Bayes Optimal Classifier & Naïve Bayes

CSE 446: Machine Learning
Emily Fox
University of Washington
March 3, 2017

Classification

GPA, 446 grade, ..

Learn: f: **X** $\mapsto$ Y ← hired / not hired

- **X** – features
- Y – target classes

Suppose you know P(Y|**X**) exactly, how should you classify?

- Bayes optimal classifier:

$X = x = \{ GPA = 3.8, 446 \text{ grade} = 3.9 \}$

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

In logistic regression, model for $P(Y|X=x) = \frac{1}{1+e^{w \cdot x}}$

Model $P(Y|X)$ directly... "discriminitive model"

Recall: Bayes rule

"likelihood" "prior"

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

normalizer

Which is shorthand for:

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

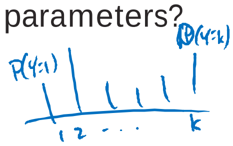
How hard is it to learn the optimal classifier?

- Data =

Sky $x[1]$	Temp $x[2]$	Humid ...	Wind	Water	Forecast $x[6]$	EnjoySpt Y
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	Cold	Change	Yes

- How do we represent these? How many parameters?
  - Prior, P(Y):
    - Suppose Y is composed of k classes

      $k-1$



$\sum P(Y=y) = 1$

  - Likelihood, P(**X**|Y):
    - Suppose **X** is composed of d binary features

$P(X=x \mid Y=y) \leftarrow$ for each class $(Y=y)$, dist over feature values

$P(S=s, T=w, H=n, W=s, Wa=w, F=s \mid E=y)$

$k(2^d - 1) \leftarrow$ a lot of params! need a lot of data,

\# classes, \# params per class

- Complex model → High variance with limited data!!!

Conditional Independence

$X \perp Y$: independence
$P(X,Y) = P(X)P(Y)$

X is **conditionally independent** of Y given Z, if the probability distribution governing X is independent of the value of Y, given the value of Z

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

e.g.,

$$P(Thunder \mid Rain, Lightening) = P(Thunder \mid Lightening)$$

$R \perp T$? No! But $R \perp T \mid L$

Equivalent to: $P(T,R \mid L) = P(T \mid \cancel{R}, L)\,P(R \mid L) = P(T \mid L)\,P(R \mid L)$

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

5 ©2017 Emily Fox CSE 446: Machine Learning

What if features are independent?

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

- Predict Lightening
- From two **conditionally independent** features
  - Thunder
  - Rain

Estimate: $P(X \mid Y)$: $P(T,R \mid L)$ ← w/ no assump: $2 \cdot (2^2 - 1) = 6$ params

But $T \perp R \mid L$

$$P(T,R \mid L) = \underbrace{P(T \mid L)}_{\substack{2\cdot(2^1-1) \\ = 2 \text{ params}}} \underbrace{P(R \mid L)}_{2 \text{ params}} \Bigg\} \ 4 \text{ params!}$$

6 ©2017 Emily Fox CSE 446: Machine Learning

The Naïve Bayes assumption

- Naïve Bayes assumption:
  - Features are independent given class:

$$P(\mathbf{X}[1], \mathbf{X}[2] \mid Y) = P(\mathbf{X}[1] \mid \mathbf{X}[2], Y) P(\mathbf{X}[2] \mid Y)$$
$$= P(\mathbf{X}[1] \mid Y) P(\mathbf{X}[2] \mid Y)$$

  - More generally:

$$\boxed{P(\mathbf{X}[1], \ldots, \mathbf{X}[d] \mid Y) = \prod_{j=1}^{d} \underbrace{P(\mathbf{X}[j] \mid Y)}_{\substack{k \text{ params} \\ (k \text{ classes, binary } X[j])}}}$$

prod. of individual feature likelihoods

- How many parameters now?
  - Suppose **X** is composed of d binary features

No assump: $k(2^d - 1)$ | Naive Bayes: $k \cdot d$

nice reduction! (might be too aggressive)

The Naïve Bayes classifier

- Given:
  - Prior P(Y)
  - d conditionally independent features **X**[j] given the class Y
  - For each **X**[j], we have likelihood P(**X**[j]|Y)
- Decision rule:

$$\arg\max_y P(y \mid x) = \arg\max_y \frac{P(x \mid y) P(y)}{P(x)}$$

P(x) ← doesn't matter

$$\hat{y} = f_{NB}(\mathbf{x}) = \arg\max_y P(y) P(\mathbf{x}[1], \ldots, \mathbf{x}[d] \mid y)$$
$$= \arg\max_y P(y) \prod_j P(\mathbf{x}[j] \mid y)$$

Naive Bayes assumption

- If assumption holds, NB is optimal classifier!

MLE for the parameters of NB

- Given dataset
  - Count(A=a,B=b) == # examples where A=a and B=b
- MLE for NB, simply:
  - Prior: P(Y=y) = $\frac{\text{Count}(Y=y)}{N}$

    e.g. P(Y='hired')
  - Likelihood: P(**X**[j]=**x**[j] | Y=y) = $\frac{\text{Count}(X=x, Y=y)}{\text{Count}(Y=y)}$

P(Grade=high | Y='hired')

$\to P(X|Y) = \frac{P(X,Y)}{P(Y)} = \frac{\text{count}(X=x, Y=y)/N}{\text{Count}(Y=y)/N}$

Subtleties of NB classifier 1 – Violating the NB assumption

- Usually, features are not conditionally independent:

Always

$$\boxed{P(\mathbf{X}[1], \ldots, \mathbf{X}[d] \mid Y) \neq \prod_j P(\mathbf{X}[j] \mid Y)}$$

- Actual probabilities P(Y|**X**) often biased towards 0 or 1
- Nonetheless, NB is one of the most used classifier out there
  - NB often performs well, even when assumption is violated
  - [Domingos & Pazzani '96] discuss some conditions for good performance

Subtleties of NB classifier 2 – Insufficient training data

- What if you never see a training instance where **X**[1]=a when Y=b?
  - e.g., Y={SpamEmail}, **X**[1]={'Viagra'}
  - P(**X**[1]=a | Y=b) = 0

$P(Y|X) \propto P(Y)\prod_{j=1}^{d} P(X[j]|Y)$ one 0 $\Rightarrow$ all 0

- Thus, no matter what the values **X**[2],..., **X**[d] take:
  - P(Y=b | **X**[1]=a, **X**[2],..., **X**[d]) = 0

'free', 'credit', 'Please', 'Sir', 'Mam' ....

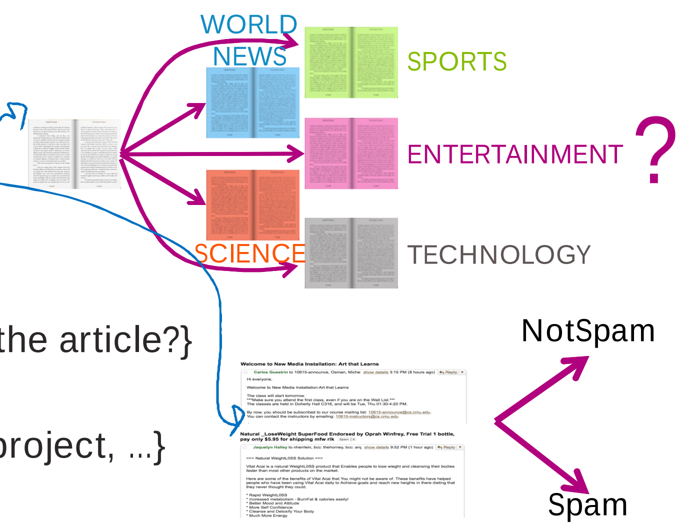
- "Solution": smoothing
  - Add "fake" counts, usually uniformly distributed
  - Equivalent to Bayesian learning

$\text{smoothCount}(X[j]=x, Y=y)$
$= \text{Count}(X[j]=x, Y=y)$
$+ \alpha\,\text{Unif}(X[j], Y)$

$|X[j]| = R_j$
$|Y| = k$

$\frac{1}{R_j k}$

Text classification



- Classify e-mails
  - Y = {Spam,NotSpam}
- Classify news articles
  - Y = {what is the topic of the article?}
- Classify webpages
  - Y = {Student, professor, project, ...}
- What about the features **X**?
  - The text! (very naively)

Features **X** are entire document – **X**[j] for jth word in article

Article from rec.sport.hockey group

```
Path: cantaloupe.srv.cs.cmu.edu!das-news.harvard.e
From: xxx@yyy.zzz.edu (John Doe)
Subject: Re: This year's biggest and worst (opinic
Date: 5 Apr 93 09:53:39 GMT

I can only comment on the Kings, but the most
obvious candidate for pleasant surprise is Alex
Zhitnik. He came highly touted as a defensive
defenseman, but he's clearly much more than that.
Great skater and hard shot (though wish he were
more accurate). In fact, he pretty much allowed
the Kings to trade away that huge defensive
liability Paul Coffey. Kelly Hrudey is only the
biggest disappointment if you thought he was any
good to begin with. But, at best, he's only a
mediocre goaltender. A better choice would be
Tomas Sandstrom, though not through any fault of
his own, but because some thugs in Toronto decided
```


NB for text classification

- P(**X**|Y) is huge!!! ← $10{,}000^{1000}$ possible values!
  - Article at least 1000 words, **X**={**X**[1],..., **X**[1000]}
  - **X**[j] represents jth word in document
    - i.e., the domain of **X**[j] is entire vocabulary, e.g., Webster Dictionary (or more), 10,000 words, etc.

- NB assumption helps a lot!!!
  - P(**X**[j]=**x**[j]|Y=y) is the probability of observing word **x**[j] in a document on topic y

$$f_{NB}(\mathbf{x}) = \arg\max_y P(y) \prod_{j=1}^{LengthDoc} \underbrace{P(\mathbf{x}[j] \mid y)}_{\text{prob. of word in doc given class}}$$

P("hockey" | Y=sports)

Bag of words model

- Typical additional assumption: Position in document doesn't matter

jth word ("hockey") kth word ("hockey")

P(**X**[j]=**x**[j] | Y=y) = P(**X**[k]=**x**[j] | Y=y)

  - "Bag of words" model – order of words on the page ignored
  - Sounds really silly, but often works very well!

$$P(y) \prod_{j=1}^{LengthDoc} P(\mathbf{x}[j] \mid y)$$

Naive Bayes is a bag-of-words assumption for this raw text data represent.

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

Bag of words model

- Typical additional assumption: Position in document doesn't matter

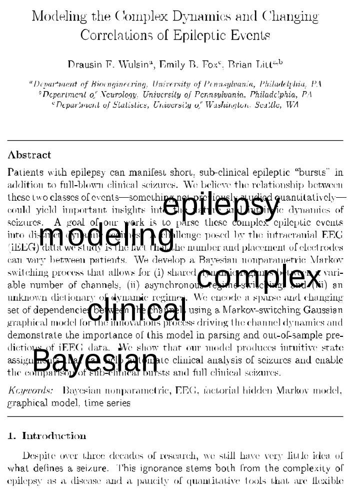
P(**X**[j]=**x**[j] | Y=y) = P(**X**[k]=**x**[j] | Y=y)

  - "Bag of words" model – order of words on the page ignored
  - Sounds really silly, but often works very well!

$$P(y) \prod_{j=1}^{LengthDoc} P(\mathbf{x}[j] \mid y)$$

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

Bag-of-words representation



Modeling the Complex Dynamics and Changing Correlations of Epileptic Events

Drausin F. Wulsin[a], Emily B. Fox[c], Brian Litt[a,b]

[a]Department of Bioengineering, University of Pennsylvania, Philadelphia, PA
[b]Department of Neurology, University of Pennsylvania, Philadelphia, PA
[c]Department of Statistics, University of Washington, Seattle, WA

Abstract

Patients with epilepsy can manifest short, sub-clinical epileptic "bursts" in addition to full-blown clinical seizures. We believe the relationship between these two classes of events—something not previously studied quantitatively—could yield important insights into the nature and intrinsic dynamics of seizures. A goal of our work is to parse these complex epileptic events into distinct dynamic regimes. A challenge posed by the intracranial EEG (iEEG) data we study is the fact that the number and placement of electrodes can vary between patients. We develop a Bayesian nonparametric Markov switching process that allows for (i) shared dynamic regimes between a variable number of channels, (ii) asynchronous regime-switching, and (iii) an unknown dictionary of dynamic regimes. We encode a sparse and changing set of dependencies between the channels using a Markov-switching Gaussian graphical model for the innovations process driving the channel dynamics and demonstrate the importance of this model in parsing and out-of-sample predictions of iEEG data. We show that our model produces intuitive state assignments that can help automate clinical analysis of seizures and enable the comparison of sub-clinical bursts and full clinical seizures.

Keywords: Bayesian nonparametric, EEG, factorial hidden Markov model, graphical model, time series

1. Introduction

Despite over three decades of research, we still have very little idea of what defines a seizure. This ignorance stems both from the complexity of epilepsy as a disease and a paucity of quantitative tools that are flexible

epilepsy
modeling
complex
clinical
Bayesian

17

Bag-of-words representation

Modeling the Complex Dynamics and Changing Correlations of Epileptic Events

Drausin F. Wulsin[a], Emily B. Fox[c], Brian Litt[a,b]

[a]Department of Bioengineering, University of Pennsylvania, Philadelphia, PA
[b]Department of Neurology, University of Pennsylvania, Philadelphia, PA
[c]Department of Statistics, University of Washington, Seattle, WA

Abstract

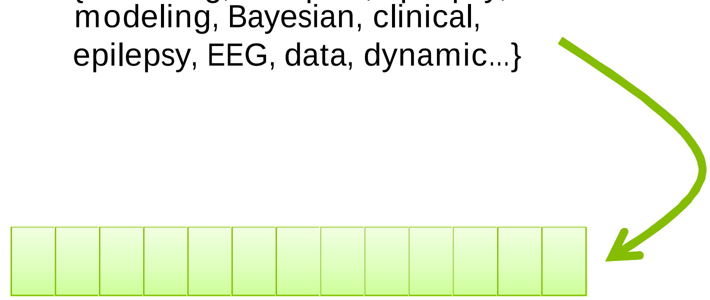
Patients with epilepsy can manifest short, sub-clinical epileptic "bursts" in addition to full-blown clinical seizures. We believe the relationship between these two classes of events—something not previously studied quantitatively—could yield important insights into the nature and intrinsic dynamics of seizures. A goal of our work is to parse these complex epileptic events into distinct dynamic regimes. A challenge posed by the intracranial EEG (iEEG) data we study is the fact that the number and placement of electrodes can vary between patients. We develop a Bayesian nonparametric Markov switching process that allows for (i) shared dynamic regimes between a variable number of channels, (ii) asynchronous regime-switching, and (iii) an unknown dictionary of dynamic regimes. We encode a sparse and changing set of dependencies between the channels using a Markov-switching Gaussian graphical model for the innovations process driving the channel dynamics and demonstrate the importance of this model in parsing and out-of-sample predictions of iEEG data. We show that our model produces intuitive state assignments that can help automate clinical analysis of seizures and enable the comparison of sub-clinical bursts and full clinical seizures.

Keywords: Bayesian nonparametric, EEG, factorial hidden Markov model, graphical model, time series

1. Introduction

Despite over three decades of research, we still have very little idea of what defines a seizure. This ignorance stems both from the complexity of epilepsy as a disease and a paucity of quantitative tools that are flexible

{modeling, complex, epilepsy, modeling, Bayesian, clinical, epilepsy, EEG, data, dynamic…}



18

NB with bag of words for text classification

(better model: n-grams
don't use all, just most popular)

- Learning phase:
  - Prior P(Y)
    - Count how many documents you have from each topic (+ prior)
  - P(**X**[j]|Y)

    smoothing
    - For each topic, count how many times you saw word in documents of this topic (+ prior)
- Test phase:
  - For each document
    - Use naïve Bayes decision rule

$$f_{NB}(\mathbf{x}) = \arg\max_y P(y) \prod_{j=1}^{LengthDoc} P(\mathbf{x}[j] \mid y)$$

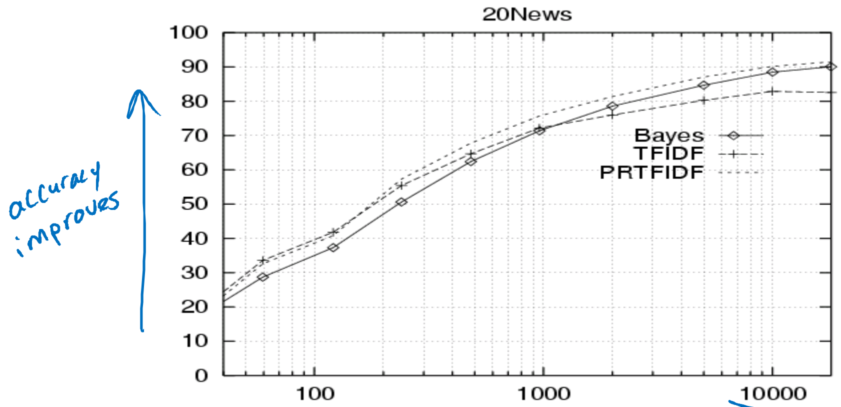
Twenty News Groups results

Given 1000 training documents from each group
Learn to classify new documents according to which newsgroup it came from

comp.graphics	misc.forsale
comp.os.ms-windows.misc	rec.autos
comp.sys.ibm.pc.hardware	rec.motorcycles
comp.sys.mac.hardware	rec.sport.baseball
comp.windows.x	rec.sport.hockey
alt.atheism	sci.space
soc.religion.christian	sci.crypt
talk.religion.misc	sci.electronics
talk.politics.mideast	sci.med
talk.politics.misc	
talk.politics.guns	

Naive Bayes: 89% classification accuracy

Learning curve for Twenty News Groups

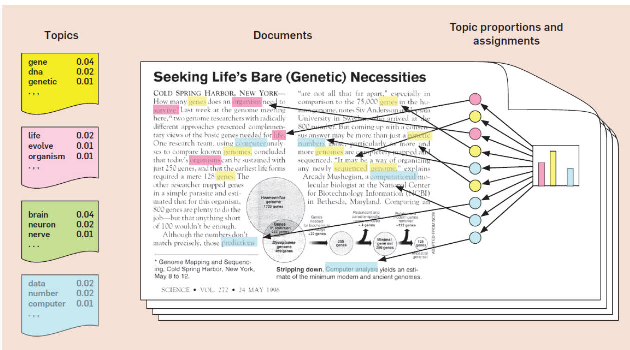


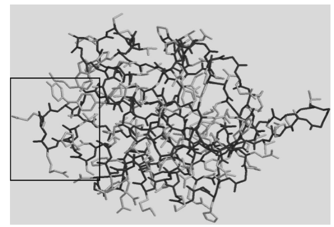
Accuracy vs. Training set size (1/3 withheld for test)

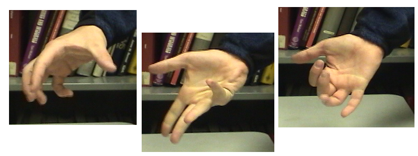
Bayesian Networks – Representation

CSE 446: Machine Learning
Emily Fox
University of Washington
March 3, 2017

Learning from structured data







TrueSkill: A Bayesian Skill Rating System

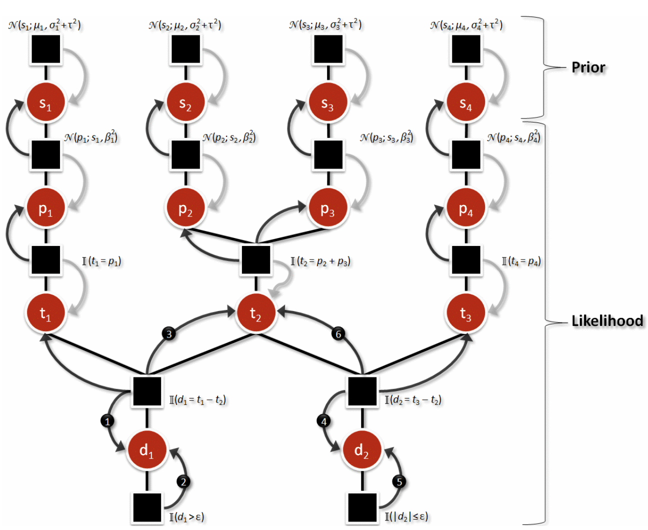
Herbrich et al., 2007

Skill

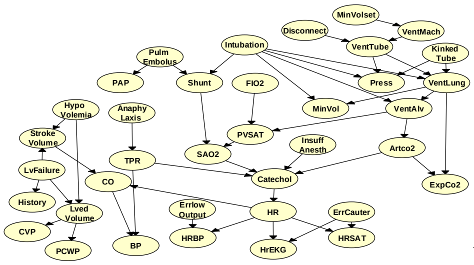
Player performance

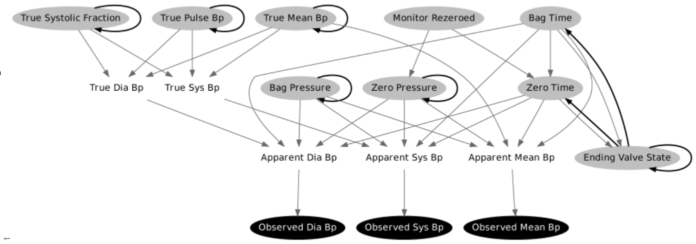
Team performance

Observed team performance difference

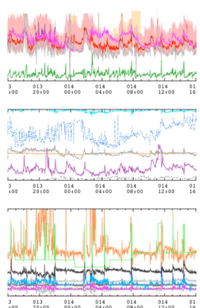


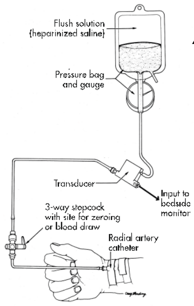
ICU Monitoring





Beinlich et al., 1989

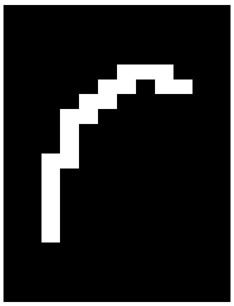
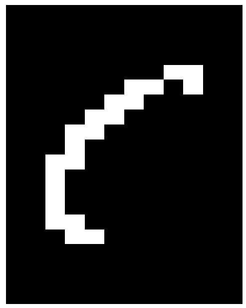




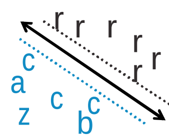
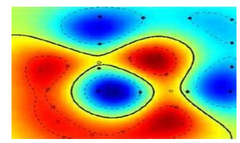
Aleks, Russell, et al., 2008

Digging in:
Learning with and without context/structure

Without context: Handwriting recognition

Character recognition,
e.g., kernel SVMs

Without context: Webpage classification

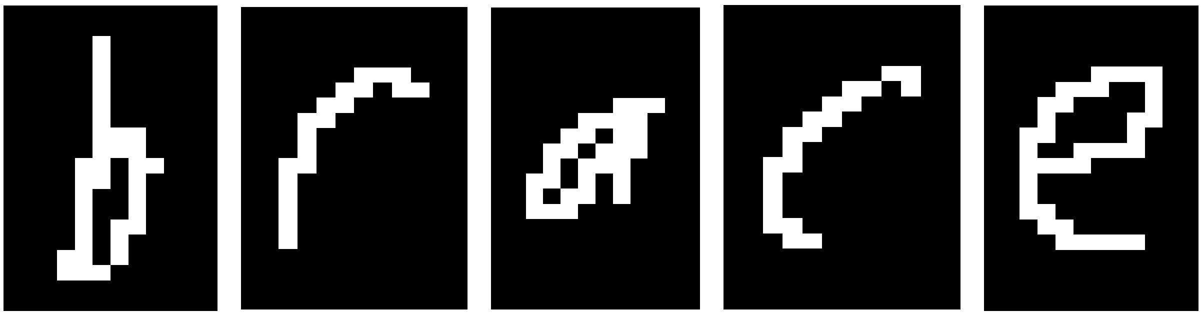
Company website

University website

Personal website

…

With context: Handwriting recognition



With context: Webpage classification

Modeling structured relationships via Bayesian networks

Today – Bayesian networks

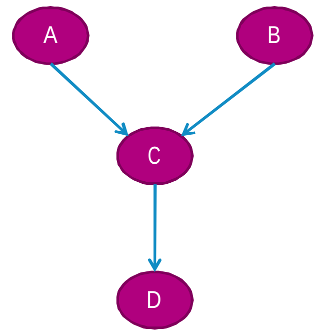
- Provided a huge advancement in AI/ML
- Generalizes naïve Bayes and logistic regression
- Compact representation for exponentially-large probability distributions
- Exploit conditional independencies

Bayesian network representation

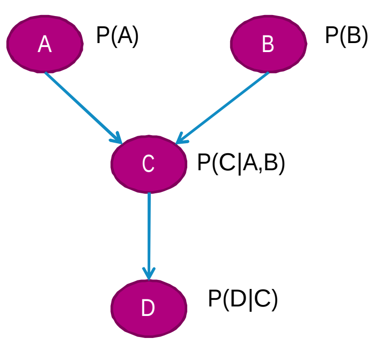
Compact representation of a probability distribution.

Vertices: Random Variables
Edges: Conditional dependencies
"probabilistic relationships"

Directed Acyclic Graph



Bayesian network probability factorization



P(A,B,C,D) = P(A) P(B) P(C|A,B) P(D|C)

One **CPT** (conditional probability table) for each variable

P(variable | parents of variable)

implies the factorization:

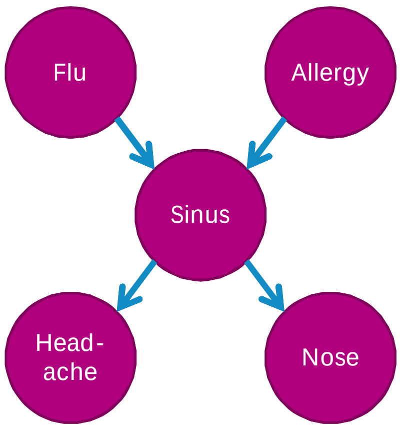
$$P(X) = \prod_{i=1}^{|X|} P(X_i|\mathbf{parents}(X_i))$$

What a Bayesian network represents (in detail) and what does it buy you?

Causal structure

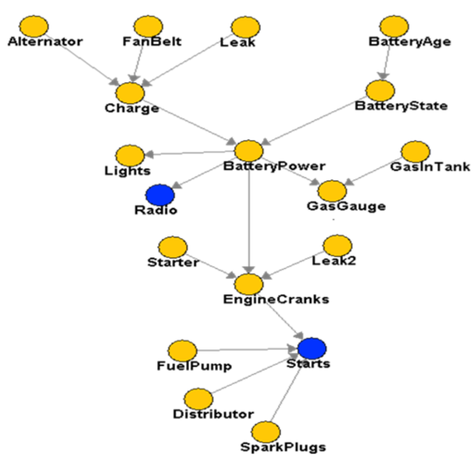
- Suppose we know the following:
  - The flu causes sinus inflammation
  - Allergies cause sinus inflammation
  - Sinus inflammation causes a runny nose
  - Sinus inflammation causes headaches
- How are these connected?

Possible queries



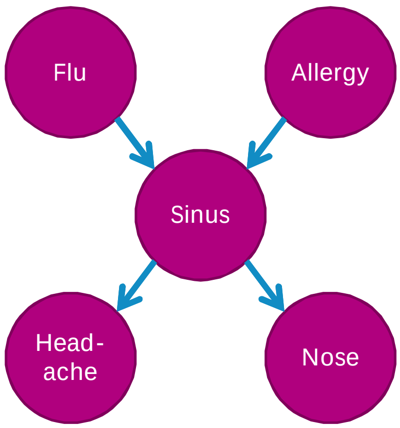
- Inference
- Most probable explanation
- Active data collection

CarStarts? Bayesian network

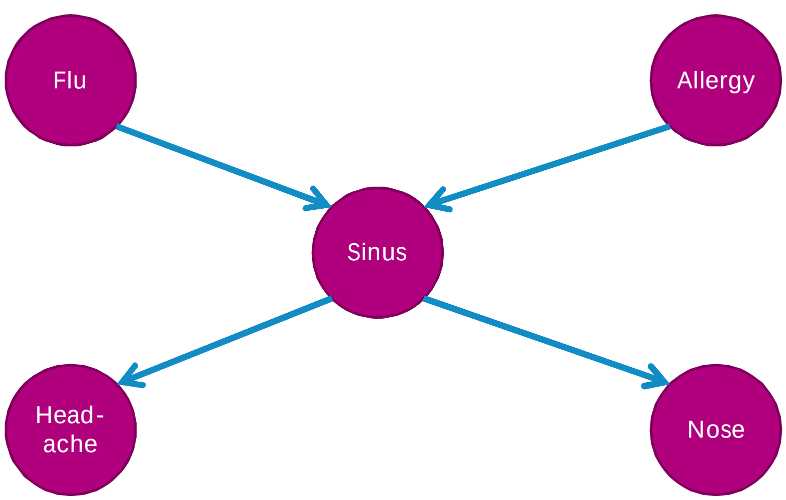


- 18 binary attributes
- Inference
  - P(BatteryAge|Starts=f)
- 2^{16} terms, why so fast?
- Not impressed?
  - HailFinder BN – more than 3^{54} = 58149737003040059690390169 terms

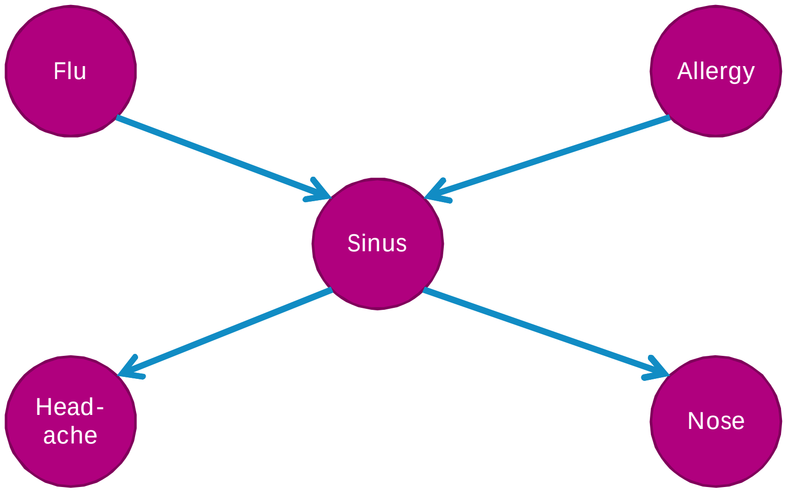
Factored joint distribution – A preview



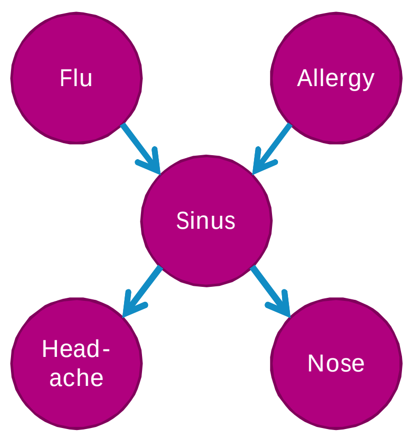
What are these probabilities? Conditional probability tables (CPTs)



Number of parameters



Factorization speeds up inference



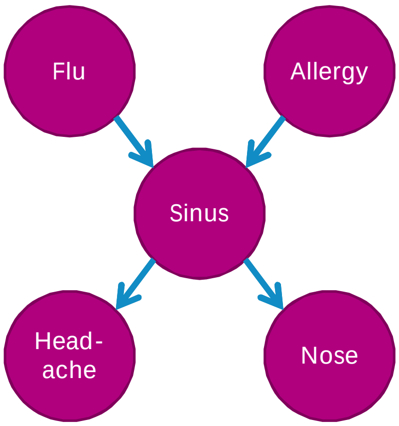
Exploit distributivity:

$$P(F=x_F|N=t) \propto \sum_{x_A,x_S,x_H} P(F=x_F, A=x_A, S=x_S, H=x_H, N=t)$$

$$= \sum_{x_A,x_S,x_H} P(F=x_F)P(A=x_A)P(S=x_S \mid F=x_F, A=x_A)P(H=x_H \mid S=x_S)P(N=t \mid S=x_S)$$

$$= P(F=x_F)\sum_{x_A} P(A=x_A)\sum_{x_S} P(S=x_S \mid F=x_F, A=x_A)P(N=t \mid S=x_S)\sum_{x_H} P(H=x_H \mid S=x_S)$$

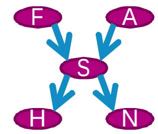
Key: Independence assumptions



Knowing sinus separates variables from each other

Marginal and conditional independence

(Marginal) Independence



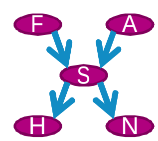
- Flu and Allergy are (marginally) independent

Flu = t	
Flu = f	

Allergy = t	
Allergy = f	

	Flu = t	Flu = f
Allergy = t		
Allergy = f		

Conditional independence

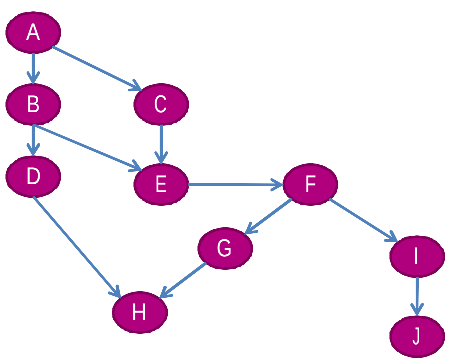


- Flu and Headache are not (marginally) ind.
- Flu and Headache are independent given Sinus infection
- More generally:

Conditional independence statements encoded by Bayesian networks

What is a Bayes net assuming?

Local Markov Assumption: A variable X is independent of its non-descendents given its parents



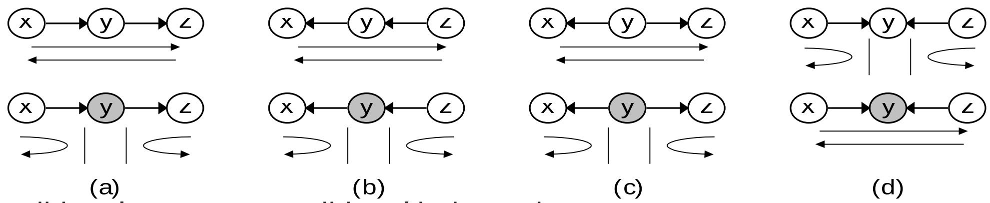
$E \perp A \mid B,C$
$E \perp D \mid B,C$
$F \perp B \mid E$

Allows you to read off some simple conditional independence relationships

48

Conditional independence in Bayes nets

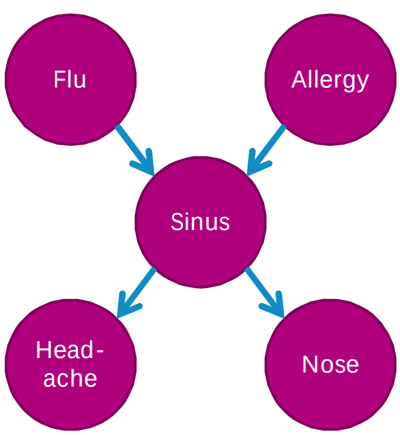
- Consider 4 different junction configurations



- Conditional versus unconditional independence:

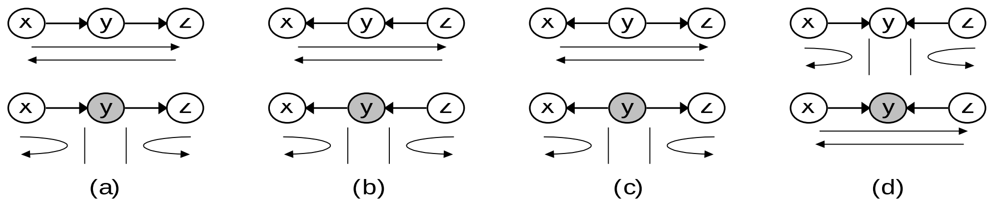
Explaining away example

Local Markov Assumption:
A variable X is independent of its non-descendents given its parents



Bayes ball algorithm

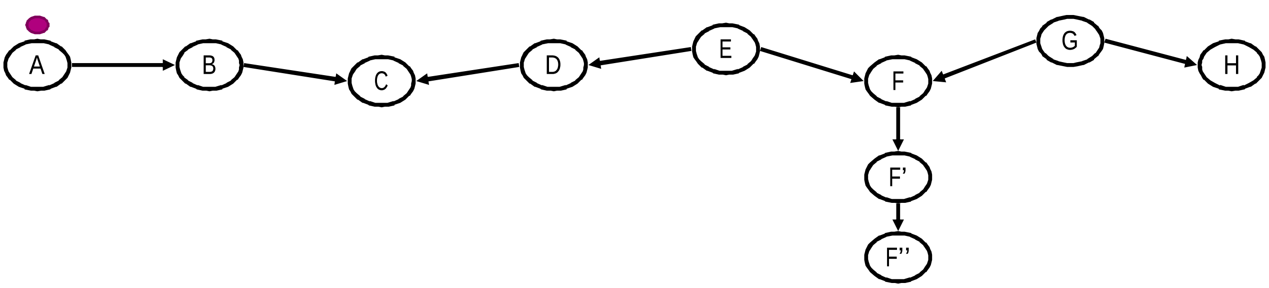
- Consider 4 different junction configurations



- Bayes ball algorithm:

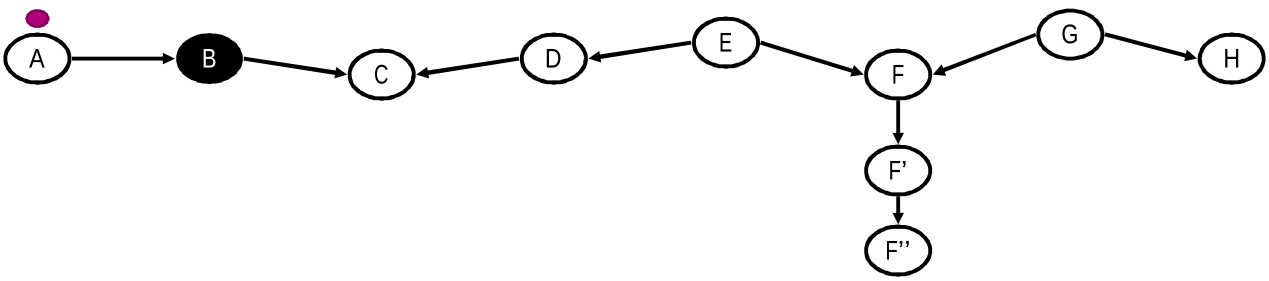
Bayes ball example

A path from A to H is Active if the Bayes ball can get from A to H



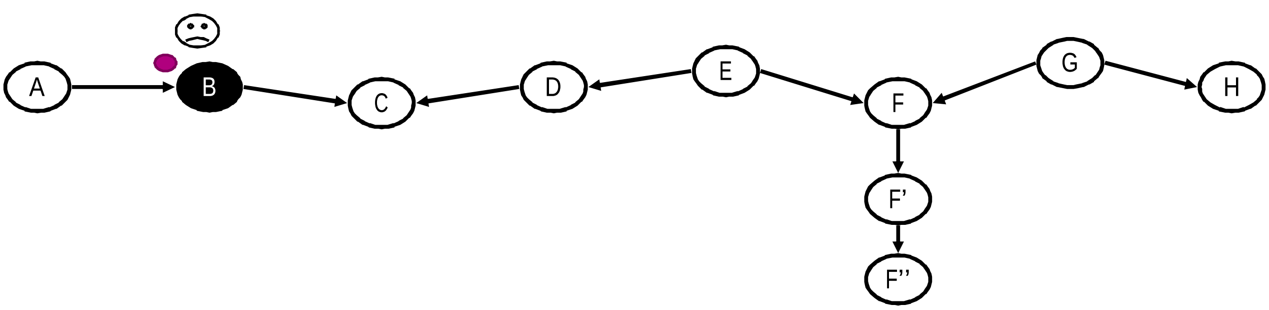
Bayes ball example

A path from A to H is Active if the Bayes ball can get from A to H



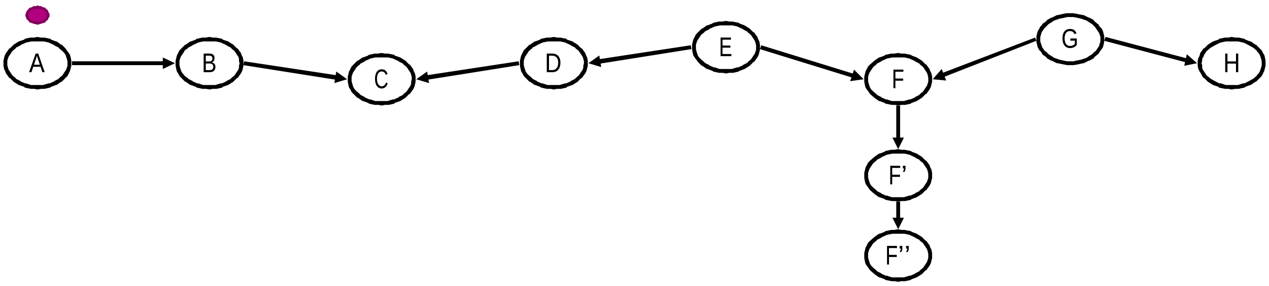
Bayes ball example

A path from A to H is Active if the Bayes ball can get from A to H



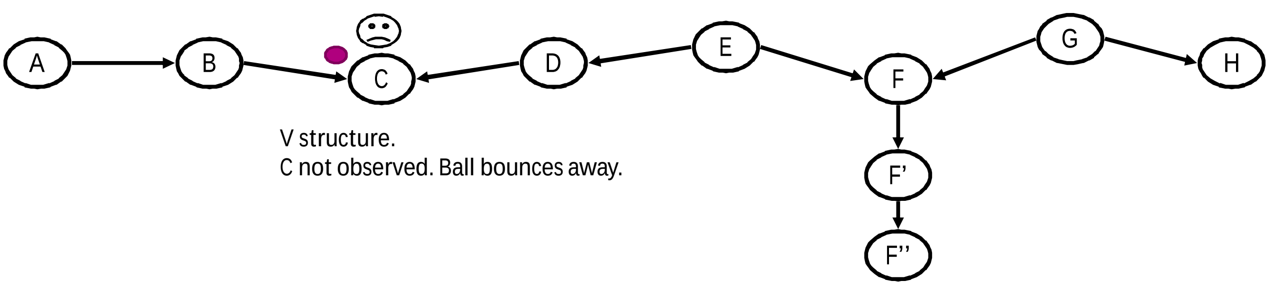
Bayes ball example

A path from A to H is Active if the Bayes ball can get from A to H



Bayes ball example

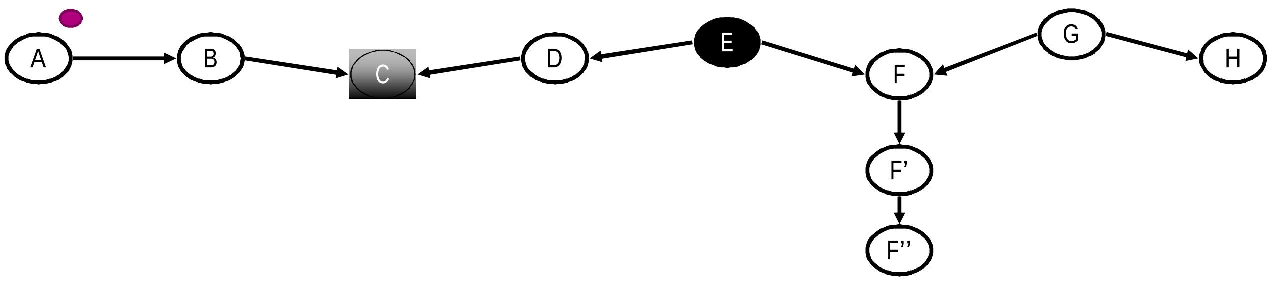
A path from A to H is Active if the Bayes ball can get from A to H



V structure.
C not observed. Ball bounces away.

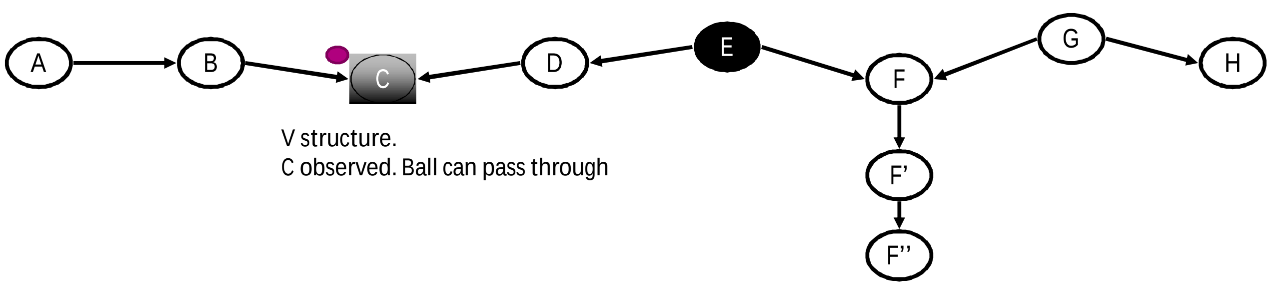
Bayes ball example

A path from A to H is Active if the Bayes ball can get from A to H



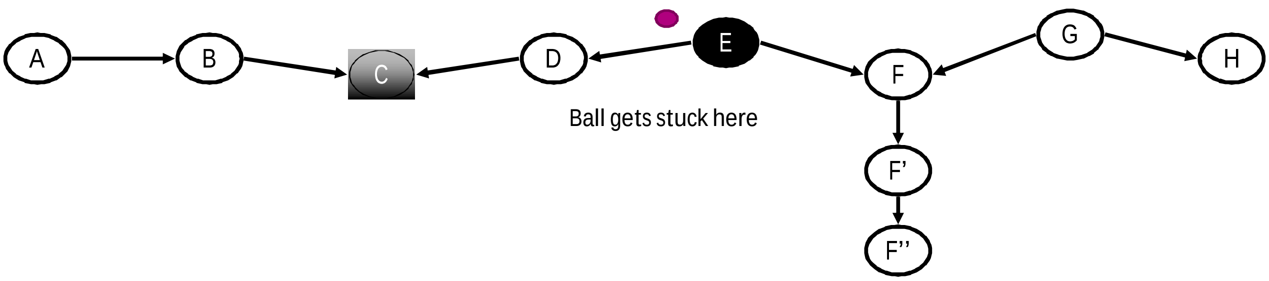
Bayes ball example

A path from A to H is Active if the Bayes ball can get from A to H



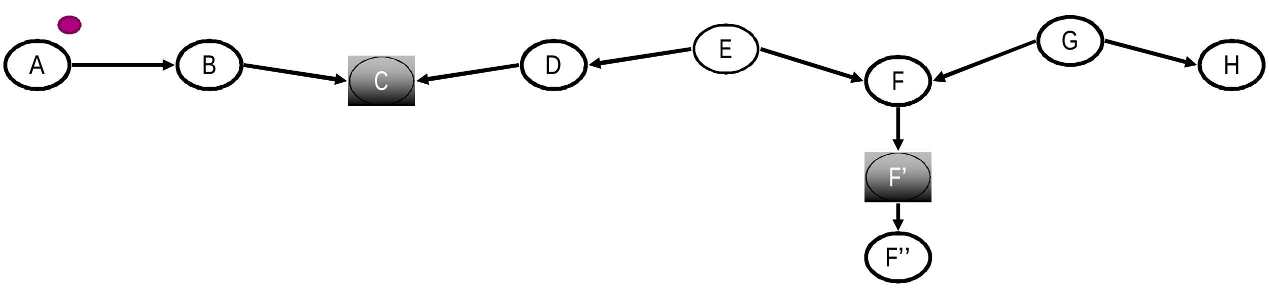
Bayes ball example

A path from A to H is Active if the Bayes ball can get from A to H



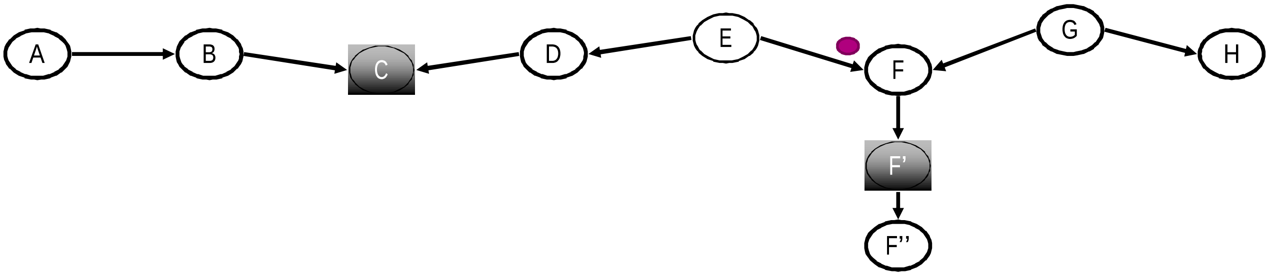
Bayes ball example

A path from A to H is Active if the Bayes ball can get from A to H



Bayes ball example

A path from A to H is Active if the Bayes ball can get from A to H



V structure.
Descendent of F observed.
Ball can pass through

Bayes ball example

A path from A to H is Active if the Bayes ball can get from A to H