural networks or ...

When we model $P(Y \mid X)$ directly, we have a discriminative model.

Bayes Rule

$$P(Y \mid X) = \frac{P(X \mid Y)P(Y)}{P(X)}$$

Which is a shorthand for:

$$\forall(i,j) \quad P(Y = y_i \mid X = x_j) = \frac{P(X = x_j \mid Y = y_i)P(Y = y_i)}{P(X = x_j)}$$

Common terminology:

- $P(Y)$ - prior
- $P(X \mid Y)$ - likelihood
- $P(X)$ - normalization

Learning Bayes optimal classifier

Sky	Temp	Humid	Wind	Water	Forecst	EnjoySpt
Sunny	Warm	Normal	Strong	Warm	Same	Yes
Sunny	Warm	High	Strong	Warm	Same	Yes
Rainy	Cold	High	Strong	Warm	Change	No
Sunny	Warm	High	Strong	Cool	Change	Yes

How do we represent the data? How many parameters do we need?

Prior, $P(Y)$:

- suppose that Y is composed of k classes

Likelihood, $P(X \mid Y)$

- suppose there are p binary features

Learning Bayes optimal classifier

How do we represent the data? How many parameters do we need?

Prior, $P(Y)$:

- suppose that Y is composed of k classes
- we need $k-1$ parameters: $P(Y=y)$ for $y=1,\ldots,k-1$

Likelihood, $P(X \mid Y)$

- suppose there are p binary features
- for each class $(Y=y)$, we need to have a distribution over features $P(X=x \mid Y=y)$
- total number of parameters $k \cdot (2^p - 1)$
- this is huge number of parameters, and we would need a lot of data (and time and storage)

Complex model! High variance with limited data!!!

Conditional independence

Independence of two random variables: $X \perp Y$

$$P(X, Y) = P(X) \cdot P(Y)$$
$$P(Y \mid X) = P(Y)$$

Y and X do not contain information about each other.

Observing Y does not help predicting X.

Observing X does not help predicting Y.

Examples:

- Independent: Winning on roulette this week and next week.
- Dependent: Russian roulette

Conditional independence

X is **conditionally independent** of Y given Z, if for all values of (i, j, k) that random variables X, Y, and Z can take, we have

$$P(X = i, Y = j \mid Z = k) = P(X = i \mid Z = k) \cdot P(Y = j \mid Z = k)$$

Knowing Z makes X and Y independent. We write $X \perp Y \mid Z$.

Examples:

Shoe size and reading skills are dependent. Given *age*, shoe size and reading skills are independent.

Storks deliver babies. Highly statistically significant correlation exists between stork populations and human birth rates across Europe.

Conditional independence

London taxi drivers:

A survey has pointed out a positive and significant correlation between the number of accidents and wearing coats. They concluded that coats could hinder movements of drivers and be the cause of accidents. A new law was prepared to prohibit drivers from wearing coats when driving.

Finally another study pointed out that people wear coats when it rains...

$$P(\text{accidents,coat} \mid \text{rain}) = P(\text{accidents} \mid \text{rain}) \cdot P(\text{coat} \mid \text{rain})$$

Conditional independence

An equivalent definition

X is **conditionally independent** of Y given Z, if for all values of (i, j, k) that random variables X, Y, and Z can take, we have

$$P(X = i \mid Y = j, Z = k) = P(X = i \mid Z = k)$$

Example:

$$P(\text{thunder} \mid \text{rain, lightning}) = P(\text{thunder} \mid \text{lightning})$$

Thunder and rain are not independent. However, if I tell you that there is lightning, they become independent.

How can we use conditional independence in classification?

Goal: Predict Thunder

Features are conditionally independent

- lightning
- rain

Recall: $P(T \mid L, R) \propto P(L, R \mid T) \cdot P(T)$

How many parameters do we need to estimate?

How can we use conditional independence in classification?

How many parameters do we need to estimate?

Without conditional independence, we need 6 parameters to represent $P(L, R \mid T)$.

However, we have $L \perp R \mid T$, so

$$P(L, R \mid T) = P(L \mid T) \cdot P(R \mid T)$$

and we need only 4 parameters.

The Naïve Bayes assumption

Features are independent given class:

$$P(X_1, X_2 \mid Y) = P(X_1 \mid Y) \cdot P(X_2 \mid Y)$$

More generally, if we have p features:

$$P(X_1, \ldots, X_p \mid Y) = \prod_{i=1}^{p} P(X_i \mid Y)$$

The likelihood is product of individual features likelihoods.

How many parameters do we need now?

The Naïve Bayes assumption

How many parameters for $P(X_1, \ldots, X_p \mid Y)$?

- Without assumption we need $k \cdot (2^p - 1)$ parameters

With the Naïve Bayes assumption

$$P(X_1, \ldots, X_p \mid Y) = \prod_{i=1}^{p} P(X_i \mid Y)$$

we need $p \cdot k$ parameters.

Nice reduction! May be to aggressive.

The Naïve Bayes classifier

Given:

- Prior $P(Y)$
- p conditionally independent features X given the class Y
- For each X_i, we have likelihood $P(X_i \mid Y)$

Decision rule:

$$
\begin{aligned}
\hat{y} &= \arg\max_y P(Y \mid X) \\
&= \arg\max_y \frac{P(X \mid Y) \cdot P(Y)}{P(X)} \\
&= \arg\max_y P(X \mid Y) \cdot P(Y) \\
&= \arg\max_y P(Y) \cdot \prod_{i=1}^{p} P(X_i \mid Y)
\end{aligned}
$$

The Naïve Bayes classifier

Given:

- Prior $P(Y)$
- p conditionally independent features X given the class Y
- For each X_i, we have likelihood $P(X_i \mid Y)$

Decision rule:

$$\hat{y} = \arg\max_y P(Y) \cdot \prod_{i=1}^{p} P(X_i \mid Y)$$

If the Naïve Bayes assumption holds, NB is optimal classifier!

How do we estimate the parameters of NB?

We count! For a given dataset

$\mathrm{Count}(A = a, B = b) \equiv$ number of examples where $A = a$ and $B = b$

Prior

$$P(Y = y) = \frac{\mathrm{Count}(Y = y)}{n}$$

Likelihood

$$P(X_i = x_i \mid Y = y) = \frac{\mathrm{Count}(X_i = x_i, Y = y)}{\mathrm{Count}(Y = y)}$$

Subtleties of NB

Usually (always), features are not conditionally independent.

$$P(X_1, \ldots, X_p \mid Y) \neq \prod_{i=1}^{p} P(X_i \mid Y)$$

Actual probabilities $P(Y \mid X)$ often biased towards 0 or 1.

Nonetheless, NB is the single most used classifier out there. NB often performs well, even when the assumption is violated.

Subtleties of NB

What if you never see a training instance where $X_1 = a$ when $Y = b$?

- For example, $Y = \{\text{SpamEmail}\}, X_1 = \{\text{'Enlargement'}\}$
- $P(X_1 = a \mid Y = b) = 0$

What does that imply for classification of test examples?

- For a test example X, what is

$$P(Y = b \mid X_1 = a, X_2, \ldots, X_p)?$$

- Does the probability above depend on the values $X_2, \ldots, X_p$?

Solution: smoothing

- Add "fake" counts

$$\mathrm{SmoothCount}(X_i = x_i, Y = y) = \mathrm{Count}(X_i = x_i, Y = y) + 1$$

Text classification

- classify e-mails (spam, ham)
- classify news articles (what is the topic of the article)
- classify reviews (positive or negative review)

Features X are entire documents (reviews):

> *I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet.*

NB for text classification

$P(X \mid Y)$ is huge.

- Documents contain many words
- There are many possible words in the vocabulary

The Naïve assumption helps a lot

- $P(X_i = x_i \mid Y = y)$ is simply the probability of observing word x_i in a document on topic y
- $P(\text{"hockey"} \mid Y = \text{sports})$

$$\hat{y} = \arg\max_y P(y) \cdot \prod_{i=1}^{\text{LengthDoc}} P(x_i \mid y)$$

Bag of words representation

I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet.

x **love** xxxxxxxxxxxxxxxx **sweet** xxxxxxx **satirical** xxxxxxxxxx xxxxxxxxxxx **great** xxxxxxx xxxxxxxxxxxxxxxxxxx **fun** xxxx xxxxxxxxxxxxx **whimsical** xxxx **romantic** xxxx **laughing** xxxxxxxxxxxxxxxxxxxxxxxxxxxxxx xxxxxxxxxxxxxx **recommend** xxxxx xxxxxxxxxxxxxxxxxxxxxxxxxxxxxx xx **several** xxxxxxxxxxxxxxxxx xxxxx **happy** xxxxxxxxx **again** xxxxxxxxxxxxxxxxxxxxxxxxxxxxxx xxxxxxxxxxxxxxxxx

Bag of words representation

Position in a document does not matter

$$P(X_i = x_i \mid Y = y) = P(X_k = x_i \mid Y)$$

- "Bag of words" representation ignores the order of words in a document
- Sounds really silly, but often works very well!

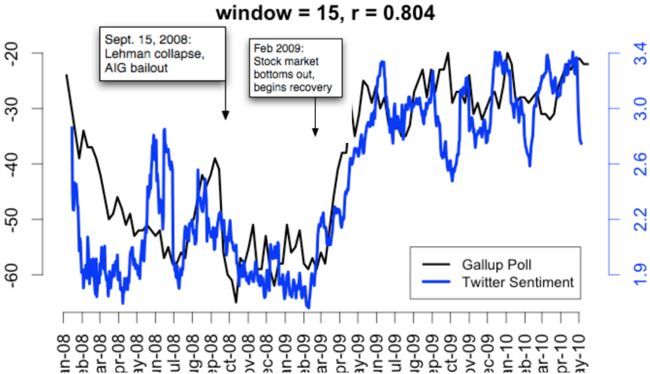
The following two documents are the same

When the lecture is over, remember to wake up the person sitting next to you in the lecture room.

in is lecture lecture next over person remember room sitting the the the to to up wake when you

Sentiment analysis

Twitter sentiment versus Gallup Poll of Consumer Confidence

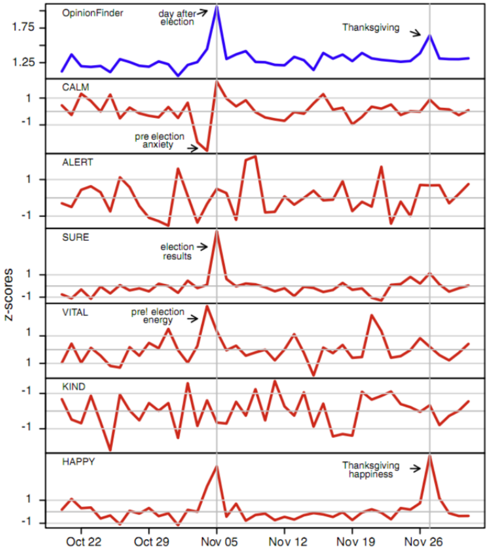


Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. 2010.

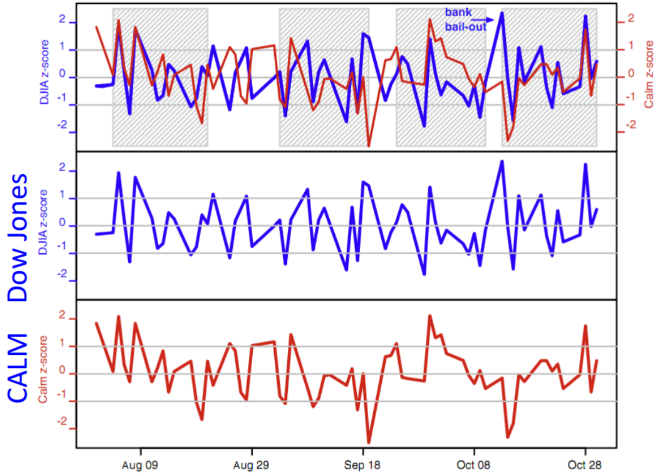
From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. In ICWSMP2010

Sentiment analysis

Twitter mood predicts the stock market. Johan Bollen, Huina Mao, Xiao-Jun Zeng



Sentiment analysis: CALM predicts DJIA 3 days later



Sentiment analysis

R detour: See *NB_reviews.R*

Application of NB to Large Movie Review Dataset.

`http://ai.stanford.edu/~amaas/data/sentiment/index.html`

Bayesian networks

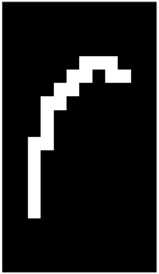
One of the most exciting advancements in statistical AI in the last 10-15 years

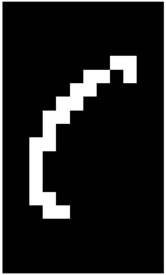
Generalizes naïve Bayes and logistic regression classifiers

Compact representation for exponentially-large probability distributions

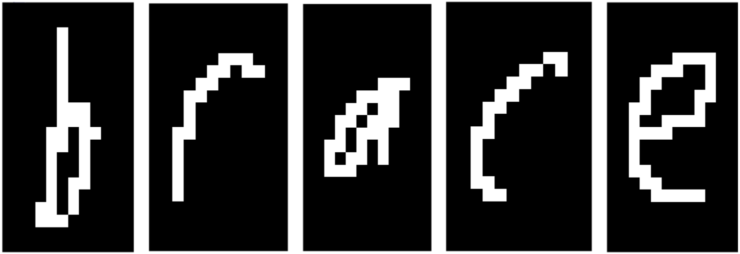
Exploit conditional independencies

Handwritten character recognition





Handwritten character recognition



Applications

- Speech recognition
- Diagnosis of diseases
- Study Human genome
- Modeling fMRI data
- Fault diagnosis
- Modeling sensor network data
- Modeling protein-protein interactions
- Weather prediction
- Computer vision
- many, many more . . .

Causal structure

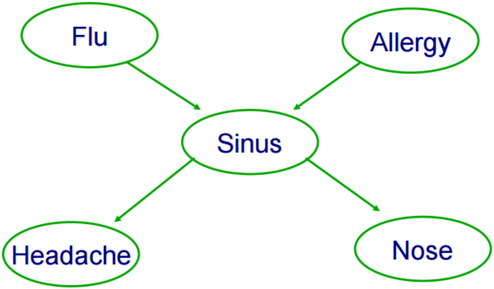
Suppose we know the following:

- The flu causes sinus inflammation
- Allergies cause sinus inflammation
- Sinus inflammation causes a runny nose
- Sinus inflammation causes headaches

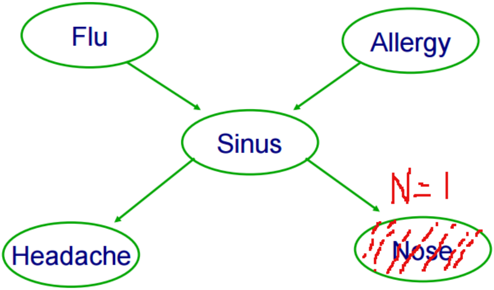
How are these connected?

Causal structure

- The flu causes sinus inflammation
- Allergies cause sinus inflammation
- Sinus inflammation causes a runny nose
- Sinus inflammation causes headaches



What can we do with this?



- Inference $P(F = 1 \mid N = 1)$
- Most probable explanation
  $\max_{f,a,s,h} P(F = f, A = a, S = s, H = h \mid N = 1)$
- Active data collection: What variable should I observe next?

Probabilistic graphical models

Key ideas:

- Conditional independence assumptions are useful
- Naïve Bayes is extreme
- Graphical models express sets of conditional independence assumptions via graph structure
- **Graph structure** + **Conditional Probability Tables** (CPTs) define joint probability distributions over sets of variables/nodes

Two types of graphical models:

- directed graphs (known as Bayesian Networks)
- undirected graphs (known as Markov Random Fields)

Topics in Graphical Models

Representation

- Which joint probability distributions does a graphical models represent?

Inference

- How to answer questions about the joint probability distribution?
  - Marginal distribution of a node variable
  - Most likely assignment of node variables

Learning

- How to learn the parameters and structure of a graphical model?

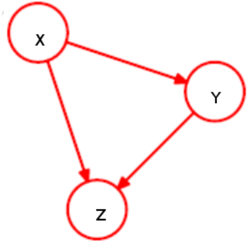
Representation

Which joint probability distributions does a graphical model represent?

Chain rule

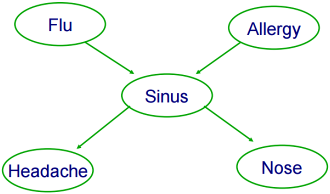
- For any arbitrary distribution
  $P(X, Y, Z) = P(X) \cdot P(Y \mid X) \cdot P(Z \mid X, Y)$

Fully connected directed graph



Representation

Absence of edges in a graphical model conveys useful information.



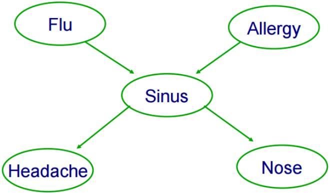
$$P(F, A, S, H, N) = P(F) \cdot P(A) \cdot P(S \mid F, A) \cdot P(H \mid S) \cdot P(N \mid S)$$

How many terms does the left hand side have?
How many parameters?

What about probabilities?

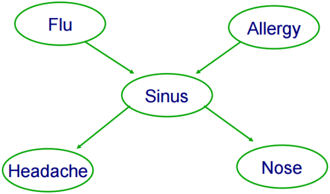
How do we specify the joint probability distribution?

We use conditional probability tables



	F=f, A=f	F=t, A=f	F=f, A=t	F=t,A=t
S=t	0.9	0.8	0.7	0.3
S=f	0.1	0.2	0.3	0.7

Number of parameters?



- more bias
- less flexible
- need less data to learn
- more accurate on smaller datasets

Representation

Which joint probability distributions does a graphical model represent?

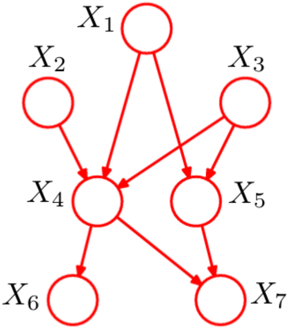
Bayesian Network is a directed acyclic graph (DAG) that, together with CPTs, provides a compact representation for a joint distribution $P(X_1, \ldots, X_p)$.

Conditional probability tables specify $P(X_i \mid \mathrm{parents}(i))$.

$$P(X_1, \ldots, X_p) = \prod_{i=1}^{p} P(X_i \mid \mathrm{parents}(i))$$

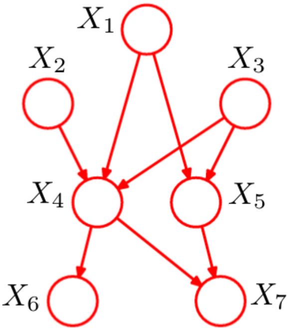
Local Markov Assumption: A variable X is independent of its non-descendants given its parents (only the parents).

Example



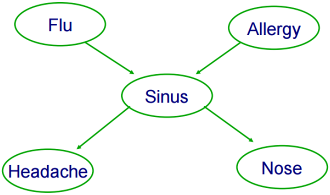
What is $P(X_1, \ldots, X_7)$?

Example



$$P(X_1, \dots, X_7) = P(X_1)P(X_2)P(X_3)P(X_4 \mid X_1, X_2, X_3) \cdot \\ P(X_5 \mid X_1, X_3)P(X_6 \mid X_4)P(X_7 \mid X_4, X_5)$$

Key ingredient: Markov independence assumptions



Local Markov Assumption: If you have no sinus infection, then flu has no influence on headache (flu causes headache but only through sinus).

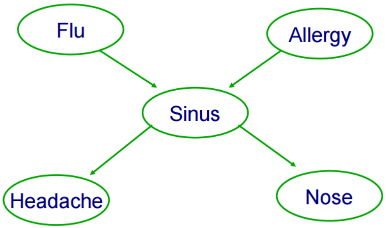
If you tell $H = 1$, that changes probability of Flu.

However, if you first tell me that $S = 1$, then H does not affect probability of Flu.

Markov independence assumptions

Local Markov Assumption: A variable X is independent of its non-descendants given its parents (only the parents).

	parents	non-desc	assumption
S	F,A	-	-
H	S	F,A,N	$H \perp \{F,A,N\} \mid S$
N	S	F,A,H	$N \perp \{F,A,H\} \mid S$
F	-	A	$F \perp A$
A	-	F	$A \perp F$



Joint distribution revisited

Local Markov Assumption: A variable X is independent of its non-descendants given its parents (only the parents).

P(F, A, S, H, N)

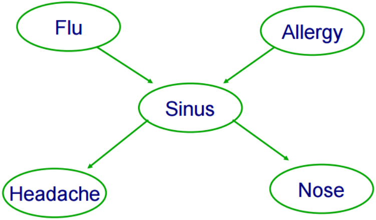
= P(F) P(A|F) P(S|F,A) P(H|S,F,A) P(N|S,F,A,H)

Chain rule

= P(F) P(A) P(S|F,A) P(H|S) P(N|S)

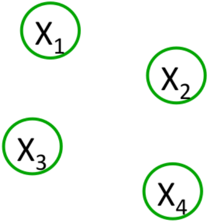
Markov Assumption

$F \perp A$, $H \perp \{F,A\} | S$, $N \perp \{F,A,H\} | S$

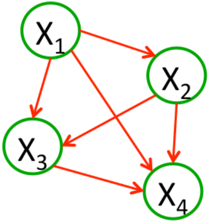


Two special cases

Fully disconnected graph



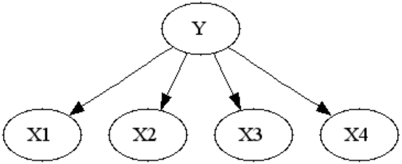
Fully connected graph



What independent assumptions are made?

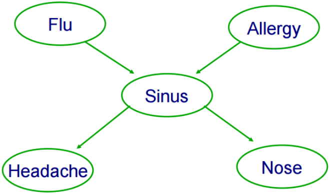
Naïve Bayes revisited

Local Markov Assumption: A variable X is independent of its non-descendants given its parents (only the parents).



What independent assumptions are made?

Exaplaining away



$F \perp A \qquad P(F \mid A = 1) = P(F)$

How about $F \perp A \mid S$?
Is it the case that $P(F \mid A = 1, S = 1) = P(F \mid S = 1)$? No!

$P(F = 1 \mid S = 1)$ is high, but $P(F = 1 \mid A = 1, S = 1)$ not as high, since $A = 1$ explains away $S = 1$.

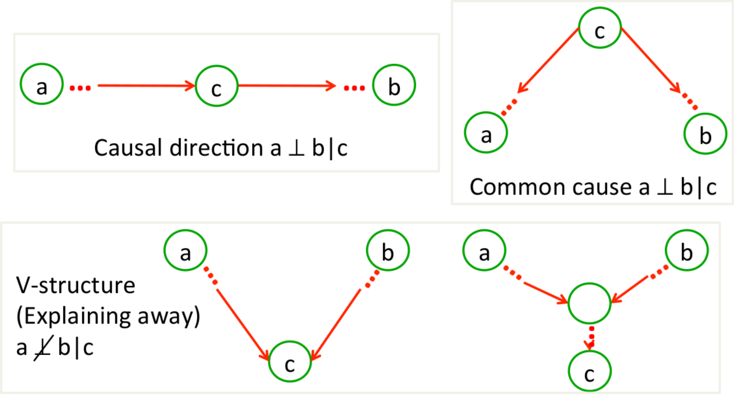
In fact, $P(F = 1 \mid A = 1, S = 1) < P(F = 1 \mid S = 1)$.

Independencies encoded in BN

The only assumption we make is the local Markov assumption.

But many other independencies can be derived.

Three important configurations



Bayesian Networks: Recap

A compact representation for large probability distributions

Semantics of a BN

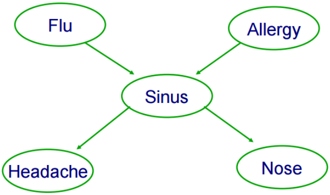
- conditional independence assumptions

Representation

- Variables
- Graph
- CPTs

Why are BNs useful?

Probabilistic inference



Query: $P(X \mid e)$

- We want answer for every assignment of X given evidence e.

Definition of conditional probability

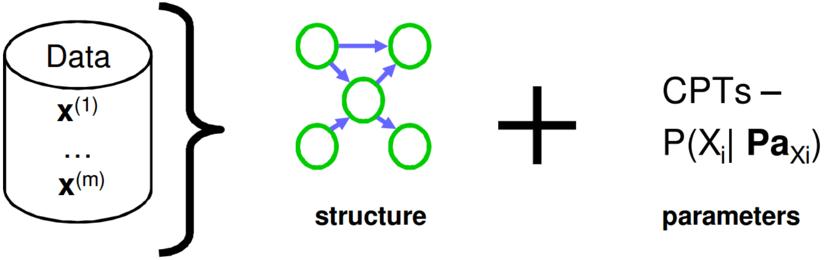
$$P(X \mid e) = \frac{P(X, e)}{P(e)} \propto P(X, e)$$

Marginalization

How do we compute $P(F, N = t)$?

How to do it quickly?

Learning


Data
$\mathbf{x}^{(1)}$
...
$\mathbf{x}^{(m)}$
structure
+
CPTs –
$P(X_i | \mathbf{Pa}_{X_i})$
parameters