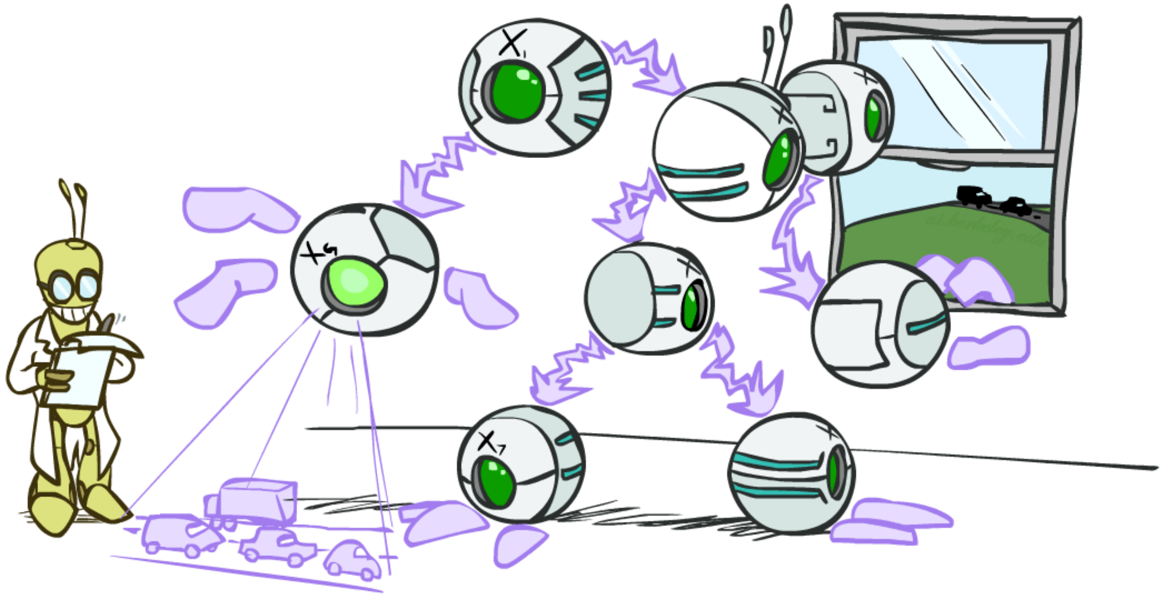
Bayesian Networks: Inference
Machine Learning



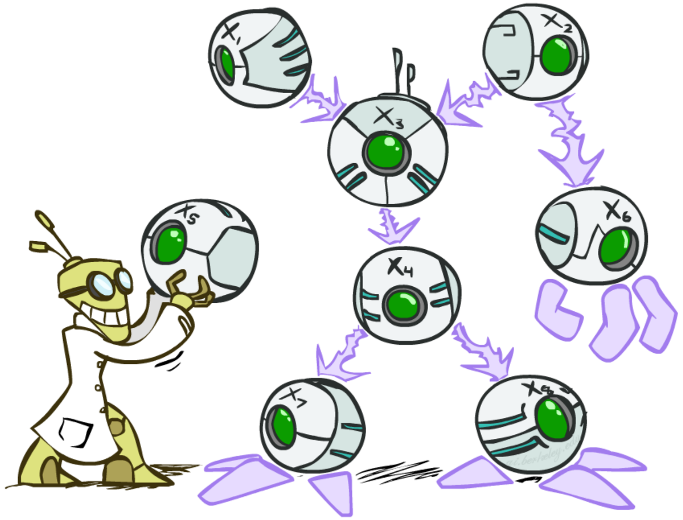
These slides are based on the slides created by Dan Klein and Pieter Abbeel for CS188 Intro to AI at UC Berkeley - http://ai.berkeley.edu.
The artwork is by Ketrina Yim.

Bayes' Nets

- ✓ Representation
- ✓ Conditional Independences
- Probabilistic Inference
- Learning Bayes' Nets from Data

Bayes' Net Representation

- A directed, acyclic graph, one node per random variable
- A conditional probability table (CPT) for each node
- Bayes' nets implicitly encode joint distributions
  - As a product of local conditional distributions



Inference

- Inference: calculating some useful quantity from the Bayes'net

- Examples:

  - Posterior probability

    $$P(Q|E_1 = e_1, \ldots E_k = e_k)$$

  - Most likely explanation:

    $$\text{argmax}_q \; P(Q = q|E_1 = e_1 \ldots)$$

Inference by Enumeration

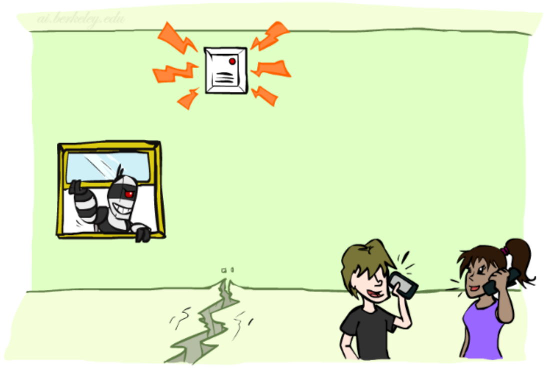
- General case:
  - Evidence variables: $E_1 \dots E_k = e_1 \dots e_k$
  - Query variable(s) : Q
  - Hidden variables: $H_1 \dots H_r$

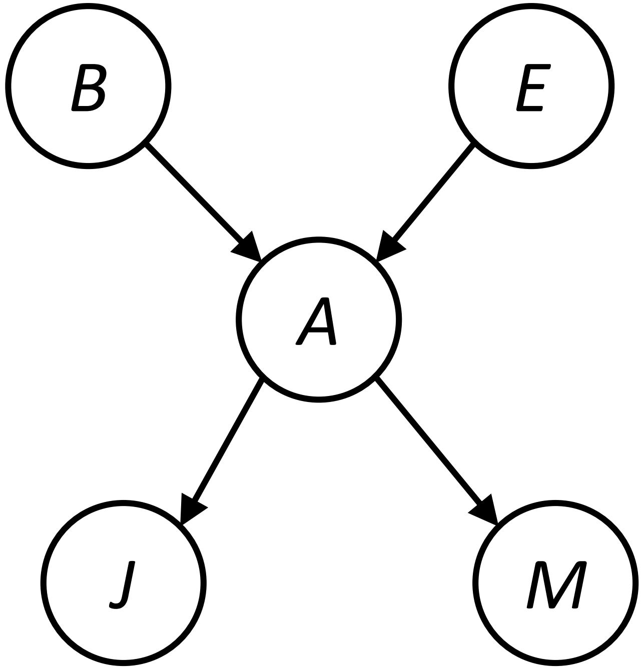
$$\left.\begin{array}{l} E_1 \dots E_k = e_1 \dots e_k \\ Q \\ H_1 \dots H_r \end{array}\right\} \begin{array}{c} X_1, X_2, \dots X_n \\ \textit{All variables} \end{array}$$

  - We want: $P(Q|e_1 \dots e_k)$

Inference by Enumeration in Bayes' Net

- John and Mary called.
- Was there a burglary?
- Query?
  - B
- Evidence?
  - J, M
- Hidden Variables?
  - E, A
- What probability are we looking for?
  - P(+b | +j, +m)





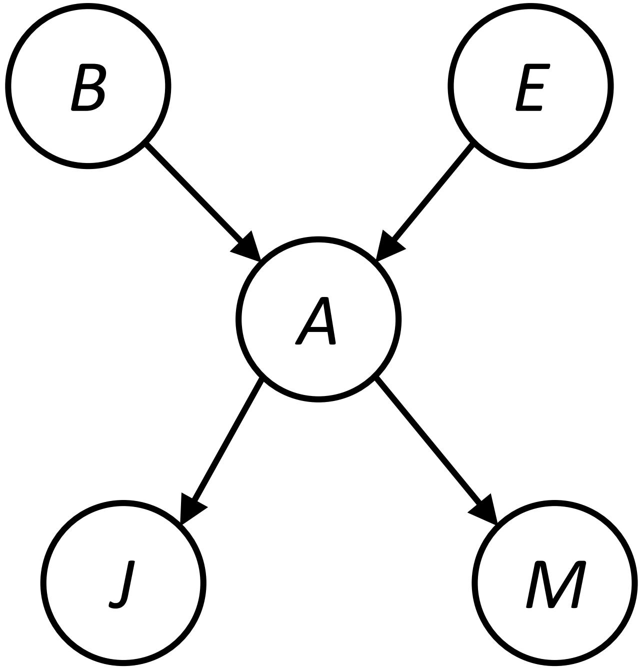
Inference by Enumeration in Bayes' Net

By definition of conditional probability:

$$P(+b \mid +j, +m) = \frac{P(+b, +j, +m)}{P(+j, +m)}$$

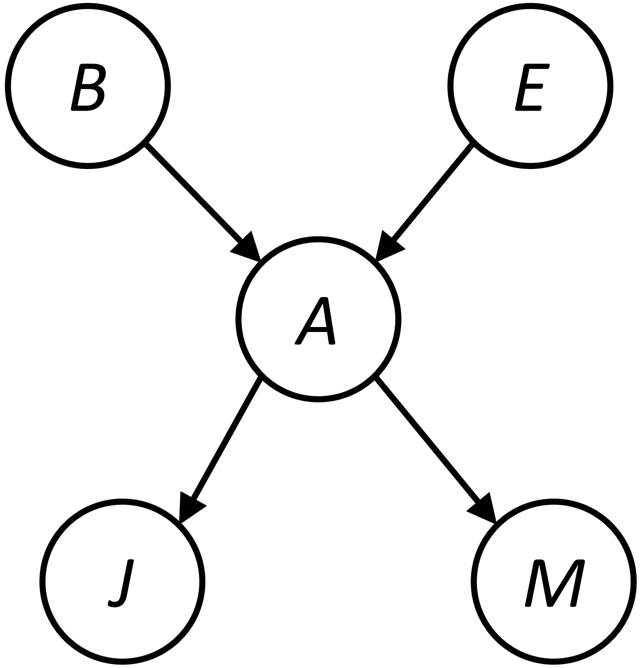
We also know that P(X) = P(X, +y) + P(X, -y):

$$\begin{aligned} P(+b, +j, +m) = \; & P(+b, +j, +m, +e, +a) \; + \\ & P(+b, +j, +m, +e, -a) \; + \\ & P(+b, +j, +m, -e, +a) \; + \\ & P(+b, +j, +m, -e, -a) \end{aligned}$$

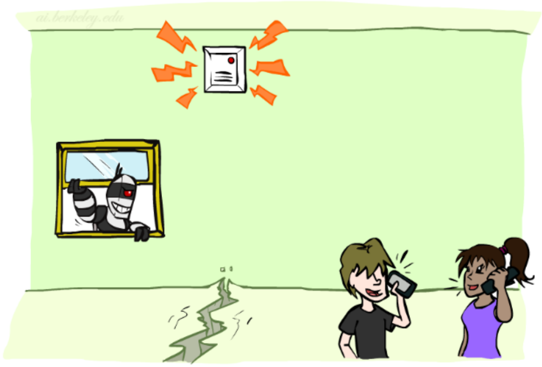


Example: Alarm Network

B	P(B)
+b	0.001
-b	0.999



E	P(E)
+e	0.002
-e	0.998



A	J	P(J|A)
+a	+j	0.9
+a	-j	0.1
-a	+j	0.05
-a	-j	0.95

A	M	P(M|A)
+a	+m	0.7
+a	-m	0.3
-a	+m	0.01
-a	-m	0.99

B	E	A	P(A|B,E)
+b	+e	+a	0.95
+b	+e	-a	0.05
+b	-e	+a	0.94
+b	-e	-a	0.06
-b	+e	+a	0.29
-b	+e	-a	0.71
-b	-e	+a	0.001
-b	-e	-a	0.999

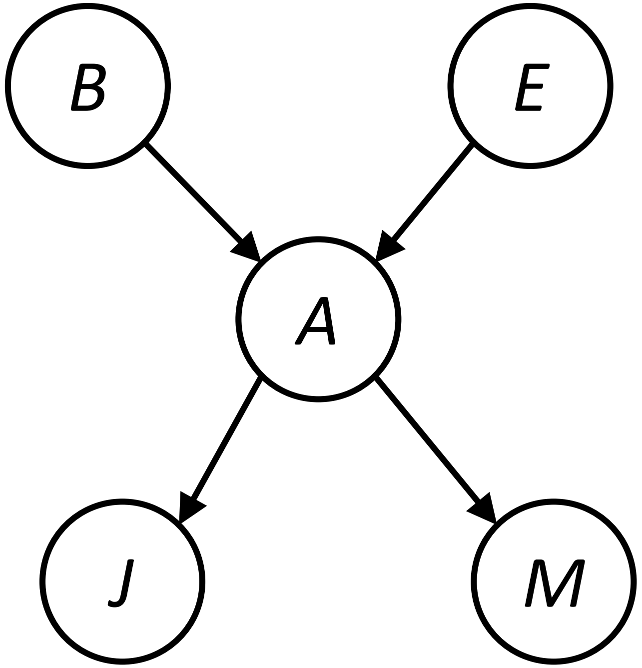
$$P(+b, +j, +m, +e, +a) =$$

$$P(+b)\ P(+e)\ P(+a|+b, +e)\ P(+j|+a)\ P(+m|+a) =$$

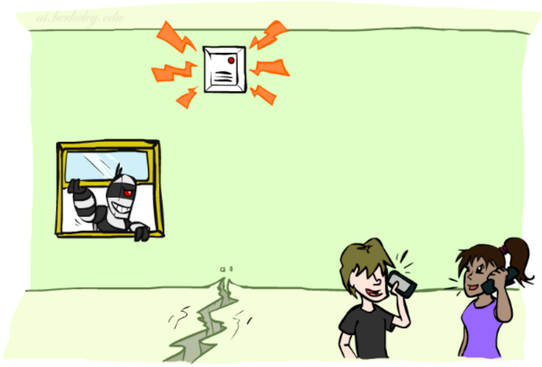
$$0.001 \times 0.002 \times 0.95 \times 0.9 \times 0.7 = 1.2 \times 10^{-06}$$

Example: Alarm Network

B	P(B)
+b	0.001
-b	0.999



E	P(E)
+e	0.002
-e	0.998



A	J	P(J|A)
+a	+j	0.9
+a	-j	0.1
-a	+j	0.05
-a	-j	0.95

A	M	P(M|A)
+a	+m	0.7
+a	-m	0.3
-a	+m	0.01
-a	-m	0.99

B	E	A	P(A|B,E)
+b	+e	+a	0.95
+b	+e	-a	0.05
+b	-e	+a	0.94
+b	-e	-a	0.06
-b	+e	+a	0.29
-b	+e	-a	0.71
-b	-e	+a	0.001
-b	-e	-a	0.999

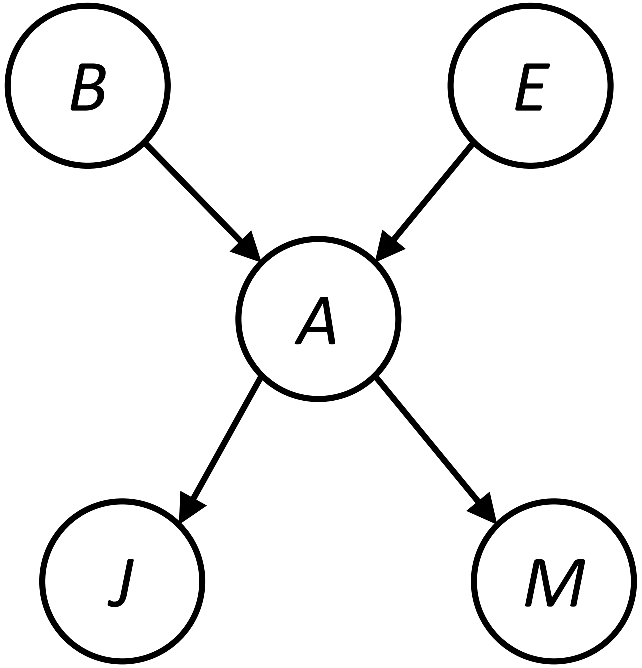
P(+b, +j, +m, +e, **-a**) =

$$P(+b)\, P(+e)\, P(-a|+b, +e)\, P(+j|-a)\, P(+m|-a) =$$

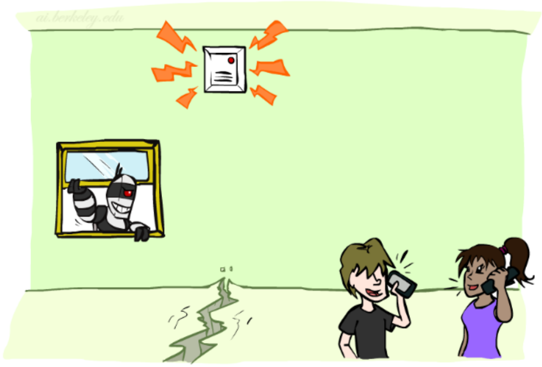
$$0.001 \times 0.002 \times 0.05 \times 0.05 \times 0.01 = 5 \times 10^{-11}$$

Example: Alarm Network

B	P(B)
+b	0.001
-b	0.999



E	P(E)
+e	0.002
-e	0.998



A	J	P(J|A)
+a	+j	0.9
+a	-j	0.1
-a	+j	0.05
-a	-j	0.95

A	M	P(M|A)
+a	+m	0.7
+a	-m	0.3
-a	+m	0.01
-a	-m	0.99

B	E	A	P(A|B,E)
+b	+e	+a	0.95
+b	+e	-a	0.05
+b	-e	+a	0.94
+b	-e	-a	0.06
-b	+e	+a	0.29
-b	+e	-a	0.71
-b	-e	+a	0.001
-b	-e	-a	0.999

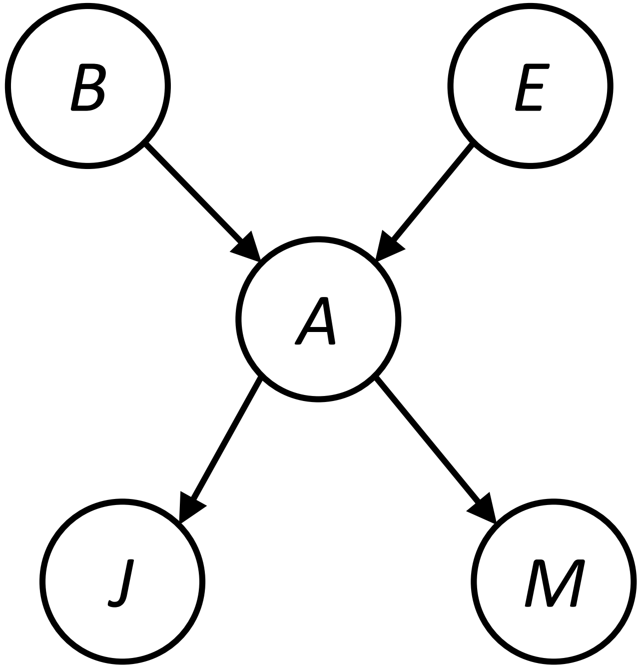
P(+b, +j, +m, **-e,** +a) =

P(+b) P(-e) P(+a|+b, -e) P(+j|+a) P(+m|+a) =

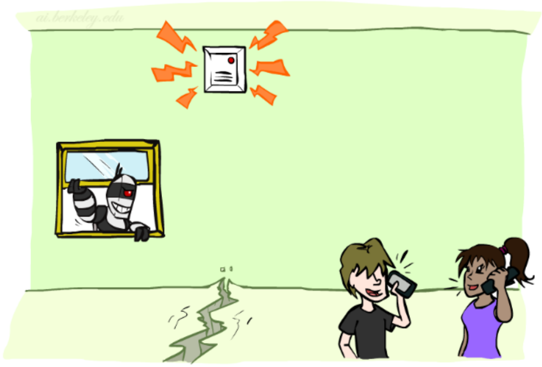
0.001 x 0.998 x 0.94 x 0.9 x 0.7 = 0.000591

Example: Alarm Network

B	P(B)
+b	0.001
-b	0.999



E	P(E)
+e	0.002
-e	0.998



A	J	P(J|A)
+a	+j	0.9
+a	-j	0.1
-a	+j	0.05
-a	-j	0.95

A	M	P(M|A)
+a	+m	0.7
+a	-m	0.3
-a	+m	0.01
-a	-m	0.99

B	E	A	P(A|B,E)
+b	+e	+a	0.95
+b	+e	-a	0.05
+b	-e	+a	0.94
+b	-e	-a	0.06
-b	+e	+a	0.29
-b	+e	-a	0.71
-b	-e	+a	0.001
-b	-e	-a	0.999

$P(+b, +j, +m, -e, \mathbf{-a}) =$

$P(+b)\, P(-e)\, P(-a|+b, -e)\, P(+j|-a)\, P(+m|-a) =$

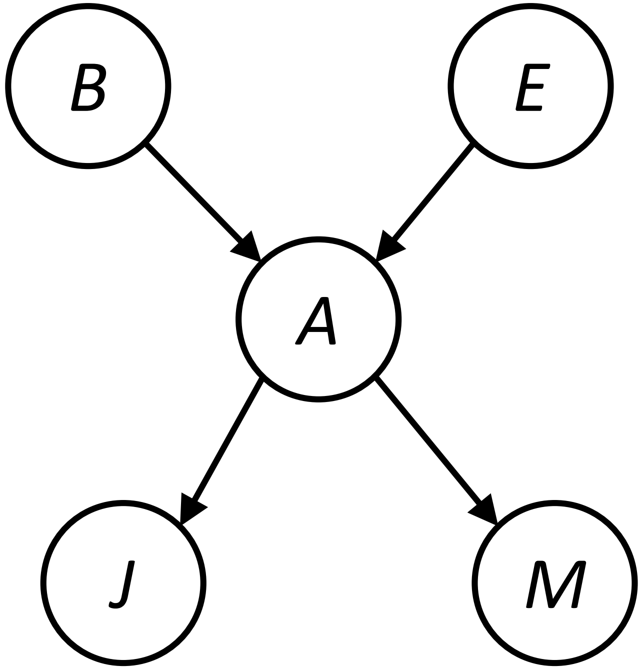
$0.001 \times 0.998 \times 0.06 \times 0.05 \times 0.01 = 3 \times 10^{-08}$

Inference by Enumeration in Bayes' Net

Putting it all together:

$$P(+b \mid +j, +m) = \frac{P(+b,+j,+m)}{P(+j,+m)}$$

$$\begin{aligned}P(+b, +j, +m) &= P(+b, +j, +m, +e, +a) + \\ &\quad P(+b, +j, +m, +e, -a) + \\ &\quad P(+b, +j, +m, -e, +a) + \\ &\quad P(+b, +j, +m, -e, -a) \\ &= 1.2 \times 10^{-06} + 5 \times 10^{-11} + \\ &\quad 0.000591 + 3 \times 10^{-08} \\ &= 0.000592\end{aligned}$$

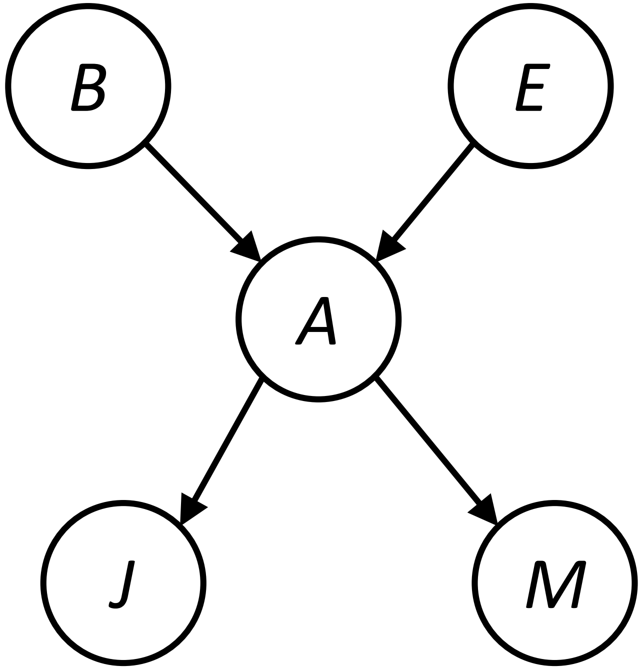


Inference by Enumeration in Bayes' Net

Similarly, we can calculate: P(-b | +j, +m)

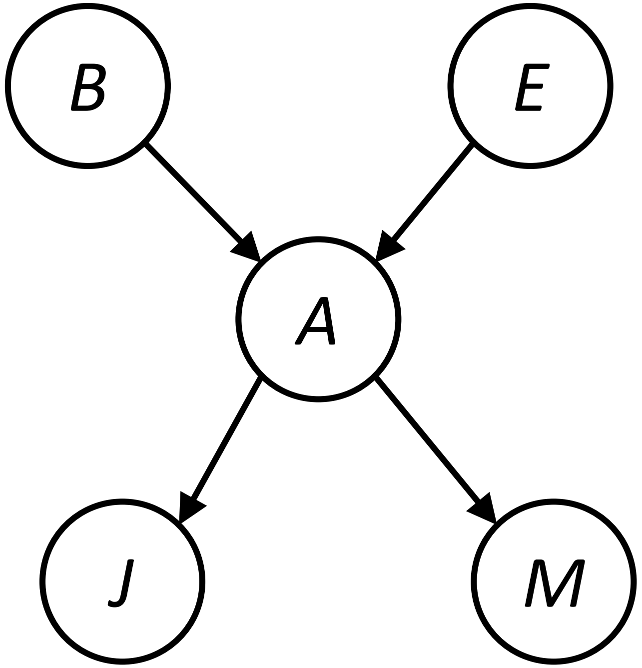
$$P(-b \mid +j, +m) = \frac{P(-b,+j,+m)}{P(+j,+m)}$$

$$\begin{aligned} P(-b, +j, +m) = {} & P(-b, +j, +m, +e, +a) + \\ & P(-b, +j, +m, +e, -a) + \\ & P(-b, +j, +m, -e, +a) + \\ & P(-b, +j, +m, -e, -a) \end{aligned}$$

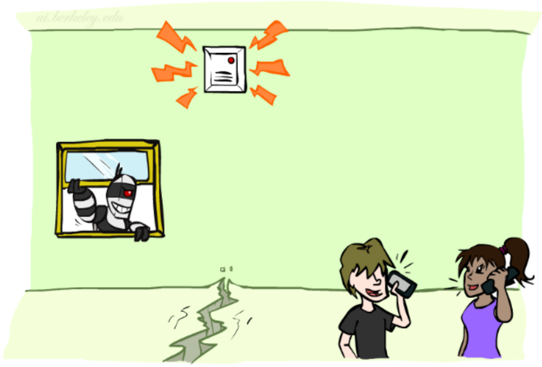


Example: Alarm Network

B	P(B)
+b	0.001
-b	0.999



E	P(E)
+e	0.002
-e	0.998



A	J	P(J|A)
+a	+j	0.9
+a	-j	0.1
-a	+j	0.05
-a	-j	0.95

A	M	P(M|A)
+a	+m	0.7
+a	-m	0.3
-a	+m	0.01
-a	-m	0.99

B	E	A	P(A|B,E)
+b	+e	+a	0.95
+b	+e	-a	0.05
+b	-e	+a	0.94
+b	-e	-a	0.06
-b	+e	+a	0.29
-b	+e	-a	0.71
-b	-e	+a	0.001
-b	-e	-a	0.999

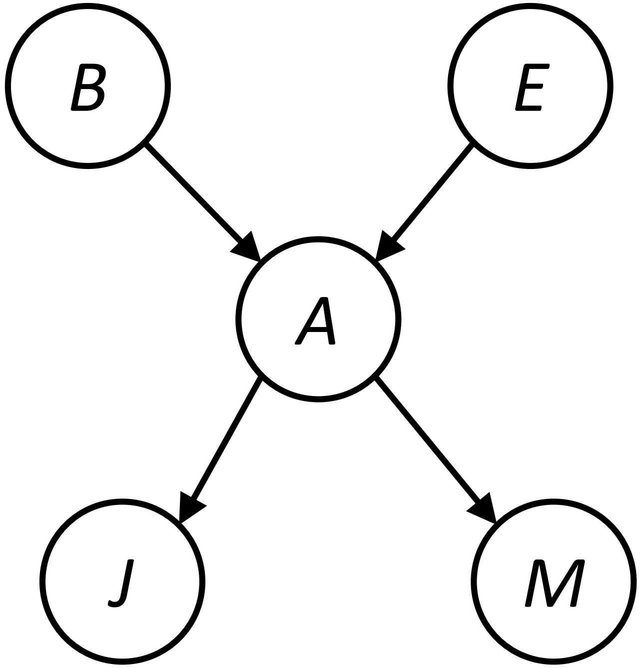
$P(-b, +j, +m, +e, +a) =$

$P(-b)\ P(+e)\ P(+a|-b, +e)\ P(+j|+a)\ P(+m|+a) =$

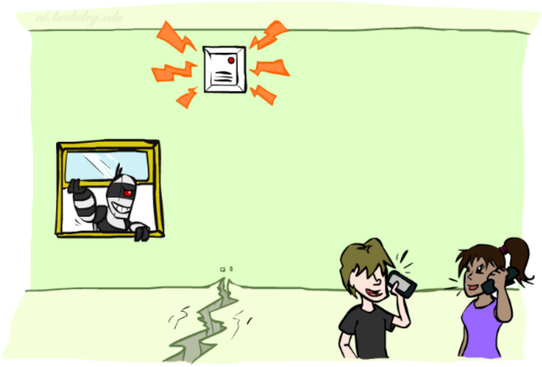
$0.999 \times 0.002 \times 0.29 \times 0.9 \times 0.7 = 0.000365$

Example: Alarm Network

B	P(B)
+b	0.001
-b	0.999



E	P(E)
+e	0.002
-e	0.998



A	J	P(J|A)
+a	+j	0.9
+a	-j	0.1
-a	+j	0.05
-a	-j	0.95

A	M	P(M|A)
+a	+m	0.7
+a	-m	0.3
-a	+m	0.01
-a	-m	0.99

B	E	A	P(A|B,E)
+b	+e	+a	0.95
+b	+e	-a	0.05
+b	-e	+a	0.94
+b	-e	-a	0.06
-b	+e	+a	0.29
-b	+e	-a	0.71
-b	-e	+a	0.001
-b	-e	-a	0.999

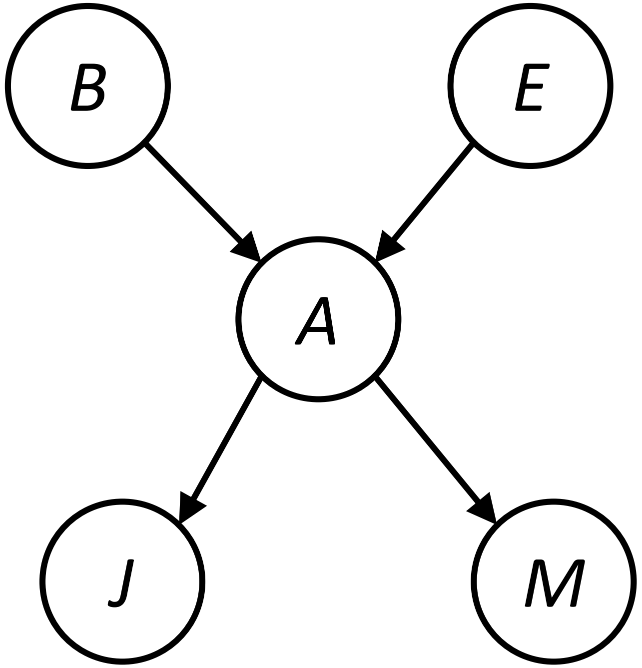
P(-b, +j, +m, +e, **-a**) =

P(-b) P(+e) P(-a|-b, +e) P(+j|-a) P(+m|-a) =

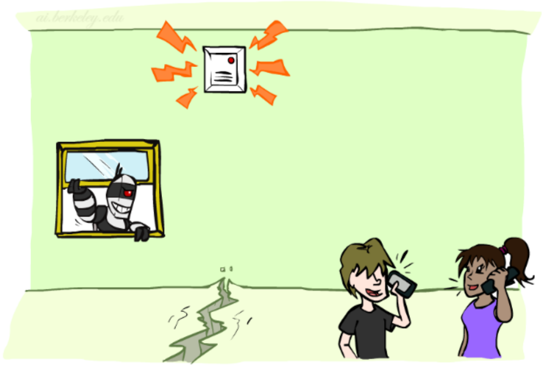
$0.999 \times 0.002 \times 0.71 \times 0.05 \times 0.01 = 7 \times 10^{-07}$

Example: Alarm Network

B	P(B)
+b	0.001
-b	0.999



E	P(E)
+e	0.002
-e	0.998



A	J	P(J|A)
+a	+j	0.9
+a	-j	0.1
-a	+j	0.05
-a	-j	0.95

A	M	P(M|A)
+a	+m	0.7
+a	-m	0.3
-a	+m	0.01
-a	-m	0.99

B	E	A	P(A|B,E)
+b	+e	+a	0.95
+b	+e	-a	0.05
+b	-e	+a	0.94
+b	-e	-a	0.06
-b	+e	+a	0.29
-b	+e	-a	0.71
-b	-e	+a	0.001
-b	-e	-a	0.999

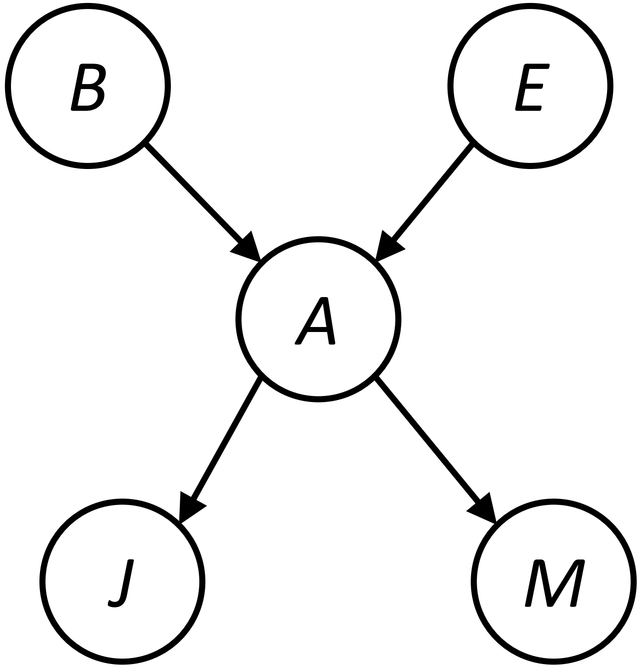
P(-b, +j, +m, **-e**, +a) =

P(-b) P(-e) P(+a|-b, -e) P(+j|+a) P(+m|+a) =

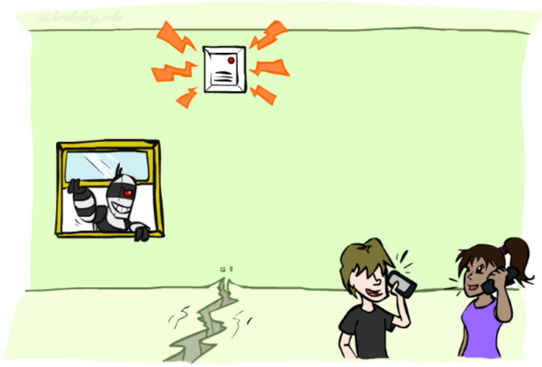
0.999 x 0.998 x 0.001 x 0.9 x 0.7 = 0.000628

Example: Alarm Network

B	P(B)
+b	0.001
-b	0.999



E	P(E)
+e	0.002
-e	0.998



A	J	P(J|A)
+a	+j	0.9
+a	-j	0.1
-a	+j	0.05
-a	-j	0.95

A	M	P(M|A)
+a	+m	0.7
+a	-m	0.3
-a	+m	0.01
-a	-m	0.99

B	E	A	P(A|B,E)
+b	+e	+a	0.95
+b	+e	-a	0.05
+b	-e	+a	0.94
+b	-e	-a	0.06
-b	+e	+a	0.29
-b	+e	-a	0.71
-b	-e	+a	0.001
-b	-e	-a	0.999

P(-b, +j, +m, -e, **-a**) =

P(-b) P(-e) P(-a|-b, -e) P(+j|-a) P(+m|-a) =

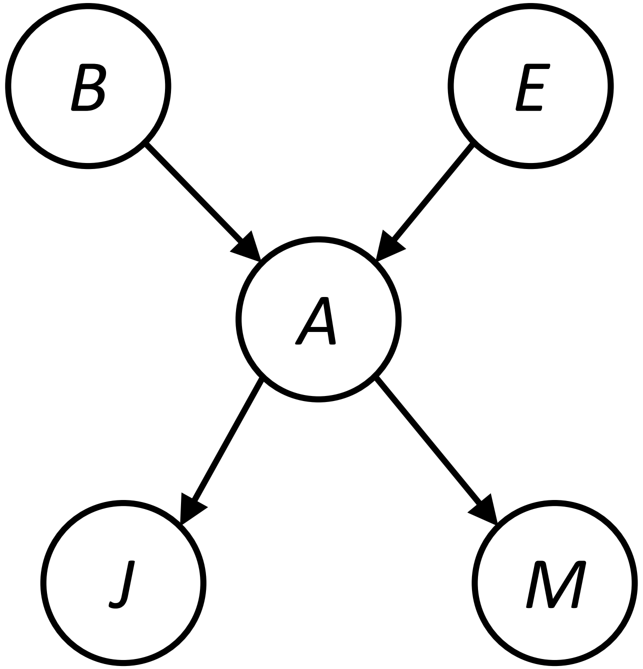
0.999 x 0.998 x 0.999 x 0.05 x 0.01 = 0.000498

Inference by Enumeration in Bayes' Net

Putting it all together:

$$P(-b \mid +j, +m) = \frac{P(-b,+j,+m)}{P(+j,+m)}$$

$$\begin{aligned} P(-b, +j, +m) = {} & P(-b, +j, +m, +e, +a) + \\ & P(-b, +j, +m, +e, -a) + \\ & P(-b, +j, +m, -e, +a) + \\ & P(-b, +j, +m, -e, -a) \\ = {} & 0.000365 + 7 \text{ x } 10^{-07} + \\ & 0.000628 + 0.000498 \\ = {} & 0.001492 \end{aligned}$$



Inference by Enumeration in Bayes' Net

P(-b, +j, +m) and P(+b, +j, +m) are the joint probabilities.

We still need to compute the conditional probabilities.

P(-b | +j, +m) = $\frac{P(-b,+j,+m)}{P(+j,+m)} = \frac{0.001492}{P(+j,+m)}$

P(+b | +j, +m) = $\frac{P(+b,+j,+m)}{P(+j,+m)} = \frac{0.000592}{P(+j,+m)}$

We normalize :

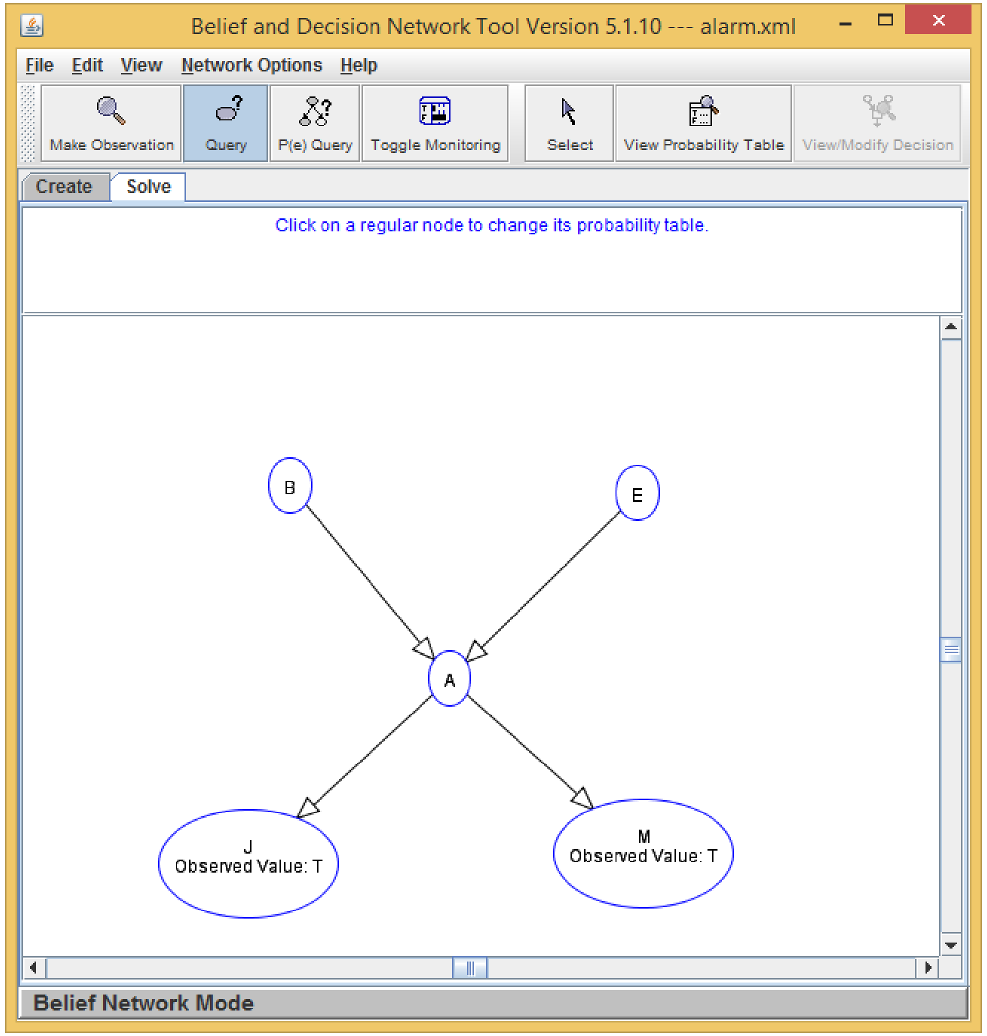
$$P(+j,+m) = P(-b,+j,+m) + P(+b,+j,+m)$$

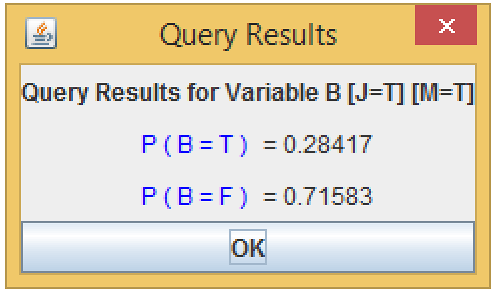
$$= 0.001492 + 0.000592 = 0.002084$$

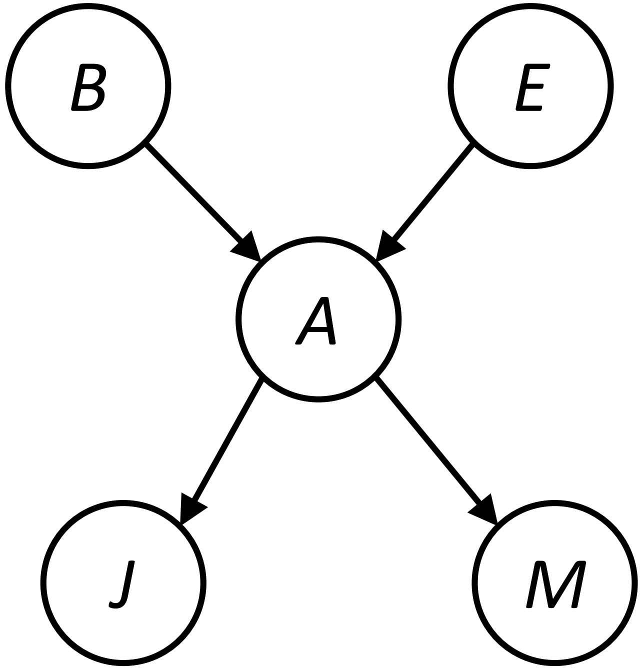
P(-b | +j, +m) = 0.716

P(+b | +j, +m) = 0.284

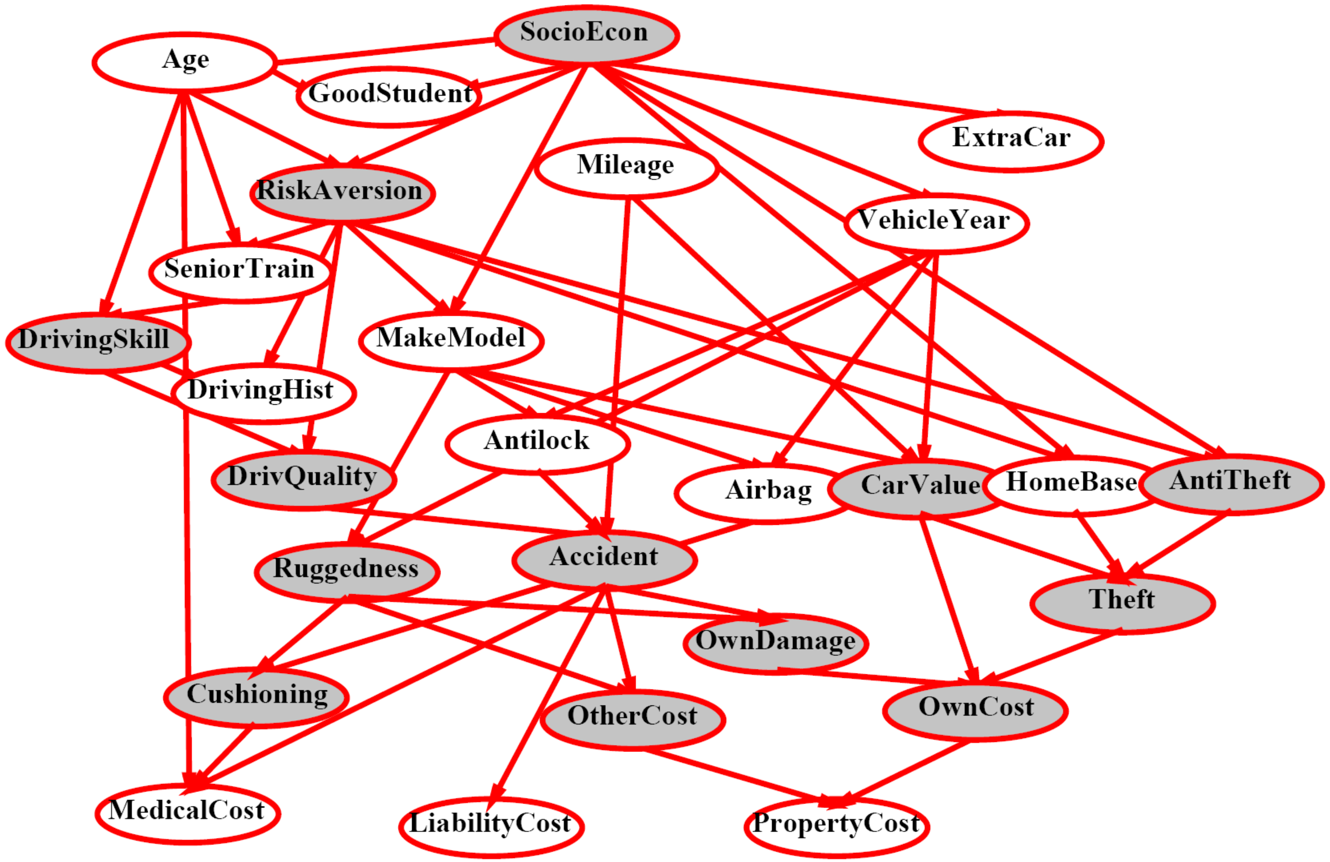
Inference in Bayes' Net – Bayes Applet aispace.org





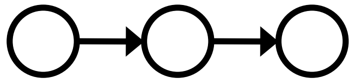


Inference by Enumeration?



Inference in Bayes' Nets

- Given unlimited time, inference by enumeration in BNs is easy
- Complexity?
  - Exponential
- There are ways to speed up enumeration
  - Build the network causally – we end up with fewer arcs



O(n)

  - variable elimination - still worst case exponential complexity
- Alternative?
  - Sampling (approximate inference)

Bayes' Nets

- ✓ Representation
- ✓ Conditional Independences
- ✓ Probabilistic Inference
- Learning Bayes' Nets from Data

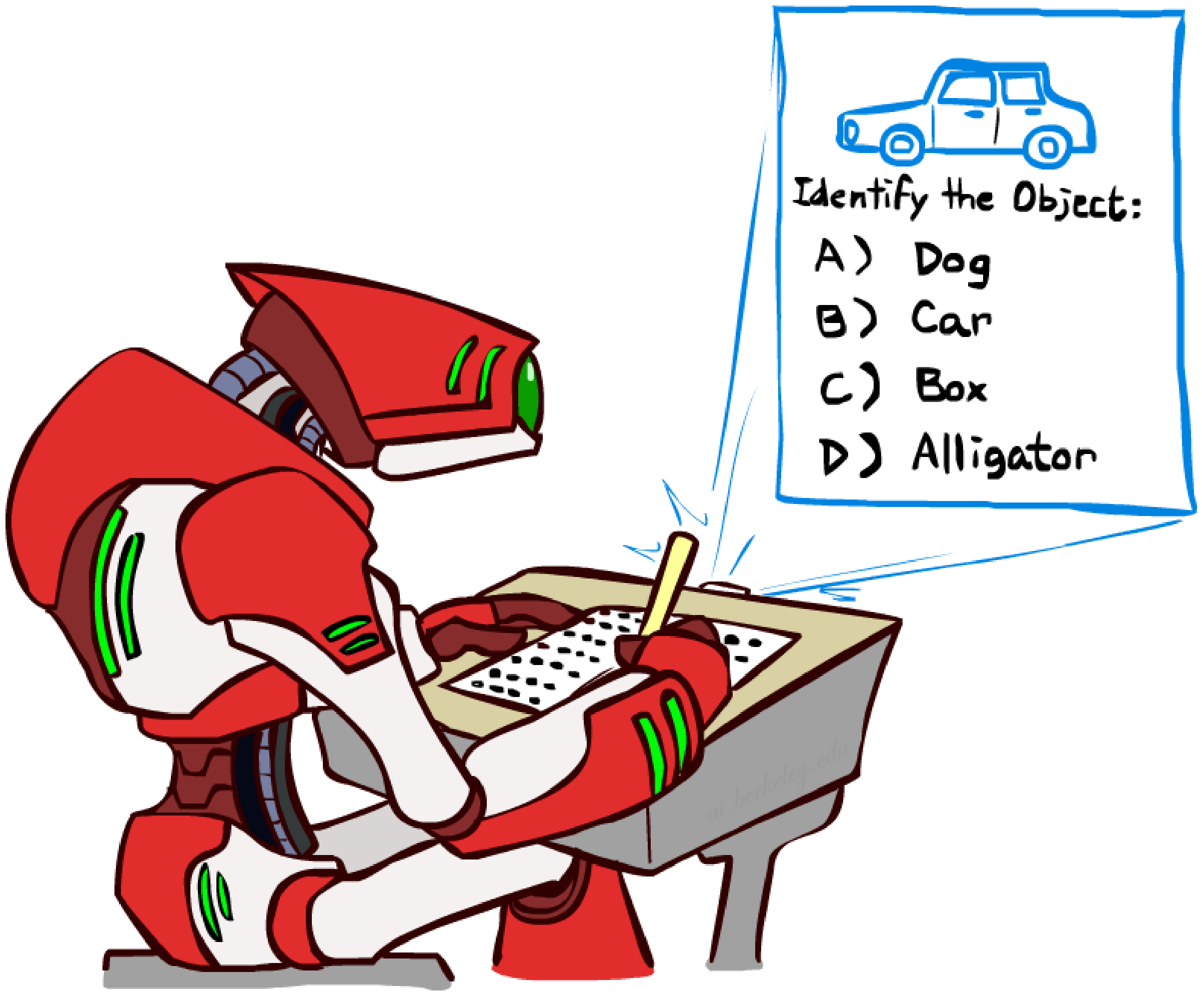
Machine Learning

- Up until now: how to use a model to make optimal decisions

- Machine learning: how to acquire a model from data / experience
  - Learning parameters (e.g. probabilities)
  - Learning structure (e.g. BN graphs – where are the arcs?)
  - Learning hidden concepts (e.g. clustering)

- Today: model-based classification with Naive Bayes

Classification

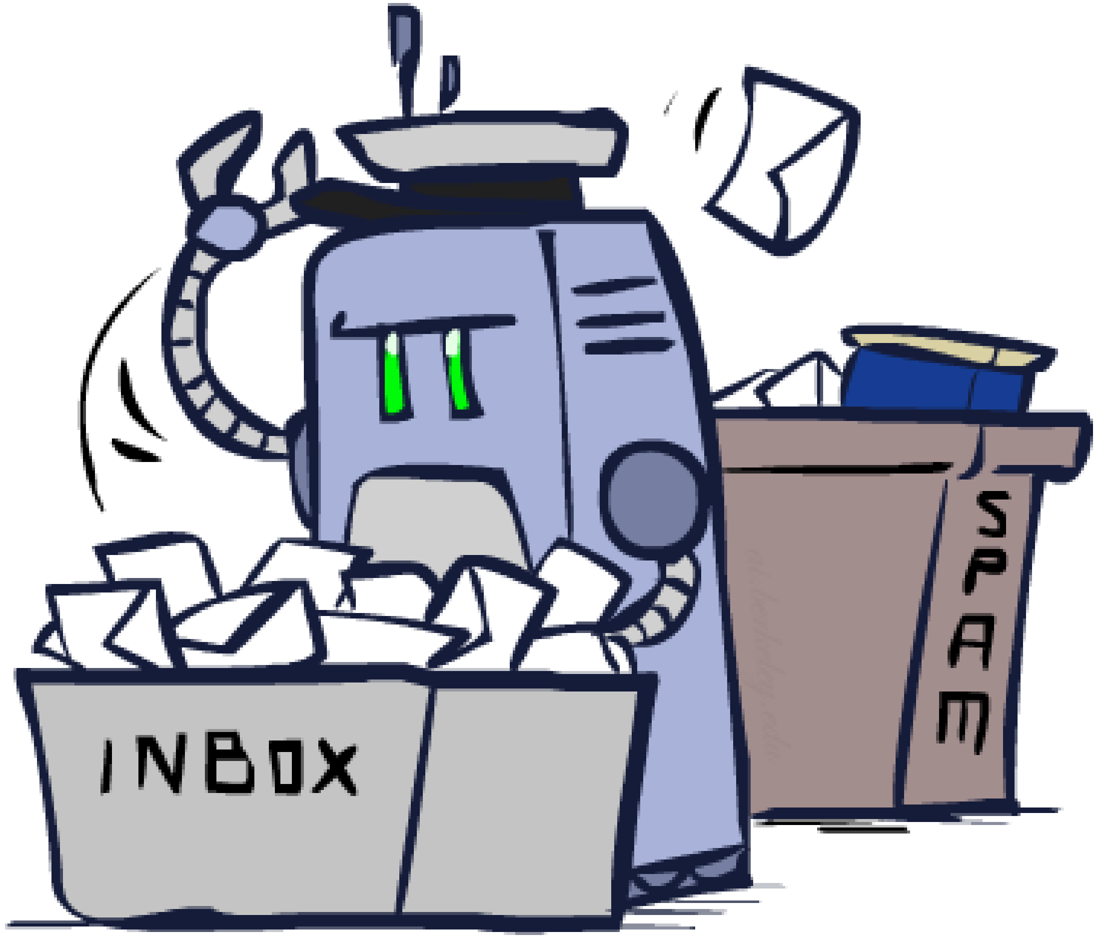


Classification Tasks

- Classification: given inputs x, predict labels (classes) y
- Examples:
  - Spam detection (input: message, classes: spam / ham)
  - OCR - Optical Character Recognition (input: images, classes: characters)
  - Medical diagnosis (input: symptoms, classes: diseases)
  - Automatic essay grading (input: document, classes: grades)
  - Fraud detection (input: account activity, classes: fraud / no fraud)
  - … many more
- Classification is an important commercial technology!

Example: Spam Filter

- Input: an email
- Output: spam/ham



Example: Spam Filter

Setup:

- Get a large collection of example emails, each labeled “spam” or “ham”
- Note: someone has to hand label all this data!
- Our goal: to learn to predict labels of new, future emails

> Dear Sir.
> First, I must solicit your confidence in this transaction, this is by virtue of its nature as being utterly confidencial and TOP SECRET. …

> Ok, I know this is blatantly OT but I'm beginning to go insane. Had an old Dell Dimension XPS sitting in the corner and decided to put it to use, I know it was working pre being stuck in the corner, but when I plugged it in, hit the power nothing happened.

Example: Spam Filter

- Features: The attributes used to make the ham / spam decision
  - Words: FREE!
  - Text Patterns: $dd, CAPS
  - Non-text: SenderInContacts
  - …

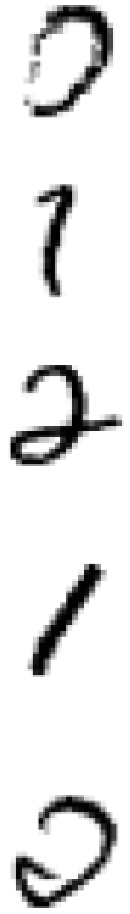
Dear Sir.
First, I must solicit your confidence in this transaction, this is by virtue of its nature as being utterly confidencial and TOP SECRET. …

Ok, I know this is blatantly OT but I'm beginning to go insane. Had an old Dell Dimension XPS sitting in the corner and decided to put it to use, I know it was working pre being stuck in the corner, but when I plugged it in, hit the power nothing happened.

Example: Digit Recognition

- Input: images / pixel grids
- Output: a digit 0-9

- Setup:
  - Get a large collection of example images, each labeled with a digit
  - Note: someone has to hand label all this data!
  - Our goal: learn to predict labels of new, future digit images



0

1

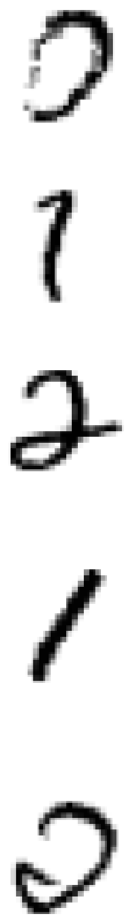
2

1

??

Example: Digit Recognition

- Features: The attributes used to make the digit decision
  - Pixels: (6,8)=ON
  - Shape Patterns: AspectRatio, NumLoops
  - …



0

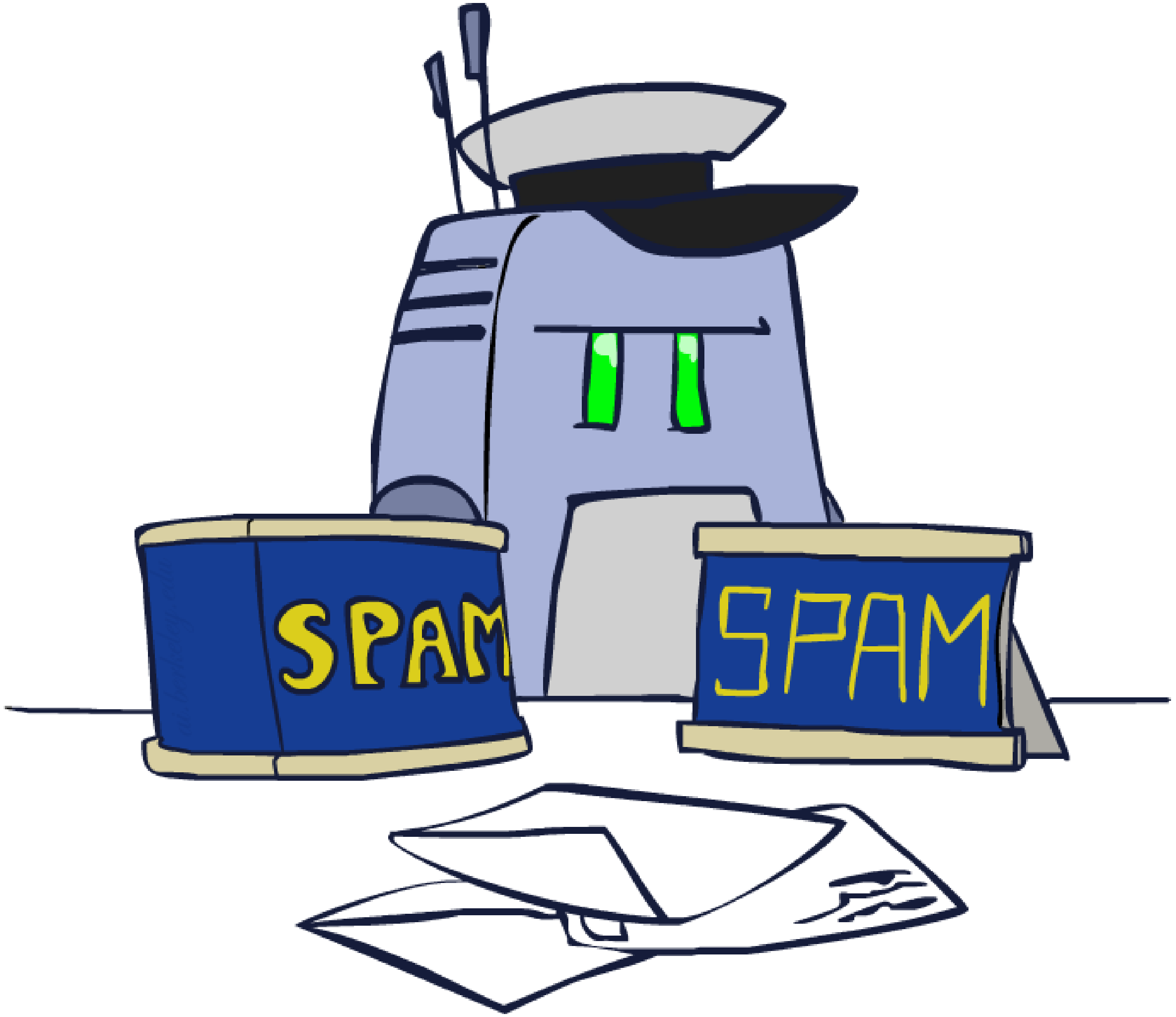
1

2

1

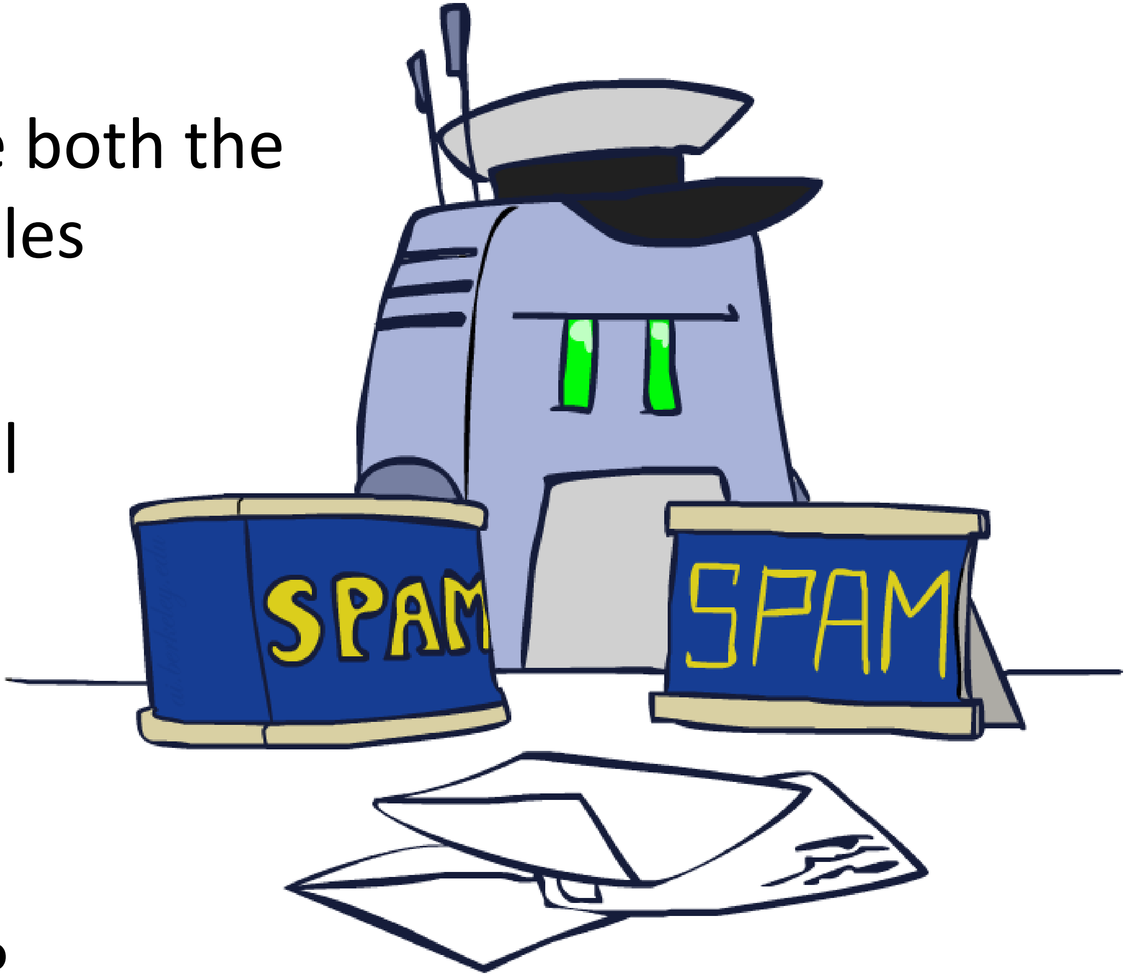
??

Model-Based Classification



Model-Based Classification

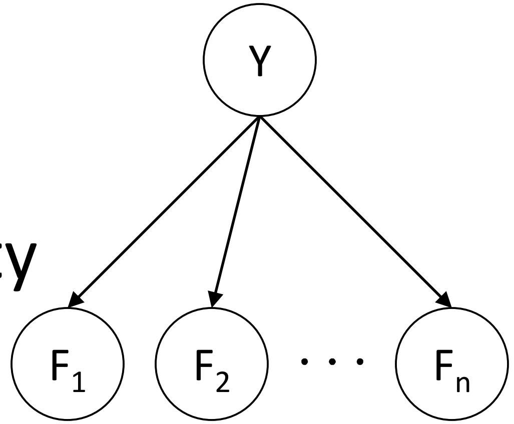
- Model-based approach
  - Build a model (e.g. Bayes' net) where both the label and features are random variables
  - Instantiate any observed features
  - Query for the distribution of the label conditioned on the features



- Challenges
  - What structure should the BN have?
  - How should we learn its parameters?

Naïve Bayes for Digits

- Naïve Bayes: Assume all features are independent effects of the label
- Simple digit recognition version:
  - One feature (variable) F_{ij} for each grid position <i,j>
  - Feature values are on / off, based on whether intensity is more or less than 0.5 in underlying image
  - Each input maps to a feature vector, e.g.



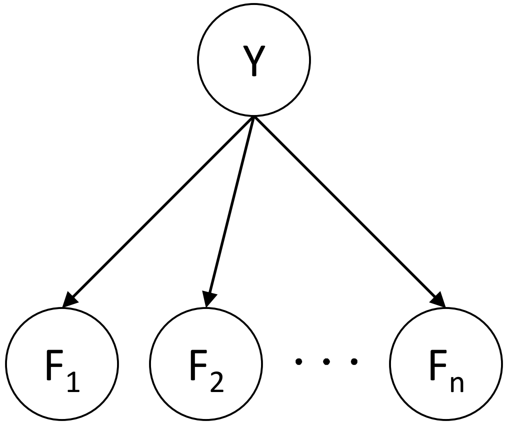
$$\rightarrow \langle F_{0,0} = 0 \;\; F_{0,1} = 0 \;\; F_{0,2} = 1 \;\; F_{0,3} = 1 \;\; F_{0,4} = 0 \;\; \ldots F_{15,15} = 0 \rangle$$

  - Here: lots of features, each is binary valued
- Naïve Bayes model: $P(Y|F_{0,0} \ldots F_{15,15}) \propto P(Y) \prod_{i,j} P(F_{i,j}|Y)$
- What do we need to learn?

General Naïve Bayes

- A general Naive Bayes model:

$$P(\mathsf{Y}, \mathsf{F}_1 \dots \mathsf{F}_n) = P(\mathsf{Y}) \prod_i P(\mathsf{F}_i | \mathsf{Y})$$



- We only have to specify how each feature depends on the class
- Total number of parameters is *linear* in n
- Model is very simplistic, but often works well anyway

Inference for Naïve Bayes

- Goal: compute posterior distribution over label variable Y
  - Step 1: get joint probability of label and evidence for each label

$$P(Y, f_1 \ldots f_n) = \begin{bmatrix} P(y_1, f_1 \ldots f_n) \\ P(y_2, f_1 \ldots f_n) \\ \vdots \\ P(y_k, f_1 \ldots f_n) \end{bmatrix} \Rightarrow \begin{bmatrix} P(y_1) \prod_i P(f_i|y_1) \\ P(y_2) \prod_i P(f_i|y_2) \\ \vdots \\ P(y_k) \prod_i P(f_i|y_k) \end{bmatrix}$$

$$\xrightarrow{+} \quad P(f_1 \ldots f_n)$$

  - Step 2: sum to get probability of evidence

$$\Downarrow$$

  - Step 3: normalize by dividing Step 1 by Step 2 $P(Y|f_1 \ldots f_n)$

General Naïve Bayes

What do we need in order to use Naïve Bayes?

- Inference method (we just saw this part)
  - Start with a bunch of probabilities: $P(Y)$ and the $P(F_i|Y)$ tables
  - Use standard inference to compute $P(Y|F_1 \ldots F_n)$
- Estimates of local conditional probability tables
  - $P(Y)$, the prior over labels
  - $P(F_i|Y)$ for each feature (evidence variable)
  - These probabilities are collectively called the *parameters* of the model: θ
  - Up until now, we assumed these appeared by magic, but…
  - …they typically come from training data counts: we'll look at this soon

Summary

- Bayes rule lets us do diagnostic queries with causal probabilities
- The naïve Bayes assumption takes all features to be independent given the class label
- We can build classifiers out of a naïve Bayes model using training data