Supervised Learning

Bayesian Learning

Probability Overview

- Events
  - discrete random variables, continuous random variables, compound events
- Axioms of probability
  - What defines a reasonable theory of uncertainty
- Independent events
- Conditional probabilities
- Bayes rule and beliefs
- Joint probability distribution
- Expectations
- Independence, Conditional independence

Random Variables

- Informally, A is a random variable if
  - A denotes something about which we are uncertain
  - perhaps the outcome of a randomized experiment

- Examples

  A = True if a randomly drawn person from our class is female

  A = The hometown of a randomly drawn person from our class

  A = True if two randomly drawn persons from our class have same birthday

- Define P(A) as "the fraction of possible worlds in which A is true" or "the fraction of times A holds, in repeated runs of the random experiment"
  - the set of possible worlds is called the sample space, S
  - A random variable A is a function defined over S

$$A: S \rightarrow \{0,1\}$$

A little formalism

More formally, we have

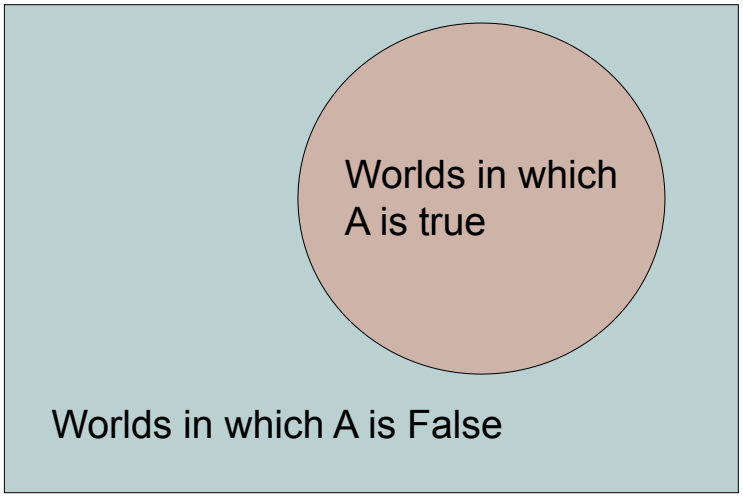
- a sample space S (e.g., set of students in our class)
  - aka the set of possible worlds

- a random variable is a function defined over the sample space
  - Gender: S → { m, f }
  - Height: S → Reals
- an event is a subset of S
  - e.g., the subset of S for which Gender=f
  - e.g., the subset of S for which (Gender=m) AND (eyeColor=blue)
- we're often interested in probabilities of specific events
- and of specific events conditioned on other specific events

Visualizing A

Sample space of all possible worlds

Its area is 1



P(A) = Area of reddish oval

The Axioms of Probability

- 0 <= P(A) <= 1
- P(True) = 1
- P(False) = 0
- P(A or B) = P(A) + P(B) - P(A and B)

[di Finetti 1931]:

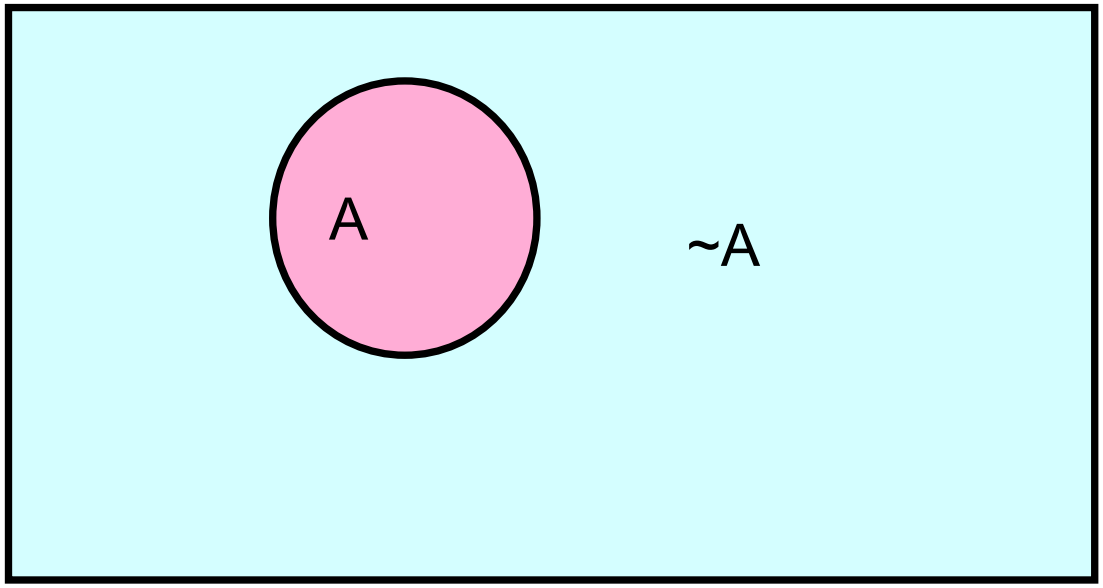
when gambling based on “uncertainty formalism A” you can be exploited by an opponent

iff

your uncertainty formalism A violates these axioms

Elementary Probability in Pictures

- P(~A) + P(A) = 1



A useful theorem

- 0 <= P(A) <= 1, P(True) = 1, P(False) = 0,
  P(A or B) = P(A) + P(B) - P(A and B)

➔ P(A) = P(A ^ B) + P(A ^ ~B)

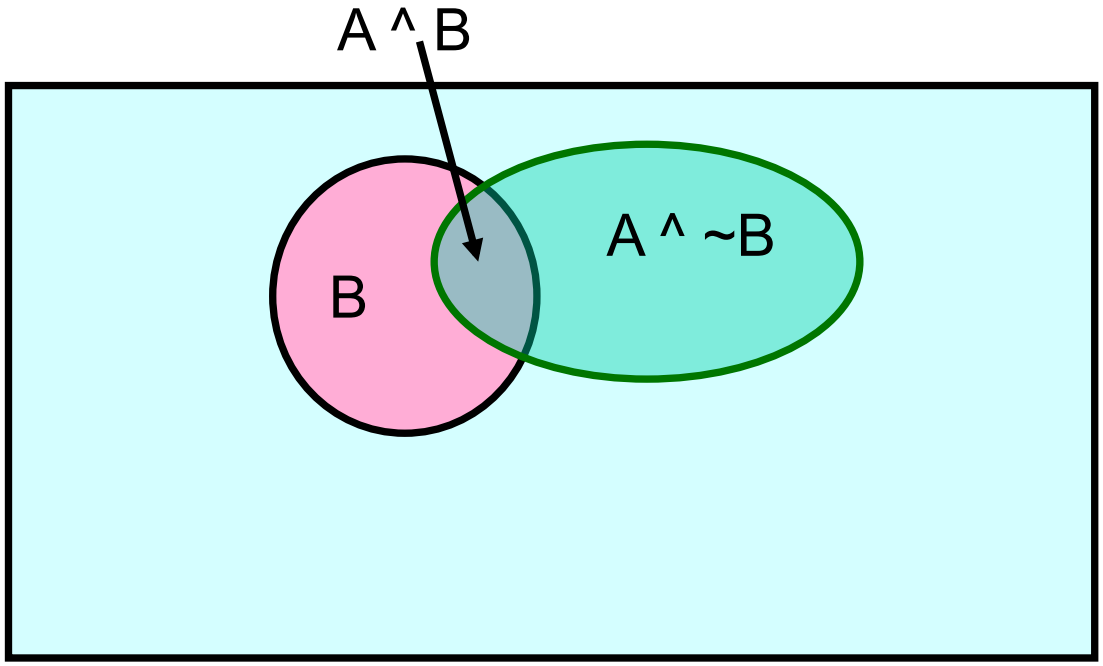
A = [A and (B or ~B)] = [(A and B) or (A and ~B)]

P(A) = P(A and B) + P(A and ~B) – P((A and B) and (A and ~B))

P(A) = P(A and B) + P(A and ~B) – P(A and B and A and ~B)

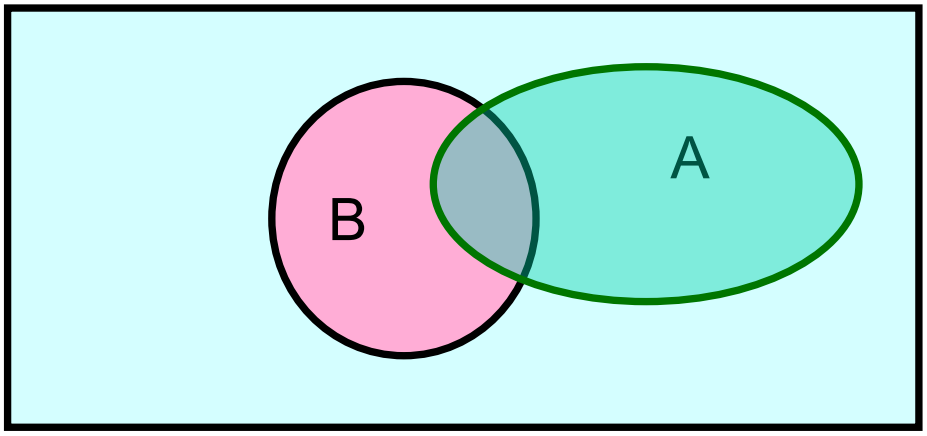
Elementary Probability in Pictures

- P(A) = P(A ^ B) + P(A ^ ~B)



Definition of Conditional Probability

$$P(A|B) = \frac{P(A \wedge B)}{P(B)}$$



Definition of Conditional Probability

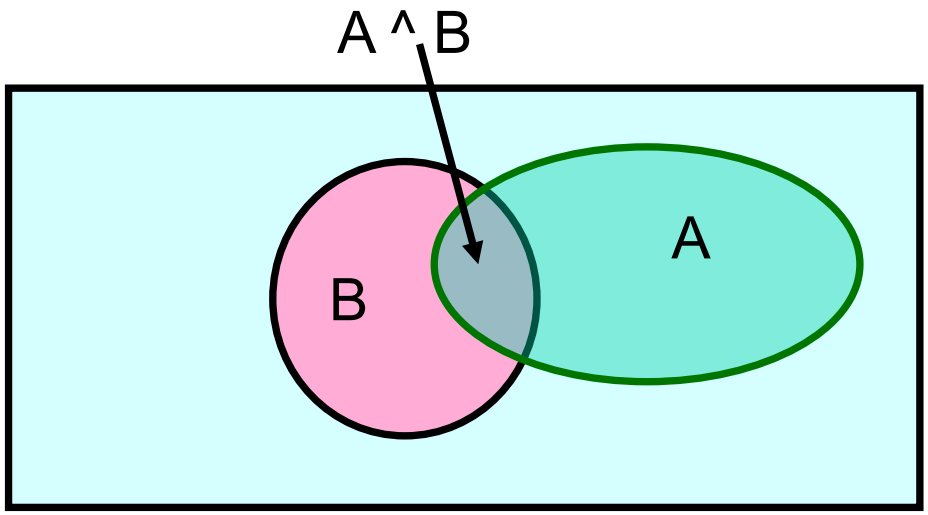
$$P(A|B) = \frac{P(A \wedge B)}{P(B)}$$

Corollary: The Chain Rule

$$P(A \wedge B) = P(A|B)\,P(B)$$

Bayes Rule

- let's write 2 expressions for P(A ^ B)



$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$ Bayes' rule

we call P(A) the "prior"

and P(A|B) the "posterior"

Bayes, Thomas (1763) An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London,* **53:370-418**

…by no means merely a curious speculation in the doctrine of chances, but necessary to be solved in order to a sure foundation for all our reasonings concerning past facts, and what is likely to be hereafter…. necessary to be considered by any that would give a clear account of the strength of *analogical* or *inductive reasoning…*

Other Forms of Bayes Rule

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B \mid A)P(A) + P(B \mid \sim A)P(\sim A)}$$

$$P(A \mid B \wedge X) = \frac{P(B \mid A \wedge X)P(A \wedge X)}{P(B \wedge X)}$$

Applying Bayes Rule

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B \mid A)P(A) + P(B \mid \sim A)P(\sim A)}$$

A = you have the flu, B = you just coughed

Assume:
P(A) = 0.05
P(B|A) = 0.80
P(B| ~A) = 0.2

what is P(flu | cough) = P(A|B)?

what does all this have to do with function approximation?

The Joint Distribution

Example: Boolean variables A, B, C

Recipe for making a joint distribution of M variables:

The Joint Distribution

Example: Boolean variables A, B, C

Recipe for making a joint distribution of M variables:

1. Make a truth table listing all combinations of values of your variables (if there are M Boolean variables then the table will have 2^M rows).

A	B	C
0	0	0
0	0	1
0	1	0
0	1	1
1	0	0
1	0	1
1	1	0
1	1	1

The Joint Distribution

Example: Boolean variables A, B, C

Recipe for making a joint distribution of M variables:

1. Make a truth table listing all combinations of values of your variables (if there are M Boolean variables then the table will have 2^M rows).
2. For each combination of values, say how probable it is.

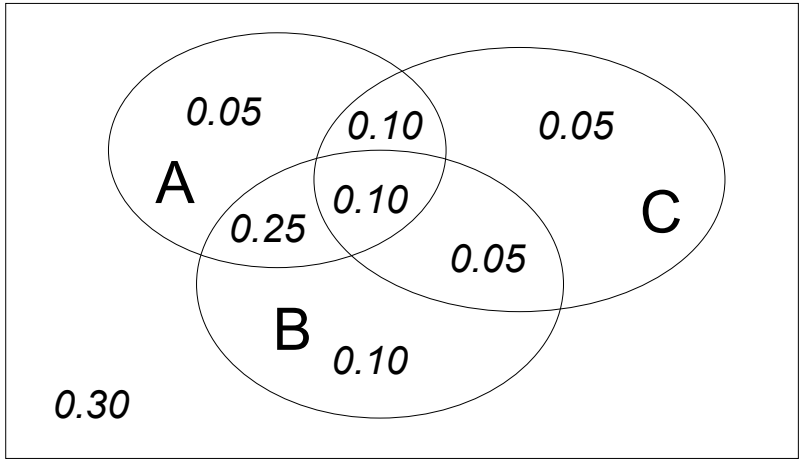
A	B	C	Prob
0	0	0	0.30
0	0	1	0.05
0	1	0	0.10
0	1	1	0.05
1	0	0	0.05
1	0	1	0.10
1	1	0	0.25
1	1	1	0.10

The Joint Distribution

Recipe for making a joint distribution of M variables:

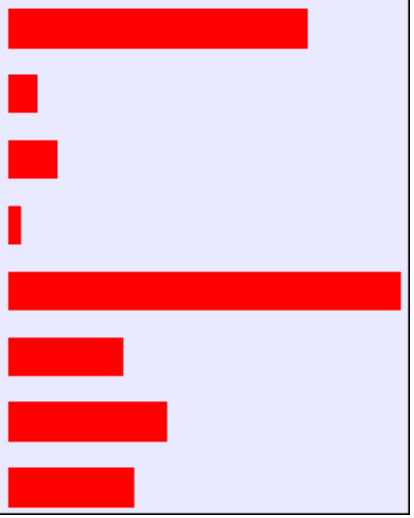
1. Make a truth table listing all combinations of values of your variables (if there are M Boolean variables then the table will have 2^M rows).
2. For each combination of values, say how probable it is.
3. If you subscribe to the axioms of probability, those numbers must sum to 1.

A	B	C	Prob
0	0	0	0.30
0	0	1	0.05
0	1	0	0.10
0	1	1	0.05
1	0	0	0.05
1	0	1	0.10
1	1	0	0.25
1	1	1	0.10



Using the Joint Distribution

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933



One you have the JD you can ask for the probability of any logical expression involving your attribute

$$P(E) = \sum_{\text{rows matching } E} P(\text{row})$$

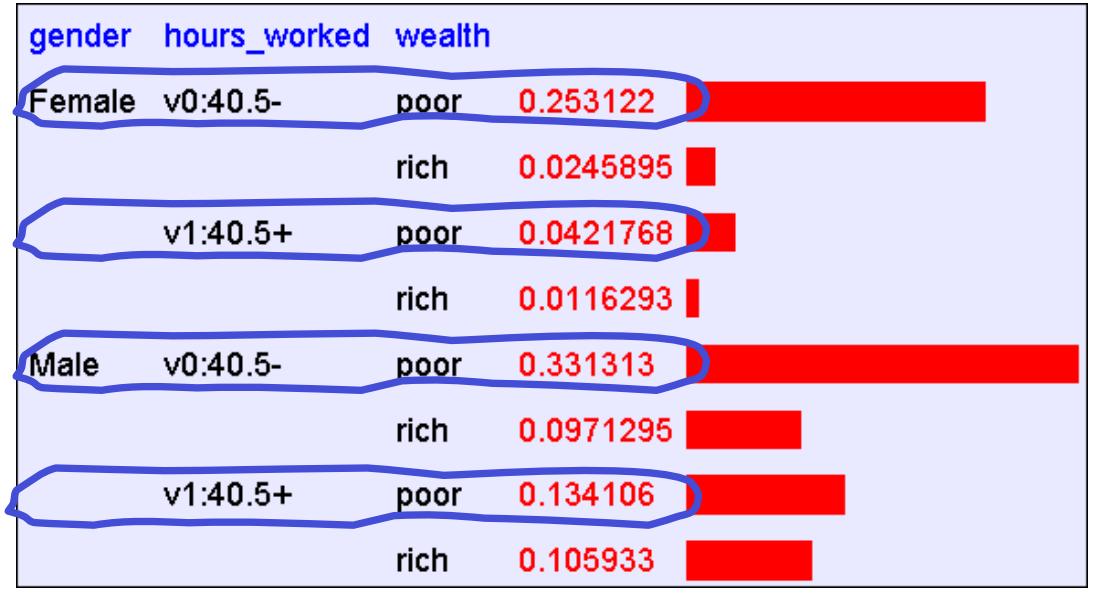
Using the Joint

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

P(Poor Male) = 0.4654

$$P(E) = \sum_{\text{rows matching } E} P(\text{row})$$

Using the Joint



gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

P(Poor) = 0.7604

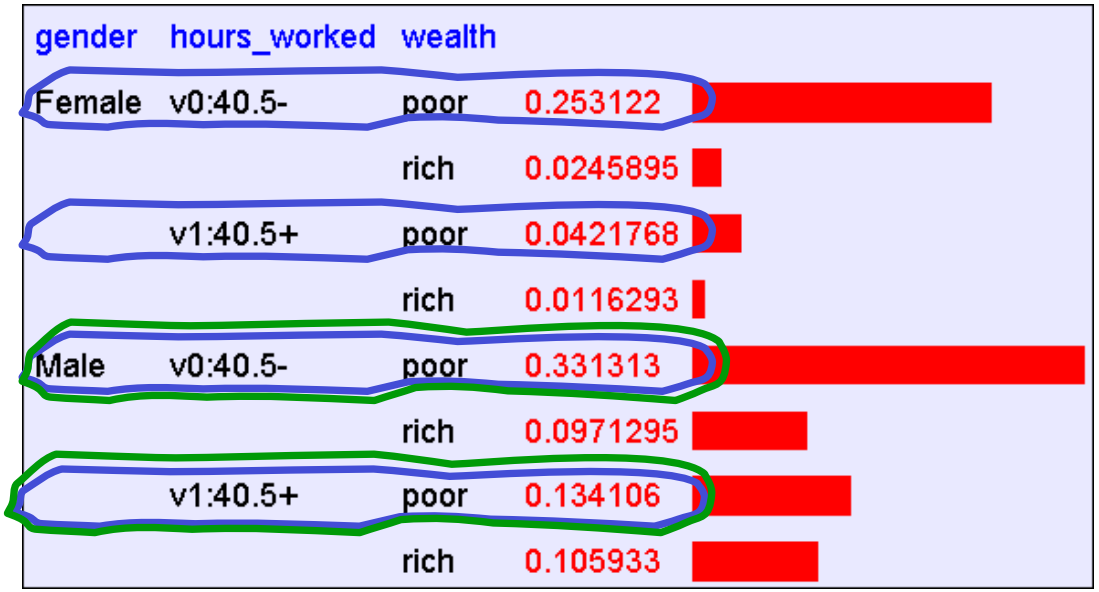
$$P(E) = \sum_{\text{rows matching } E} P(\text{row})$$

Inference with the Joint

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

$$P(E_1 \mid E_2) = \frac{P(E_1 \wedge E_2)}{P(E_2)} = \frac{\sum_{\text{rows matching } E_1 \text{ and } E_2} P(\text{row})}{\sum_{\text{rows matching } E_2} P(\text{row})}$$

Inference with the Joint



$$P(E_1 \mid E_2) = \frac{P(E_1 \wedge E_2)}{P(E_2)} = \frac{\sum_{\text{rows matching } E_1 \text{ and } E_2} P(\text{row})}{\sum_{\text{rows matching } E_2} P(\text{row})}$$

P(Male | Poor) = 0.4654 / 0.7604 = 0.612

Learning and the Joint Distribution

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

Suppose we want to learn the function f: <G, H> → W

Equivalently, P(W | G, H)

Solution: learn joint distribution from data, calculate P(W | G, H)

e.g., P(W=rich | G = female, H = 40.5-) =

[A. Moore]

sounds like the solution to
learning F: X → Y,
or P(Y | X).

Are we done?

sounds like the solution to
learning F: X → Y,
or P(Y | X).

Main problem: learning P(Y|X)
can require more data than we have

consider learning Joint Dist. with 100 attributes
# of rows in this table?
# of people on earth?
fraction of rows with 0 training examples?

What to do?

1. Be smart about how we estimate probabilities from sparse data
   - maximum likelihood estimates
   - maximum a posteriori estimates

2. Be smart about how to represent joint distributions
   - Bayes networks, graphical models

1. Be smart about how we estimate probabilities

Estimating Probability of Heads

- I show you the above coin X, and hire you to estimate the probability that it will turn up heads $(X = 1)$ or tails $(X = 0)$
- You flip it repeatedly, observing
  - it turns up heads α_1 times
  - it turns up tails α_0 times
- Your estimate for $P(X = 1)$ is....?

Estimating θ = P(X=1)

X=1 X=0

Test A:

100 flips: 51 Heads (X=1), 49 Tails (X=0)

Test B:

3 flips: 2 Heads (X=1), 1 Tails (X=0)

Estimating θ = P(X=1)

X=1 X=0

Case C: (online learning)

- keep flipping, want single learning algorithm that gives reasonable estimate after each flip

Principles for Estimating Probabilities

Principle 1 (maximum likelihood):

- choose parameters θ that maximize **P(data | θ)**
- e.g., $$\hat{\theta}^{MLE} = \frac{\alpha_1}{\alpha_1 + \alpha_0}$$

Principle 2 (maximum a posteriori prob.):

- choose parameters θ that maximize **P(θ | data)**
- e.g.

$$\hat{\theta}^{MAP} = \frac{\alpha_1 + \#\text{hallucinated\_1s}}{(\alpha_1 + \#\text{hallucinated\_1s}) + (\alpha_0 + \#\text{hallucinated\_0s})}$$

Maximum Likelihood Estimation

X=1 X=0

P(X=1) = θ P(X=0) = (1-θ)

Data D:

Flips produce data D with α_1 heads, α_0 tails

- flips are independent, identically distributed 1's and 0's (Bernoulli)
- α_1 and α_0 are counts that sum these outcomes (Binomial)

$$P(D|\theta) = P(\alpha_1, \alpha_0|\theta) = \theta^{\alpha_1}(1-\theta)^{\alpha_0}$$

Maximum Likelihood Estimate for Θ

$$\begin{aligned}\hat{\theta} &= \arg\max_{\theta} \ln P(\mathcal{D} \mid \theta) \\ &= \arg\max_{\theta} \ln \theta^{\alpha_H}(1-\theta)^{\alpha_T}\end{aligned}$$

- Set derivative to zero: $\frac{d}{d\theta} \ln P(\mathcal{D} \mid \theta) = 0$

[C. Guestrin]

$$\hat{\theta} = \arg\max_{\theta} \ln P(D|\theta)$$

$$= \arg\max_{\theta} \ln \left[\theta^{\alpha_1}(1-\theta)^{\alpha_0}\right]$$

- Set derivative to zero: $\frac{d}{d\theta} \ln P(\mathcal{D} \mid \theta) = 0$

hint: $\frac{\partial \ln \theta}{\partial \theta} = \frac{1}{\theta}$

Summary:
Maximum Likelihood Estimate

X=1 X=0

P(X=1) = θ
P(X=0) = 1-θ
(Bernoulli)

- Each flip yields boolean value for X

$$X \sim \text{Bernoulli: } P(X) = \theta^X (1-\theta)^{(1-X)}$$

- Data set D of independent, identically distributed (iid) flips produces α_1 ones, α_0 zeros (Binomial)

$$P(D|\theta) = P(\alpha_1, \alpha_0|\theta) = \theta^{\alpha_1}(1-\theta)^{\alpha_0}$$

$$\hat{\theta}^{MLE} = \text{argmax}_\theta \, P(D|\theta) = \frac{\alpha_1}{\alpha_1+\alpha_0}$$

Principles for Estimating Probabilities

Principle 1 (maximum likelihood):

- choose parameters θ that maximize $P(\text{data} \mid \theta)$

Principle 2 (maximum a posteriori prob.):

- choose parameters θ that maximize

$$P(\theta \mid \text{data}) = \frac{P(\text{data} \mid \theta)\, P(\theta)}{P(\text{data})}$$

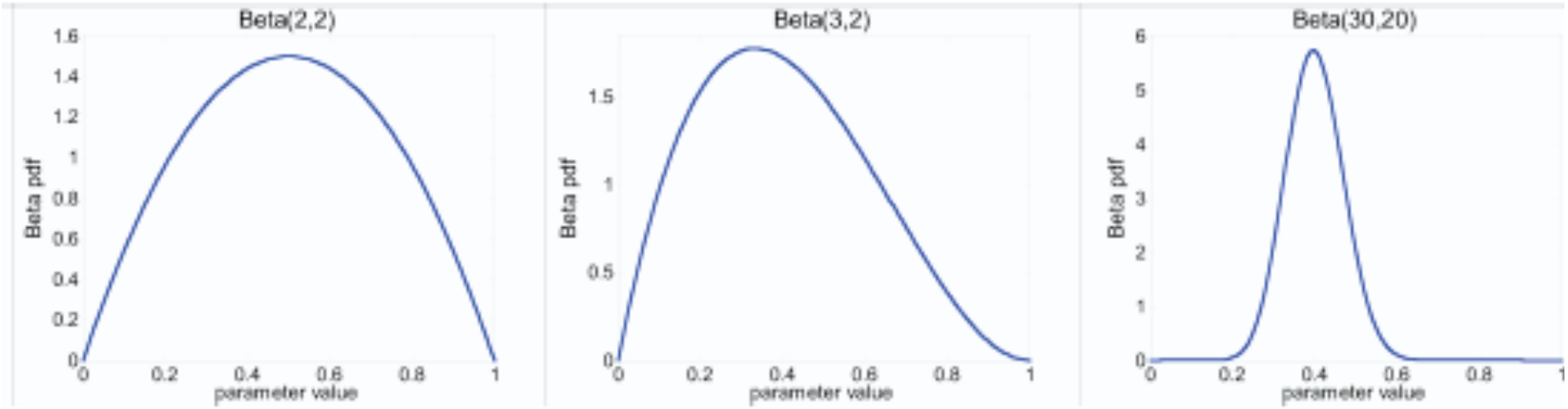
Beta prior distribution – P(θ)

$$P(\theta) = \frac{\theta^{\beta_H - 1}(1-\theta)^{\beta_T - 1}}{B(\beta_H, \beta_T)} \sim Beta(\beta_H, \beta_T)$$

- Likelihood function: $P(\mathcal{D} \mid \theta) = \theta^{\alpha_H}(1-\theta)^{\alpha_T}$
- Posterior: $P(\theta \mid \mathcal{D}) \propto P(\mathcal{D} \mid \theta)P(\theta)$

Beta prior distribution – P(θ)

$$P(\theta) = \frac{\theta^{\beta_H - 1}(1-\theta)^{\beta_T - 1}}{B(\beta_H, \beta_T)} \sim Beta(\beta_H, \beta_T)$$



[C. Guestrin]

Eg. 1 Coin flip problem

Likelihood is ~ Binomial

$$P(\mathcal{D} \mid \theta) = \theta^{\alpha_H}(1-\theta)^{\alpha_T}$$

If prior is Beta distribution,

$$P(\theta) = \frac{\theta^{\beta_H - 1}(1-\theta)^{\beta_T - 1}}{B(\beta_H, \beta_T)} \sim Beta(\beta_H, \beta_T)$$

Then posterior is Beta distribution

$$P(\theta|D) \sim Beta(\alpha_H + \beta_H, \alpha_H + \beta_H)$$

and MAP estimate is therefore

$$\hat{\theta}^{MAP} = \frac{\alpha_H + \beta_H - 1}{(\alpha_H + \beta_H - 1) + (\alpha_T + \beta_T - 1)}$$

Eg. 2 Dice roll problem (6 outcomes instead of 2)

Likelihood is ~ Multinomial($\theta = \{\theta_1, \theta_2, \ldots, \theta_k\}$)

$$P(\mathcal{D} \mid \theta) = \theta_1^{\alpha_1}\theta_2^{\alpha_2} \ldots \theta_k^{\alpha_k}$$

If prior is Dirichlet distribution,

$$P(\theta) = \frac{\theta_1^{\beta_1 - 1}\ \theta_2^{\beta_2 - 1} \ldots \theta_k^{\beta_k - 1}}{B(\beta_1, \ldots, \beta_k)} \sim \text{Dirichlet}(\beta_1, \ldots, \beta_k)$$

Then posterior is Dirichlet distribution

$$P(\theta|D) \sim \text{Dirichlet}(\beta_1 + \alpha_1, \ldots, \beta_k + \alpha_k)$$

and MAP estimate is therefore

$$\hat{\theta}_i^{MAP} = \frac{\alpha_i + \beta_i - 1}{\sum_{j=1}^{k}(\alpha_j + \beta_j - 1)}$$

Some terminology

- Likelihood function: $P(\text{data} \mid \theta)$
- Prior: $P(\theta)$
- Posterior: $P(\theta \mid \text{data})$

- Conjugate prior: $P(\theta)$ is the conjugate prior for likelihood function $P(\text{data} \mid \theta)$ if the forms of $P(\theta)$ and $P(\theta \mid \text{data})$ are the same.

Recap...

Two Principles for Estimating Parameters

- Maximum Likelihood Estimate (MLE): choose θ that maximizes probability of observed data $\mathcal{D}$

$$\hat{\theta} = \arg\max_{\theta} P(\mathcal{D} \mid \theta)$$

- Maximum a Posteriori (MAP) estimate: choose θ that is most probable given prior probability and the data

$$\hat{\theta} = \arg\max_{\theta} P(\theta \mid \mathcal{D})$$

$$= \arg\max_{\theta} = \frac{P(\mathcal{D} \mid \theta) P(\theta)}{P(\mathcal{D})}$$

Maximum Likelihood Estimate

X=1 X=0

P(X=1) = θ

P(X=0) = 1-θ

(Bernoulli)

- Each flip yields boolean value for X

$$X \sim \text{Bernoulli: } P(X) = \theta^X (1-\theta)^{(1-X)}$$

- Data set D of independent, identically distributed (iid) flips produces α_1 ones, α_0 zeros

$$P(D|\theta) = P(\alpha_1, \alpha_0|\theta) = \theta^{\alpha_1}(1-\theta)^{\alpha_0}$$

$$\hat{\theta}^{MLE} = \arg\max_{\theta} P(D|\theta) = \frac{\alpha_1}{\alpha_1 + \alpha_0}$$

Maximum A Posteriori (MAP) Estimate

- Data set D of independent, identically distributed (iid) flips produces α_1 ones, α_0 zeros

$$P(D|\theta) = P(\alpha_1, \alpha_0|\theta) = \theta^{\alpha_1}(1-\theta)^{\alpha_0}$$

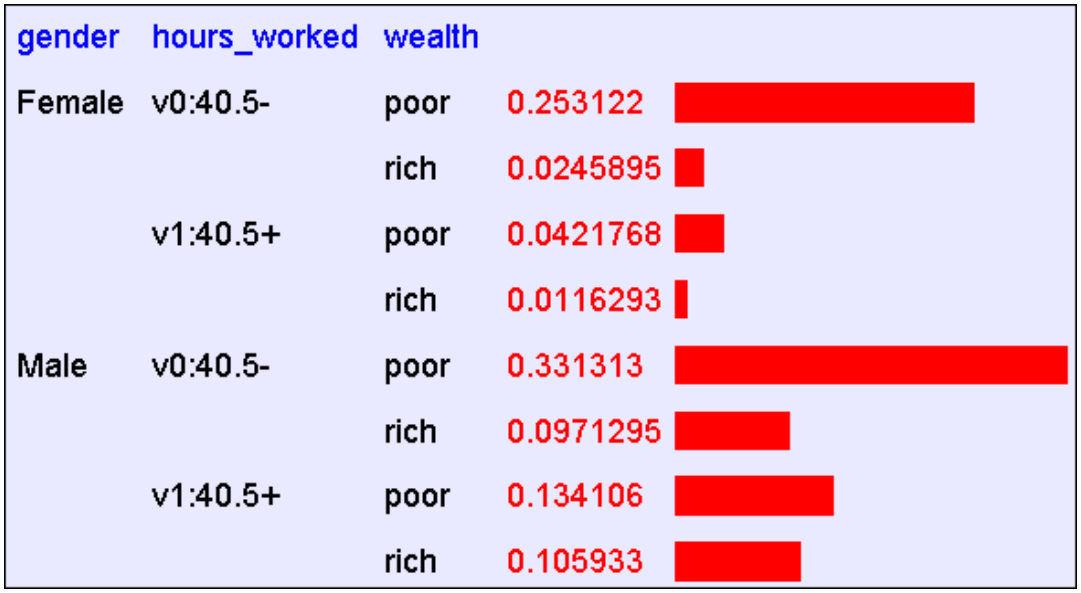
- Assume prior $P(\theta) = Beta(\beta_1, \beta_0) = \frac{1}{B(\beta_1,\beta_0)}\,\theta^{\beta_1-1}(1-\theta)^{\beta_0-1}$
- Then

$$\hat{\theta}^{MAP} = \arg\max_{\theta} P(D|\theta)P(\theta) = \frac{\alpha_1+\beta_1-1}{(\alpha_1+\beta_1-1)+(\alpha_0+\beta_0-1)}$$

(like MLE, but hallucinating β_1-1 additional heads, β_0-1 additional tails)

Let's learn classifiers by learning P(Y|X)

Consider Y=Wealth, X=<Gender, HoursWorked>



Gender	HrsWorked	P(rich | G,HW)	P(poor | G,HW)
F	<40.5	.09	.91
F	>40.5	.21	.79
M	<40.5	.23	.77
M	>40.5	.38	.62

How many parameters must we estimate?

Suppose $X = \langle X_1, \ldots X_n \rangle$
where X_i and Y are boolean RV's

Gender	HrsWorked	P(rich | G,HW)	P(poor | G,HW)
F	<40.5	.09	.91
F	>40.5	.21	.79
M	<40.5	.23	.77
M	>40.5	.38	.62

To estimate $P(Y| X_1, X_2, \ldots X_n)$

If we have 30 boolean X_i's: $P(Y \mid X_1, X_2, \ldots X_{30})$

Bayes Rule

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

Which is shorthand for:

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

Equivalently:

$$(\forall i,j) P(Y = y_i | X = x_j) = \frac{P(X = x_j | Y = y_i) P(Y = y_i)}{\sum_k P(X = x_j | Y = y_k) P(Y = y_k)}$$

Can we reduce params using Bayes Rule?

Suppose X =<X_1,… X_n>
where X_i and Y are boolean RV's

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

How many parameters to define P(X_1,… X_n | Y)?

How many parameters to define P(Y)?

Naïve Bayes

Naïve Bayes assumes

$$P(X_1 \ldots X_n|Y) = \prod_i P(X_i|Y)$$

i.e., that X_i and X_j are conditionally independent given Y, for all $i \neq j$

Conditional Independence

Definition: 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

$$(\forall i, j, k) P(X = x_i | Y = y_j, Z = z_k) = P(X = x_i | Z = z_k)$$

Which we often write

$$P(X|Y, Z) = P(X|Z)$$

E.g.,

$$P(Thunder|Rain, Lightning) = P(Thunder|Lightning)$$

Naïve Bayes uses assumption that the X_i are conditionally independent, given Y. E.g., $P(X_1|X_2, Y) = P(X_1|Y)$

Given this assumption, then:

$$P(X_1, X_2|Y) =$$

Naïve Bayes uses assumption that the X_i are conditionally independent, given Y. E.g., $P(X_1|X_2, Y) = P(X_1|Y)$

Given this assumption, then:

$$P(X_1, X_2|Y) = P(X_1|X_2, Y)P(X_2|Y)$$
$$= P(X_1|Y)P(X_2|Y)$$

in general: $P(X_1...X_n|Y) = \prod_i P(X_i|Y)$

Naïve Bayes uses assumption that the X_i are conditionally independent, given Y. E.g., $P(X_1|X_2, Y) = P(X_1|Y)$

Given this assumption, then:

$$P(X_1, X_2|Y) = P(X_1|X_2, Y)P(X_2|Y)$$
$$= P(X_1|Y)P(X_2|Y)$$

in general: $P(X_1...X_n|Y) = \prod_i P(X_i|Y)$

How many parameters to describe $P(X_1...X_n|Y)$? $P(Y)$?

- Without conditional indep assumption?
- With conditional indep assumption?

Naïve Bayes in a Nutshell

Bayes rule:

$$P(Y = y_k | X_1 \dots X_n) = \frac{P(Y = y_k) P(X_1 \dots X_n | Y = y_k)}{\sum_j P(Y = y_j) P(X_1 \dots X_n | Y = y_j)}$$

Assuming conditional independence among X_i's:

$$P(Y = y_k | X_1 \dots X_n) = \frac{P(Y = y_k) \prod_i P(X_i | Y = y_k)}{\sum_j P(Y = y_j) \prod_i P(X_i | Y = y_j)}$$

So, to pick most probable Y for $X^{new} = < X_1, \dots, X_n >$

$$Y^{new} \leftarrow \arg\max_{y_k} \; P(Y = y_k) \prod_i P(X_i^{new} | Y = y_k)$$

Naïve Bayes Algorithm – discrete X_i

- Train Naïve Bayes (examples)

  for each* value y_k

  estimate $\pi_k \equiv P(Y = y_k)$

  for each* value x_{ij} of each attribute X_i

  estimate $\theta_{ijk} \equiv P(X_i = x_{ij} | Y = y_k)$

- Classify (X^{new})

$$Y^{new} \leftarrow \arg\max_{y_k} \; P(Y = y_k) \prod_i P(X_i^{new} | Y = y_k)$$

$$Y^{new} \leftarrow \arg\max_{y_k} \; \pi_k \prod_i \theta_{ijk}$$

* probabilities must sum to 1, so need estimate only n-1 of these...

Estimating Parameters: Y, X_i discrete-valued

Maximum likelihood estimates (MLE's):

$$\hat{\pi}_k = \hat{P}(Y = y_k) = \frac{\#D\{Y = y_k\}}{|D|}$$

$$\hat{\theta}_{ijk} = \hat{P}(X_i = x_{ij} | Y = y_k) = \frac{\#D\{X_i = x_{ij} \wedge Y = y_k\}}{\#D\{Y = y_k\}}$$

Number of items in dataset D for which $Y=y_k$

Example: Live in Sq Hill? P(S|G,D,B)

- S=1 iff live in Squirrel Hill
- G=1 iff shop at SH Giant Eagle
- D=1 iff Drive or carpool to CMU
- B=1 iff Birthday is before July 1

What probability parameters must we estimate?

Example: Live in Sq Hill? P(S|G,D,E)

- S=1 iff live in Squirrel Hill
- G=1 iff shop at SH Giant Eagle
- D=1 iff Drive or Carpool to CMU
- B=1 iff Birthday is before July 1

P(S=1) :

P(D=1 | S=1) :

P(D=1 | S=0) :

P(G=1 | S=1) :

P(G=1 | S=0) :

P(B=1 | S=1) :

P(B=1 | S=0) :

P(S=0) :

P(D=0 | S=1) :

P(D=0 | S=0) :

P(G=0 | S=1) :

P(G=0 | S=0) :

P(B=0 | S=1) :

P(B=0 | S=0) :

Naïve Bayes: Subtlety #1

Often the X_i are not really conditionally independent

- We use Naïve Bayes in many cases anyway, and it often works pretty well
  - often the right classification, even when not the right probability (see [Domingos&Pazzani, 1996])

- What is effect on estimated P(Y|X)?
  - Extreme case: what if we add two copies: $X_i = X_k$

Extreme case: what if we add two copies: $X_i = X_k$

Naïve Bayes: Subtlety #2

If unlucky, our MLE estimate for $P(X_i \mid Y)$ might be zero.
(for example, X_i = *birthdate.* X_i = *Jan_25_1992*)

- Why worry about just one parameter out of many?

- What can be done to address this?

Estimating Parameters

- Maximum Likelihood Estimate (MLE): choose θ that maximizes probability of observed data $\mathcal{D}$

$$\hat{\theta} = \arg\max_{\theta} P(\mathcal{D} \mid \theta)$$

- Maximum a Posteriori (MAP) estimate: choose θ that is most probable given prior probability and the data

$$\hat{\theta} = \arg\max_{\theta} P(\theta \mid \mathcal{D})$$

$$= \arg\max_{\theta} = \frac{P(\mathcal{D} \mid \theta) P(\theta)}{P(\mathcal{D})}$$

Estimating Parameters: Y, X_i discrete-valued

Maximum likelihood estimates:

$$\hat{\pi}_k = \hat{P}(Y = y_k) = \frac{\#D\{Y = y_k\}}{|D|}$$

$$\hat{\theta}_{ijk} = \hat{P}(X_i = x_j | Y = y_k) = \frac{\#D\{X_i = x_j \wedge Y = y_k\}}{\#D\{Y = y_k\}}$$

MAP estimates (Beta, Dirichlet priors):

$$\hat{\pi}_k = \hat{P}(Y = y_k) = \frac{\#D\{Y = y_k\} + (\beta_k - 1)}{|D| + \sum_m (\beta_m - 1)}$$

$$\hat{\theta}_{ijk} = \hat{P}(X_i = x_j | Y = y_k) = \frac{\#D\{X_i = x_j \wedge Y = y_k\} + (\beta_k - 1)}{\#D\{Y = y_k\} + \sum_m (\beta_m - 1)}$$

Only difference:
"imaginary" examples

Learning to classify text documents

- Classify which emails are spam?
- Classify which emails promise an attachment?
- Classify which web pages are student home pages?

How shall we represent text documents for Naïve Bayes?

Baseline: Bag of Words Approach

aardvark	0
about	2
all	2
Africa	1
apple	0
anxious	0
...	
gas	1
...	
oil	1
…	
Zaire	0

Learning to classify document: P(Y|X) the "Bag of Words" model

- Y discrete valued. e.g., Spam or not
- $X = \langle X_1, X_2, \ldots X_n \rangle$ = document

- X_i is a random variable describing the word at position i in the document
- possible values for X_i : any word w_k in English

- Document = bag of words: the vector of counts for all w_k's
  - like #heads, #tails, but we have many more than 2 values
  - assume word probabilities are position independent (i.i.d. rolls of a 50,000-sided die)

Naïve Bayes Algorithm – discrete X_i

- Train Naïve Bayes (examples)

  for each value y_k

  estimate $\pi_k \equiv P(Y = y_k)$

  for each value x_j of each attribute X_i

  estimate $\theta_{ijk} \equiv P(X_i = x_j | Y = y_k)$

  prob that word x_j appears in position i, given $Y=y_k$

- Classify (X^{new})

$$Y^{new} \leftarrow \arg\max_{y_k} P(Y = y_k) \prod_i P(X_i^{new} | Y = y_k)$$

$$Y^{new} \leftarrow \arg\max_{y_k} \pi_k \prod_i \theta_{ijk}$$

* Additional assumption: word probabilities are position independent

$$\theta_{ijk} = \theta_{mjk} \quad \text{for all } i, m$$

MAP estimates for bag of words

Map estimate for multinomial

$$\theta_i = \frac{\alpha_i + \beta_i - 1}{\sum_{m=1}^{k} \alpha_m + \sum_{m=1}^{k} (\beta_m - 1)}$$

What β' s should we choose?

Twenty NewsGroups

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

Recap on DT:

- Concept learning. Expressing “AND” and “OR” with a tree.
- Entropy and Gain functions: Quantify the quality of the features for branching in a tree.
- ID3 algorithm:
  - Iteratively choose the best feature for each branch.
  - Complete search space (No restriction bias).
  - Incomplete search strategy (Preference bias towards short trees).
- Avoiding overfitting:
  - Build the tree and then prune it.
  - Criteria: Cross-validation/Minimum Description Length.
- Extensions:
  - Continuous values.
  - Missing values.
  - Different gain functions.
  - Regression trees.

Probability Overview

- Events
  - discrete random variables, continuous random variables, compound events
- Axioms of probability
  - What defines a reasonable theory of uncertainty
- Independent events
- Conditional probabilities
- Bayes rule and beliefs
- Joint probability distribution
- Expectations
- Independence, Conditional independence

Maximum Likelihood Estimate

X=1 X=0

P(X=1) = θ

P(X=0) = 1-θ

(Bernoulli)

- Each flip yields boolean value for X

$$X \sim \text{Bernoulli: } P(X) = \theta^X (1-\theta)^{(1-X)}$$

- Data set D of independent, identically distributed (iid) flips produces α_1 ones, α_0 zeros

$$P(D|\theta) = P(\alpha_1, \alpha_0|\theta) = \theta^{\alpha_1}(1-\theta)^{\alpha_0}$$

$$\hat{\theta}^{MLE} = \arg\max_{\theta} P(D|\theta) = \frac{\alpha_1}{\alpha_1 + \alpha_0}$$

Maximum A Posteriori (MAP) Estimate

- Data set D of independent, identically distributed (iid) flips produces α_1 ones, α_0 zeros

$$P(D|\theta) = P(\alpha_1, \alpha_0|\theta) = \theta^{\alpha_1}(1-\theta)^{\alpha_0}$$

- Assume prior $P(\theta) = Beta(\beta_1, \beta_0) = \frac{1}{B(\beta_1,\beta_0)}\,\theta^{\beta_1-1}(1-\theta)^{\beta_0-1}$
- Then

$$\hat{\theta}^{MAP} = \arg\max_{\theta} P(D|\theta)P(\theta) = \frac{\alpha_1+\beta_1-1}{(\alpha_1+\beta_1-1)+(\alpha_0+\beta_0-1)}$$

(like MLE, but hallucinating β_1-1 additional heads, β_0-1 additional tails)

Naïve Bayes in a Nutshell

Bayes rule:

$$P(Y = y_k | X_1 \dots X_n) = \frac{P(Y = y_k) P(X_1 \dots X_n | Y = y_k)}{\sum_j P(Y = y_j) P(X_1 \dots X_n | Y = y_j)}$$

Assuming conditional independence among X_i's:

$$P(Y = y_k | X_1 \dots X_n) = \frac{P(Y = y_k) \prod_i P(X_i | Y = y_k)}{\sum_j P(Y = y_j) \prod_i P(X_i | Y = y_j)}$$

So, to pick most probable Y for $X^{new} = < X_1, \dots, X_n >$

$$Y^{new} \leftarrow \arg\max_{y_k} \; P(Y = y_k) \prod_i P(X_i^{new} | Y = y_k)$$

Naïve Bayes Algorithm – discrete X_i

- Train Naïve Bayes (examples)

  for each* value y_k

  estimate $\pi_k \equiv P(Y = y_k)$

  for each* value x_{ij} of each attribute X_i

  estimate $\theta_{ijk} \equiv P(X_i = x_{ij} | Y = y_k)$

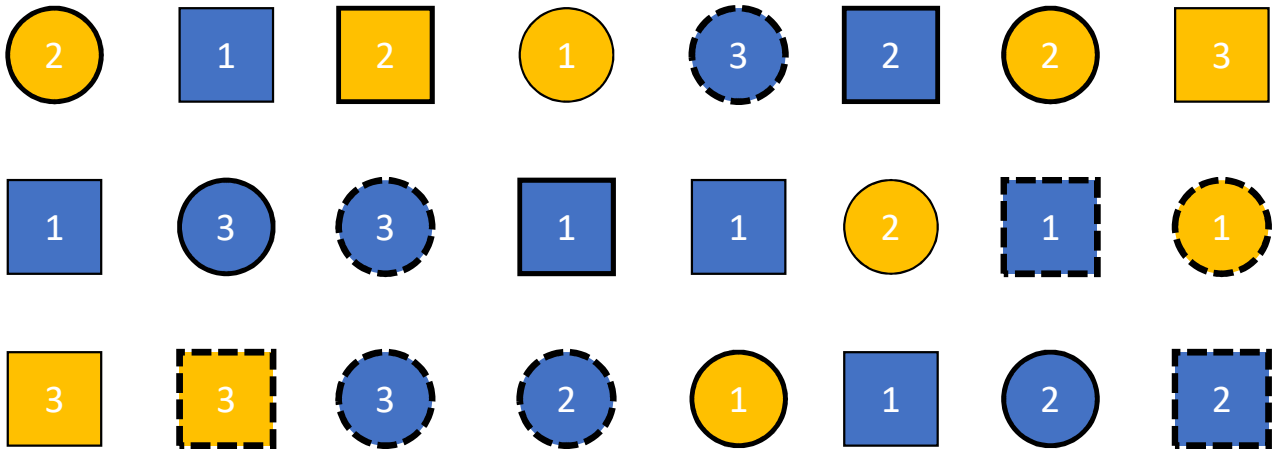
- Classify (X^{new})

$$Y^{new} \leftarrow \arg\max_{y_k} \; P(Y = y_k) \prod_i P(X_i^{new} | Y = y_k)$$

$$Y^{new} \leftarrow \arg\max_{y_k} \; \pi_k \prod_i \theta_{ijk}$$

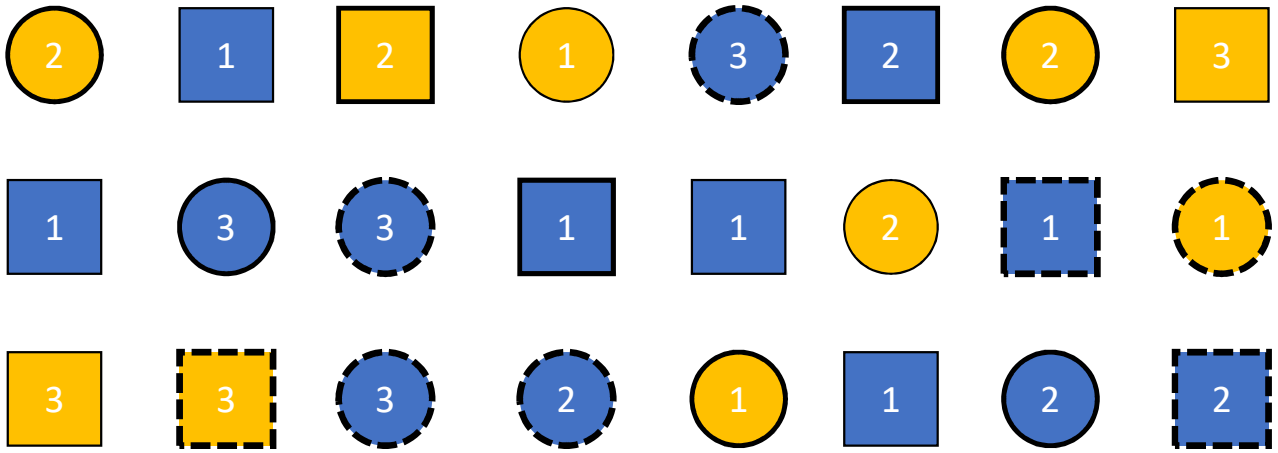
* probabilities must sum to 1, so need estimate only n-1 of these...

Applying the Naïve Bayes method (MLE):



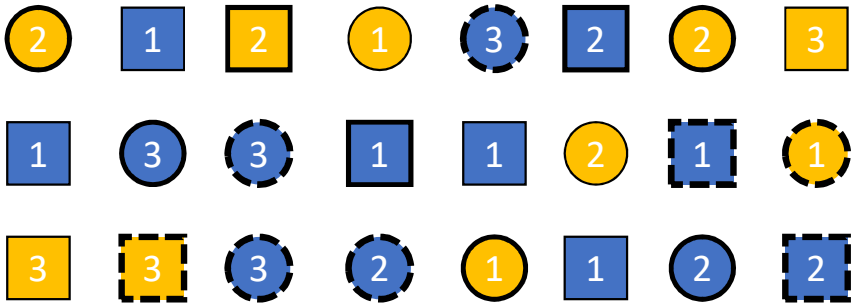
Our **data set**: Classify shapes of the objects according to their colors, border and numbers.

Applying the Naïve Bayes method (MLE):



Our **data set**: Classify **shapes** (y) of the objects according to their **colors, border and numbers** (X)

Applying the Naïve Bayes method (MLE):

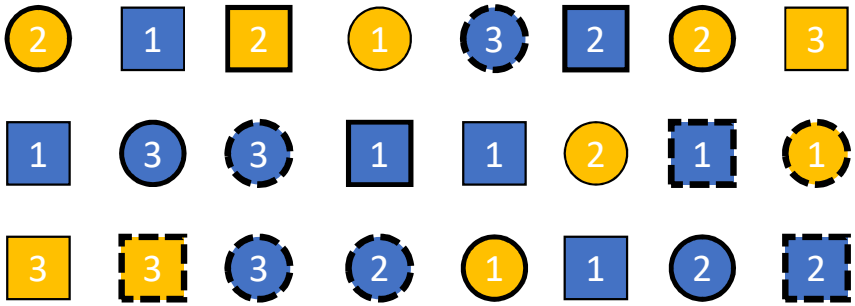


COLOR	Square	Circle
BLUE	8	6
GOLD	4	6

BORDER	Square	Circle
BOLD	3	5
THIN	6	2
DASHED	3	5

NUMBER	Square	Circle
1	6	3
2	3	5
3	3	4

Applying the Naïve Bayes method (MLE):



New case: (BLUE, THIN, 1) is a square or a circle?

$$\hat{\pi}_{\odot} = \frac{\#D\{Y = \odot\}}{|D|} = \frac{12}{24} \quad \hat{\pi}_{\square} = \frac{\#D\{Y = \square\}}{|D|} = \frac{12}{24}$$

$$P(BLUE|\odot) = \frac{\#D\{X = BLUE \wedge Y = \odot\}}{\#D\{Y = \odot\}} = \frac{6}{12}$$

COLOR	Square	Circle
BLUE	8	6
GOLD	4	6

BORDER	Square	Circle
BOLD	3	5
THIN	6	2
DASHED	3	5

NUMBER	Square	Circle
1	6	3
2	3	5
3	3	4

Applying the Naïve Bayes method (MLE):

New case: (BLUE, THIN, 1) is a square or a circle?

$$\hat{\pi}_{\odot} = \frac{\#D\{Y = \odot\}}{|D|} = \frac{12}{24} \quad \hat{\pi}_{\boxdot} = \frac{\#D\{Y = \boxdot\}}{|D|} = \frac{12}{24}$$

$$P(BLUE|\odot) = \frac{\#D\{X = BLUE \wedge Y = \odot\}}{\#D\{Y = \odot\}} = \frac{6}{12} \quad P(1|\odot) = \frac{\#D\{X = 1 \wedge Y = \odot\}}{\#D\{Y = \odot\}} = \frac{3}{12}$$

$$P(BLUE|\boxdot) = \frac{\#D\{X = BLUE \wedge Y = \boxdot\}}{\#D\{Y = \boxdot\}} = \frac{8}{12} \quad P(1|\boxdot) = \frac{\#D\{X = 1 \wedge Y = \boxdot\}}{\#D\{Y = \boxdot\}} = \frac{6}{12}$$

$$P(THIN|\odot) = \frac{\#D\{X = THIN \wedge Y = \odot\}}{\#D\{Y = \odot\}} = \frac{2}{12}$$

$$P(THIN|\boxdot) = \frac{\#D\{X = THIN \wedge Y = \boxdot\}}{\#D\{Y = \boxdot\}} = \frac{6}{12}$$

COLOR	Square	Circle
BLUE	8	6
GOLD	4	6

BORDER	Square	Circle
BOLD	3	5
THIN	6	2
DASHED	3	5

NUMBER	Square	Circle
1	6	3
2	3	5
3	3	4

Applying the Naïve Bayes method (MLE):

New case: (BLUE, THIN, 1) is a square or a circle?

$$\hat{\pi}_{\odot} = \frac{12}{24} \qquad \hat{\pi}_{\boxdot} = \frac{12}{24}$$

$$P(BLUE|\odot) = \frac{6}{12} \qquad P(THIN|\odot) = \frac{2}{12} \qquad P(1|\odot) = \frac{3}{12}$$

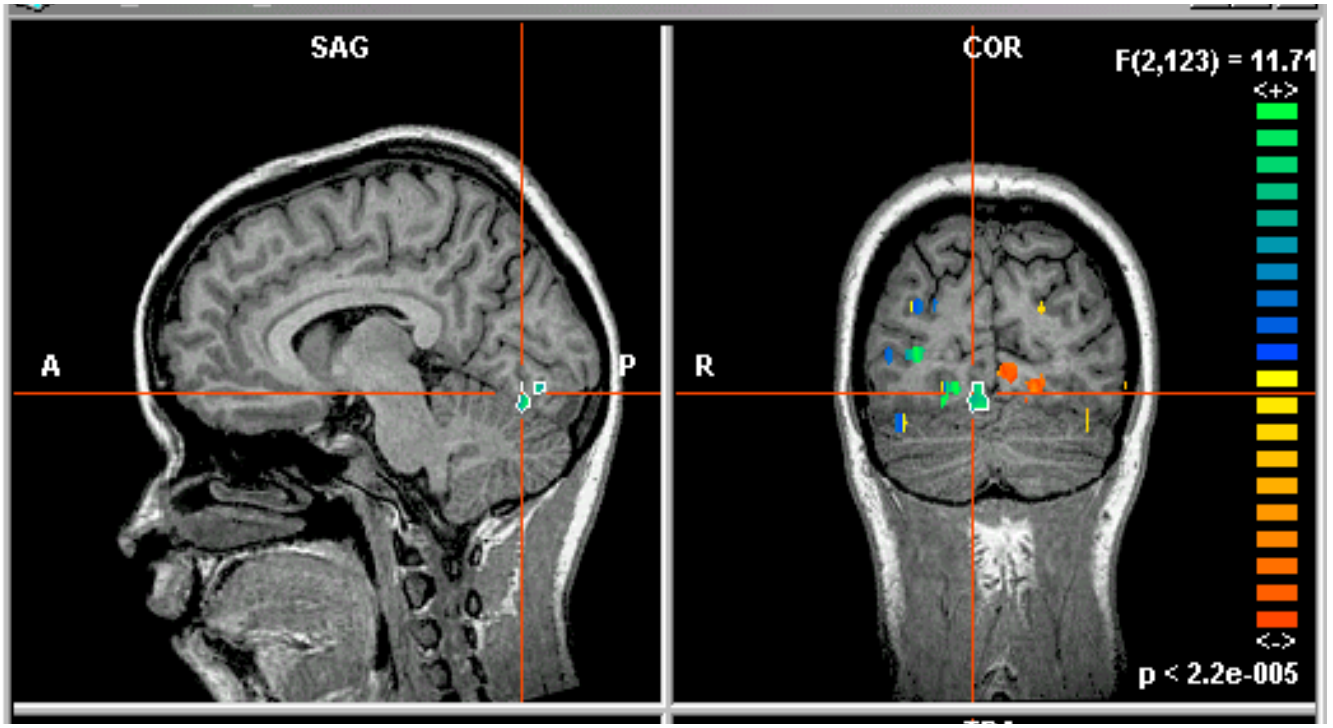
$$P(BLUE|\boxdot) = \frac{8}{12} \qquad P(THIN|\boxdot) = \frac{6}{12} \qquad P(1|\boxdot) = \frac{6}{12}$$

$$P(BLUE, THIN, 1|\odot) = \hat{\pi}_{\odot} \prod P(X|\odot) = \frac{12}{24}\frac{6}{12}\frac{2}{12}\frac{3}{12}$$

$$P(BLUE, THIN, 1|\boxdot) = \hat{\pi}_{\odot} \prod P(X|\boxdot) = \frac{12}{24}\frac{8}{12}\frac{6}{12}\frac{6}{12}$$

What if we have continuous X_i ?

Eg., image classification: X_i is real-valued ith pixel



What if we have continuous X_i ?

Eg., image classification: X_i is real-valued i[th] pixel

Naïve Bayes requires $P(X_i \mid Y{=}y_k)$, but X_i is real (continuous)

$$P(Y = y_k | X_1 \dots X_n) = \frac{P(Y = y_k) \prod_i P(X_i | Y = y_k)}{\sum_j P(Y = y_j) \prod_i P(X_i | Y = y_j)}$$

Common approach: assume $P(X_i \mid Y{=}y_k)$ follows a Normal (Gaussian) distribution

What if we have continuous X_i ?

Eg., image classification: X_i is real-valued i[th] pixel

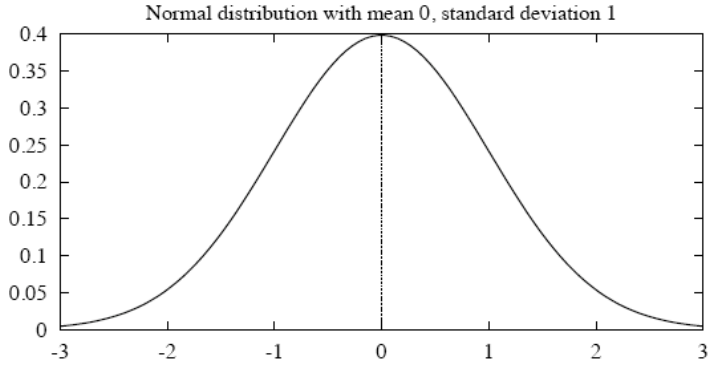
Naïve Bayes requires $P(X_i \mid Y{=}y_k)$, but X_i is real (continuous)

$$P(Y = y_k | X_1 \dots X_n) = \frac{P(Y = y_k) \prod_i P(X_i | Y = y_k)}{\sum_j P(Y = y_j) \prod_i P(X_i | Y = y_j)}$$

Common approach: assume $P(X_i \mid Y{=}y_k)$ follows a Normal (Gaussian) distribution

Gaussian Distribution

(also called “Normal”)



p(x) is a *probability density function*, whose integral (not sum) is 1

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

The probability that X will fall into the interval (a, b) is given by

$$\int_a^b p(x)dx$$

- Expected, or mean value of X, $E[X]$, is

$$E[X] = \mu$$

- Variance of X is

$$Var(X) = \sigma^2$$

- Standard deviation of X, σ_X, is

$$\sigma_X = \sigma$$

What if we have continuous X_i ?

Gaussian Naïve Bayes (GNB): assume

$$p(X_i = x | Y = y_k) = \frac{1}{\sqrt{2\pi\sigma_{ik}^2}} \; e^{-\frac{1}{2}(\frac{x-\mu_{ik}}{\sigma_{ik}})^2}$$

Sometimes assume variance

- is independent of Y (i.e., σ_i),
- or independent of X_i (i.e., σ_k)
- or both (i.e., σ)

Gaussian Naïve Bayes Algorithm – continuous X_i
(but still discrete Y)

- Train Naïve Bayes (examples)

  for each value y_k

  estimate* $\pi_k \equiv P(Y = y_k)$

  for each attribute X_i estimate $P(X_i|Y = y_k)$

  - class conditional mean μ_{ik}, variance σ_{ik}

- Classify (X^{new})

$$Y^{new} \leftarrow \arg\max_{y_k} \; P(Y = y_k) \prod_i P(X_i^{new}|Y = y_k)$$

$$Y^{new} \leftarrow \arg\max_{y_k} \; \pi_k \prod_i \mathcal{N}(X_i^{new}; \mu_{ik}, \sigma_{ik})$$

* probabilities must sum to 1, so need estimate only n-1 parameters...

Estimating Parameters: Y discrete, X_i continuous

Maximum likelihood estimates:

$$\hat{\mu}_{ik} = \frac{1}{\sum_j \delta(Y^j = y_k)} \sum_j X_i^j \delta(Y^j = y_k)$$

jth training example

ith feature

kth class

$\delta()=1$ if $(y^j=y_k)$ else 0

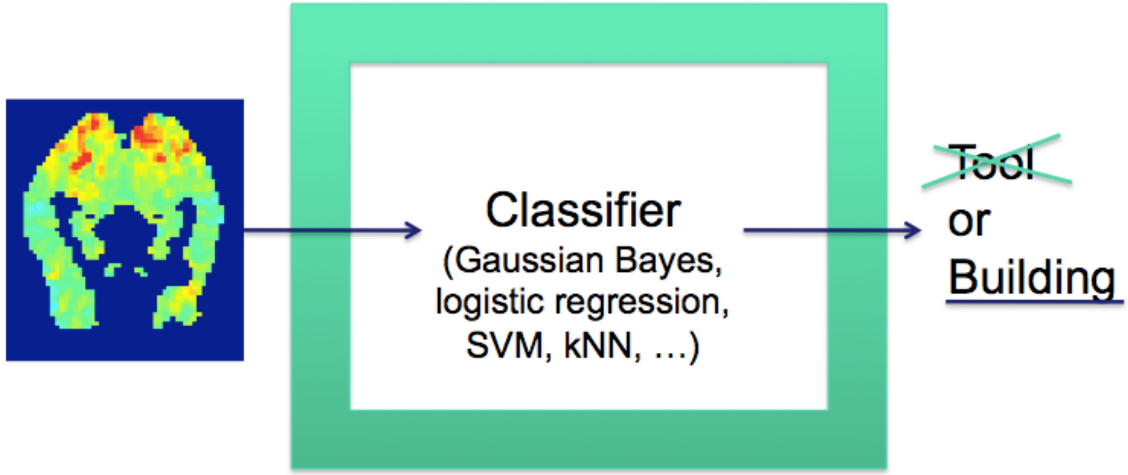
$$\hat{\sigma}_{ik}^2 = \frac{1}{\sum_j \delta(Y^j = y_k)} \sum_j (X_i^j - \hat{\mu}_{ik})^2 \delta(Y^j = y_k)$$

How many parameters must we estimate for Gaussian Naïve Bayes if Y has k possible values, X=<X1, … Xn>?

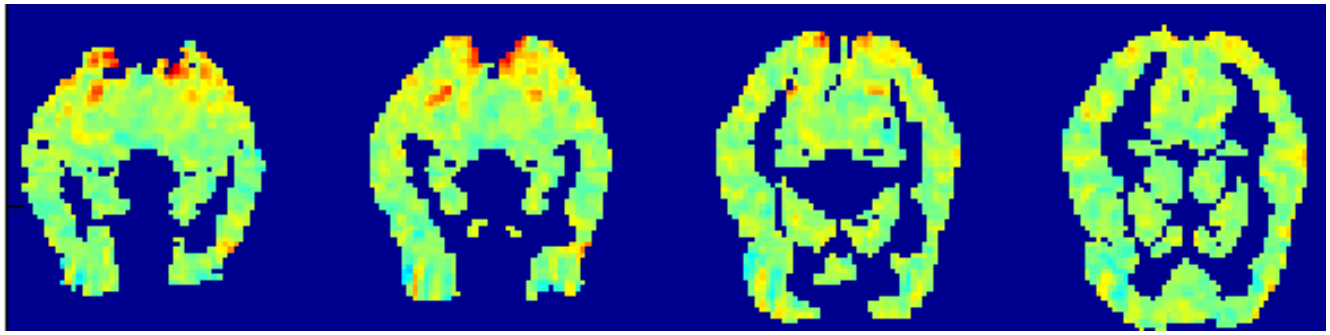
$$p(X_i = x | Y = y_k) = \frac{1}{\sqrt{2\pi\sigma_{ik}^2}} \; e^{-\frac{1}{2}(\frac{x-\mu_{ik}}{\sigma_{ik}})^2}$$

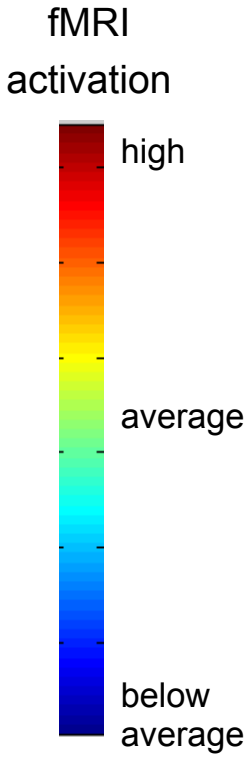
GNB Example: Classify a person's cognitive state, based on brain image

- reading a sentence or viewing a picture?
- reading the word describing a "Tool" or "Building"?
- answering the question, or getting confused?



Mean activations over all training examples for Y="bottle"

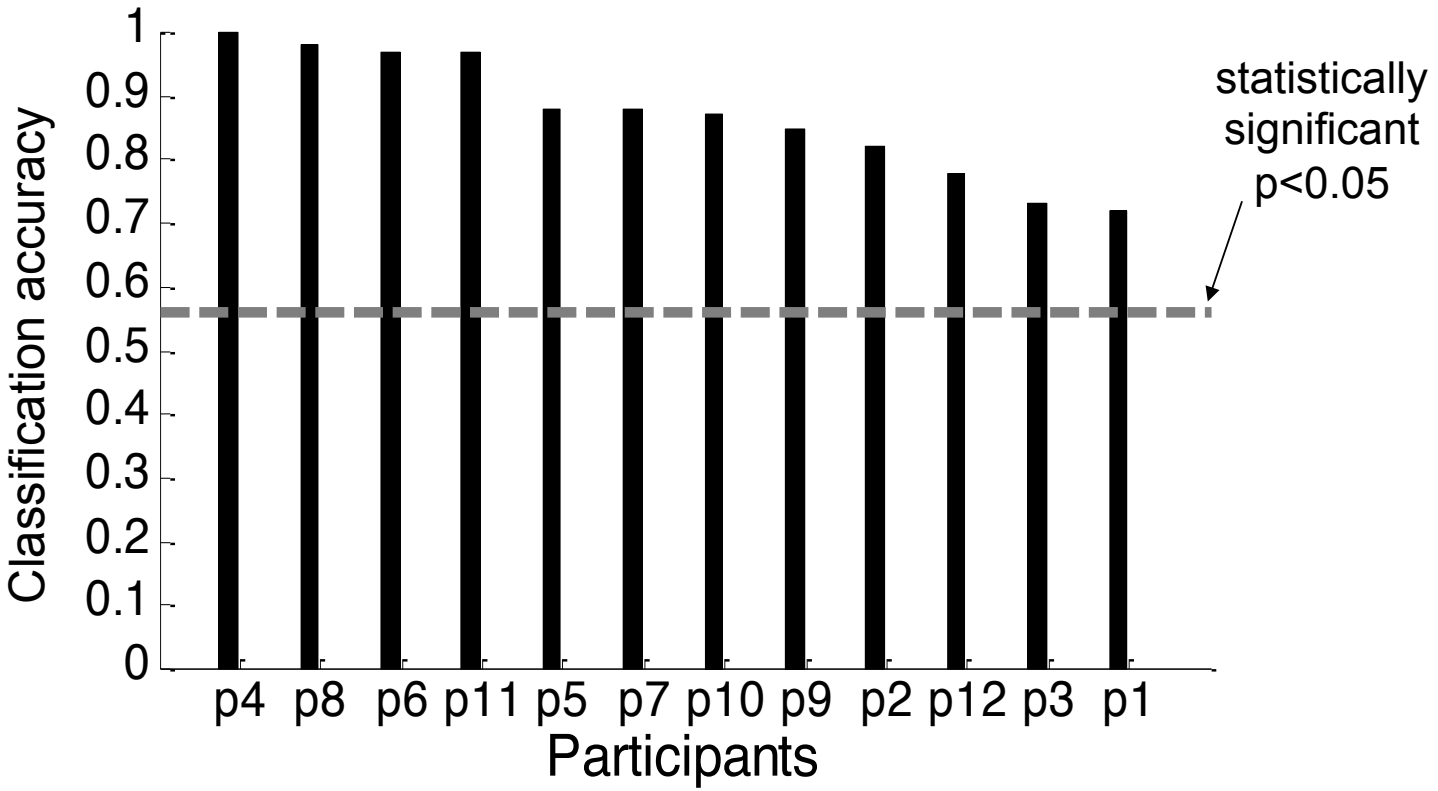




Y is the mental state (reading "house" or "bottle")
X_i are the voxel activities,

this is a plot of the μ's defining $P(X_i \mid Y=\text{"bottle"})$

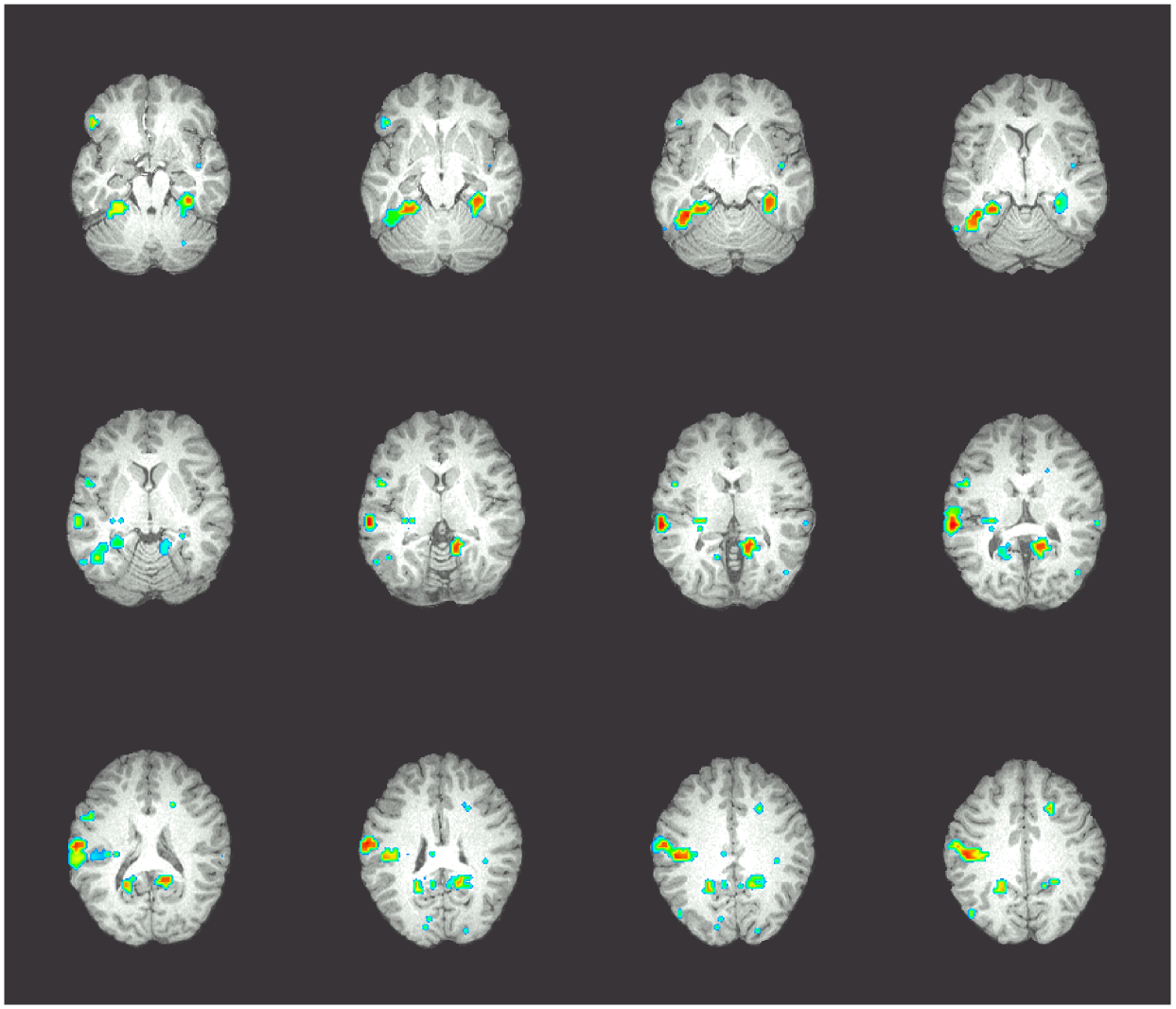
Classification task: is person viewing a “tool” or “building”?



Where is information encoded in the brain?

Accuracies of cubical 27-voxel classifiers centered at each significant voxel

[0.7-0.8]



Naïve Bayes: What you should know

- Designing classifiers based on Bayes rule
- Conditional independence
  - What it is
  - Why it's important
- Naïve Bayes assumption and its consequences
  - Which (and how many) parameters must be estimated under different generative models (different forms for P(X|Y))
    - and why this matters
- How to train Naïve Bayes classifiers
  - MLE and MAP estimates
  - with discrete and/or continuous inputs X_i

sounds like the solution to
learning F: X → Y,
or P(Y | X).

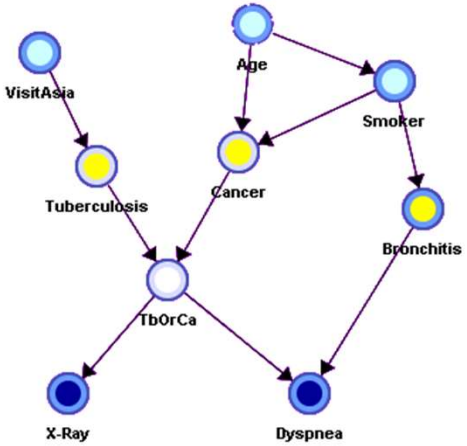
Main problem: learning P(Y|X)
can require more data than we have

consider learning Joint Dist. with 100 attributes
# of rows in this table?
# of people on earth?
fraction of rows with 0 training examples?

What to do?

1. Be smart about how we estimate probabilities from sparse data
   - maximum likelihood estimates
   - maximum a posteriori estimates

2. Be smart about how to represent joint distributions
   - Bayes networks, graphical models



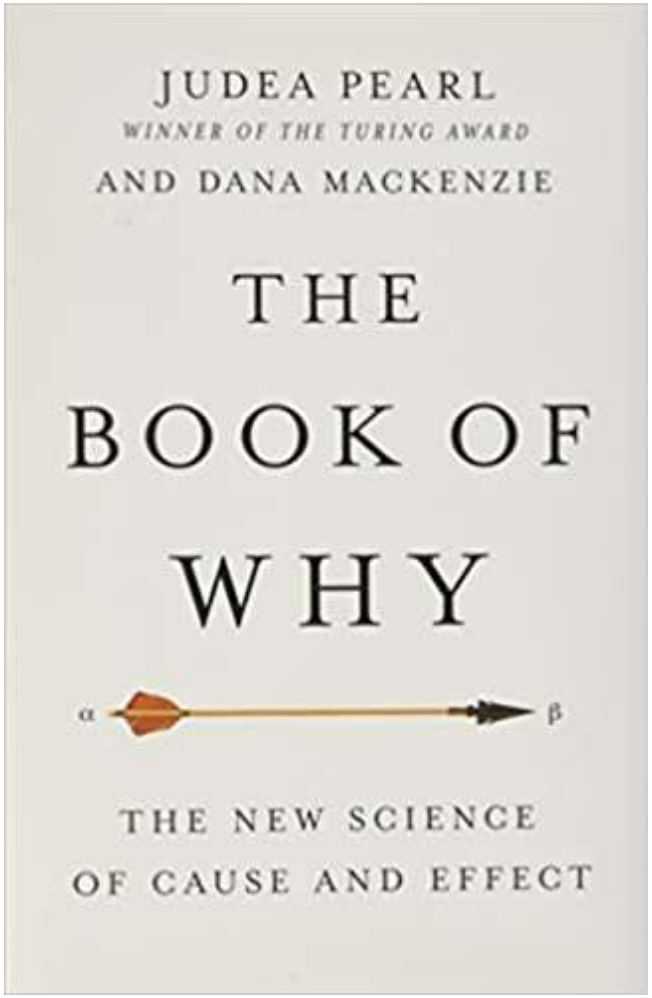
Reasoning with Bayesian Belief Networks

Overview

- Bayesian Belief Networks (BBNs) can reason with networks of propositions and associated probabilities
- BBNs encode causal associations between facts and events the propositions represent
- Useful for many AI problems
  - Diagnosis
  - Expert systems
  - Planning
  - Learning

Judea Pearl

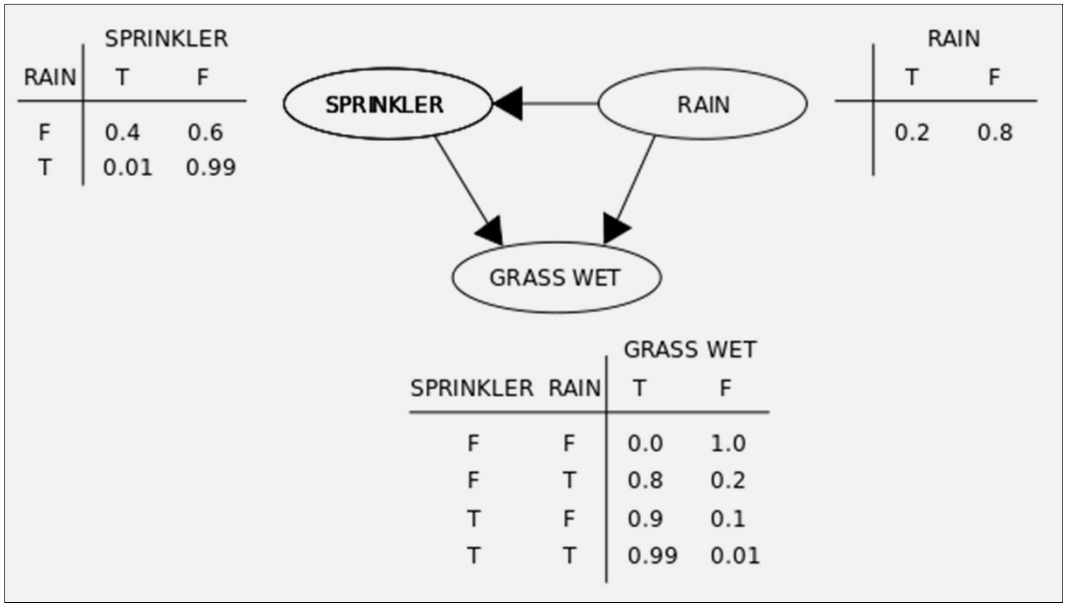
- UCLA CS professor
- Introduced Bayesian networks in the 1980
- Pioneer of probabilistic approach to AI reasoning
- First to mathematize causal modeling in empirical sciences
- Written many books on the topics, including the popular 2018 Book of Why



BBN Definition

- AKA Bayesian Network, Bayes Net
- A graphical model (as a DAG) of probabilistic relationships among a set of random variables
- Nodes are variables, links represent direct influence of one variable on another
- Nodes have prior probabilities or Conditional Probability Tables (CPTs)

source



Recall Bayes Rule

$$P(H,E) = P(H \mid E)P(E) = P(E \mid H)P(H)$$

$$P(H \mid E) = \frac{P(E \mid H)P(H)}{P(E)}$$

$$P(H \mid E) = \frac{P(E \mid H)P(H)}{P(E)}$$

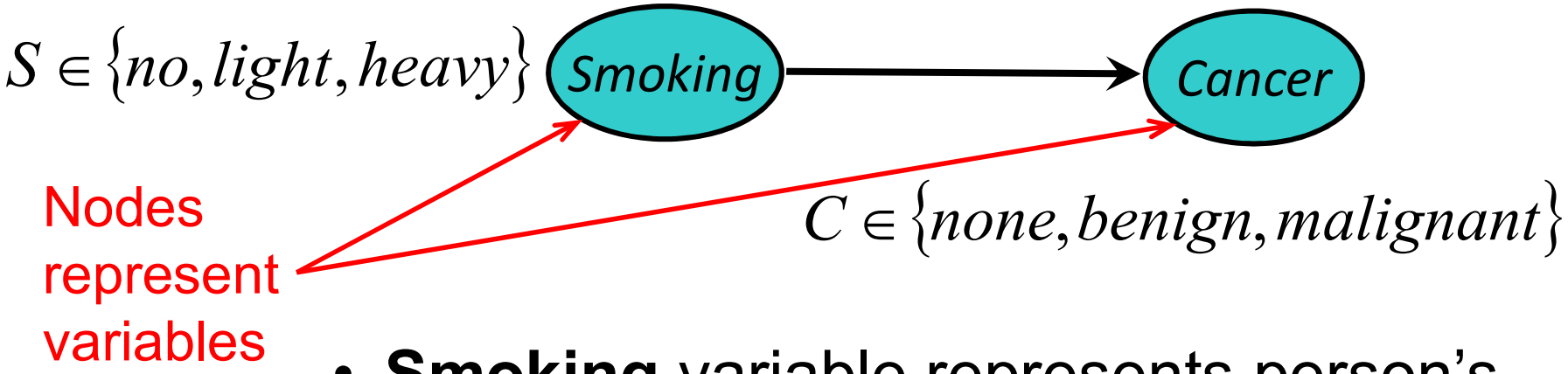
Note symmetry: can compute probability of a ***hypothesis given its evidence*** as well as probability of ***evidence given hypothesis***

Simple Bayesian Network

$S \in \{no, light, heavy\}$

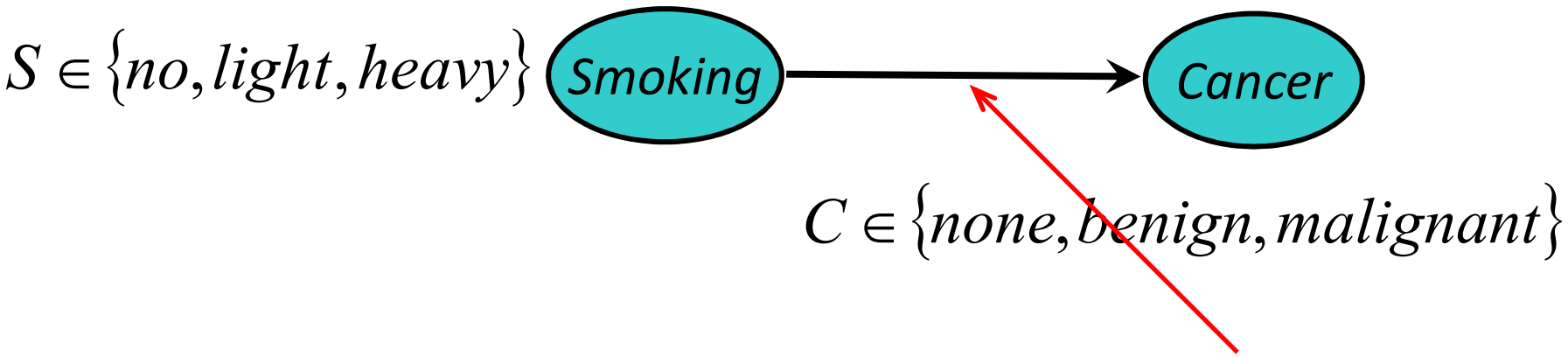
$C \in \{none, benign, malignant\}$

Simple Bayesian Network



- **Smoking** variable represents person's degree of smoking and has three possible values (no, light, heavy)
- **Cancer** variable represents person's cancer diagnosis and has three possible values (none, benign, malignant)

Simple Bayesian Network



- **tl;dr:** smoking effects cancer
- **Smoking** behavior effects the probability of **cancer** outcome
- **Smoking** behavior considered evidence for whether a person is likely to have cancer or not

Directed links represent *"causal"* relations

Simple Bayesian Network

$S \in \{no, light, heavy\}$

$C \in \{none, benign, malignant\}$

Prior probability of S

P(S=no)	0.80
P(S=light)	0.15
P(S=heavy)	0.05

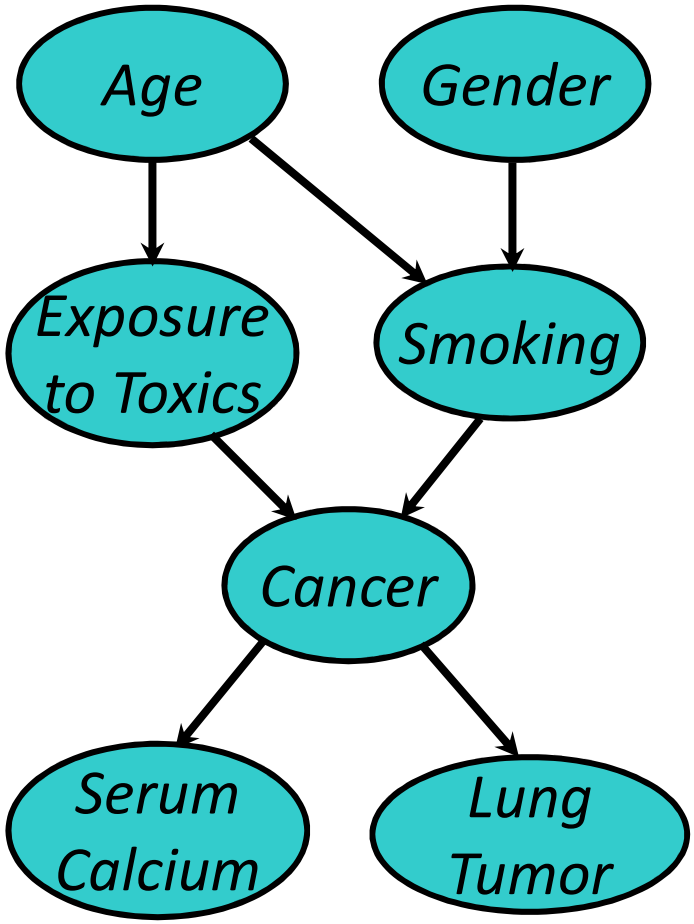
Nodes without in-links have **prior probabilities**

Nodes with in-links have **joint probability distributions**

Joint distribution of S and C

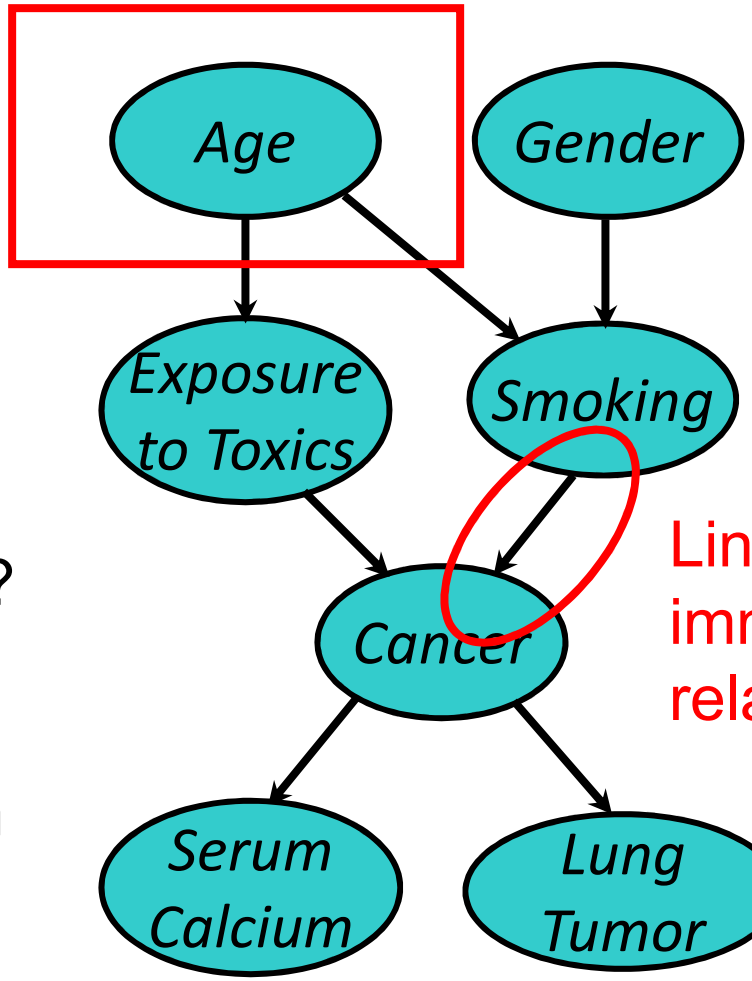
Smoking=	*no*	*light*	*heavy*
C=none	0.96	0.88	0.60
C=benign	0.03	0.08	0.25
C=malignant	0.01	0.04	0.15

More Complex Bayesian Network



More Complex Bayesian Network

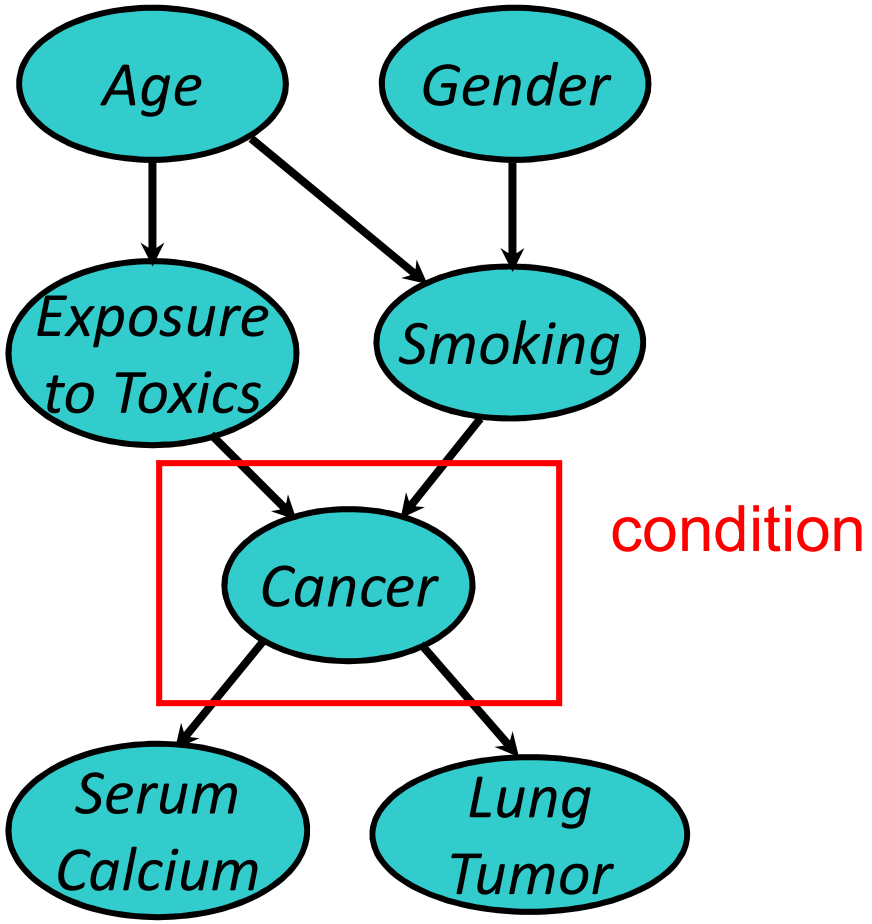
Nodes represent variables



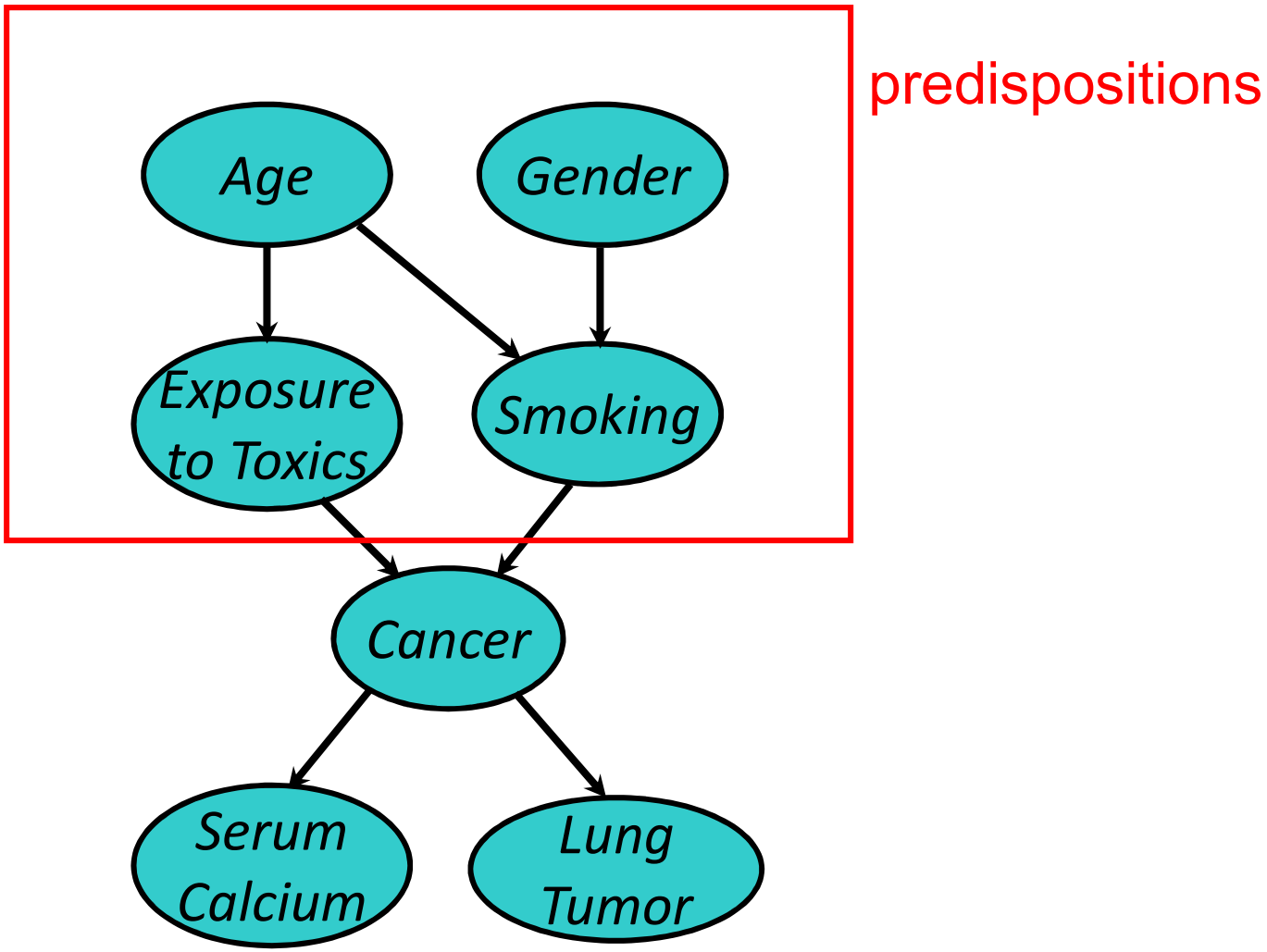
- Does gender cause smoking?
- Influence might be a better term

Links represent immediate *"causal"* relations

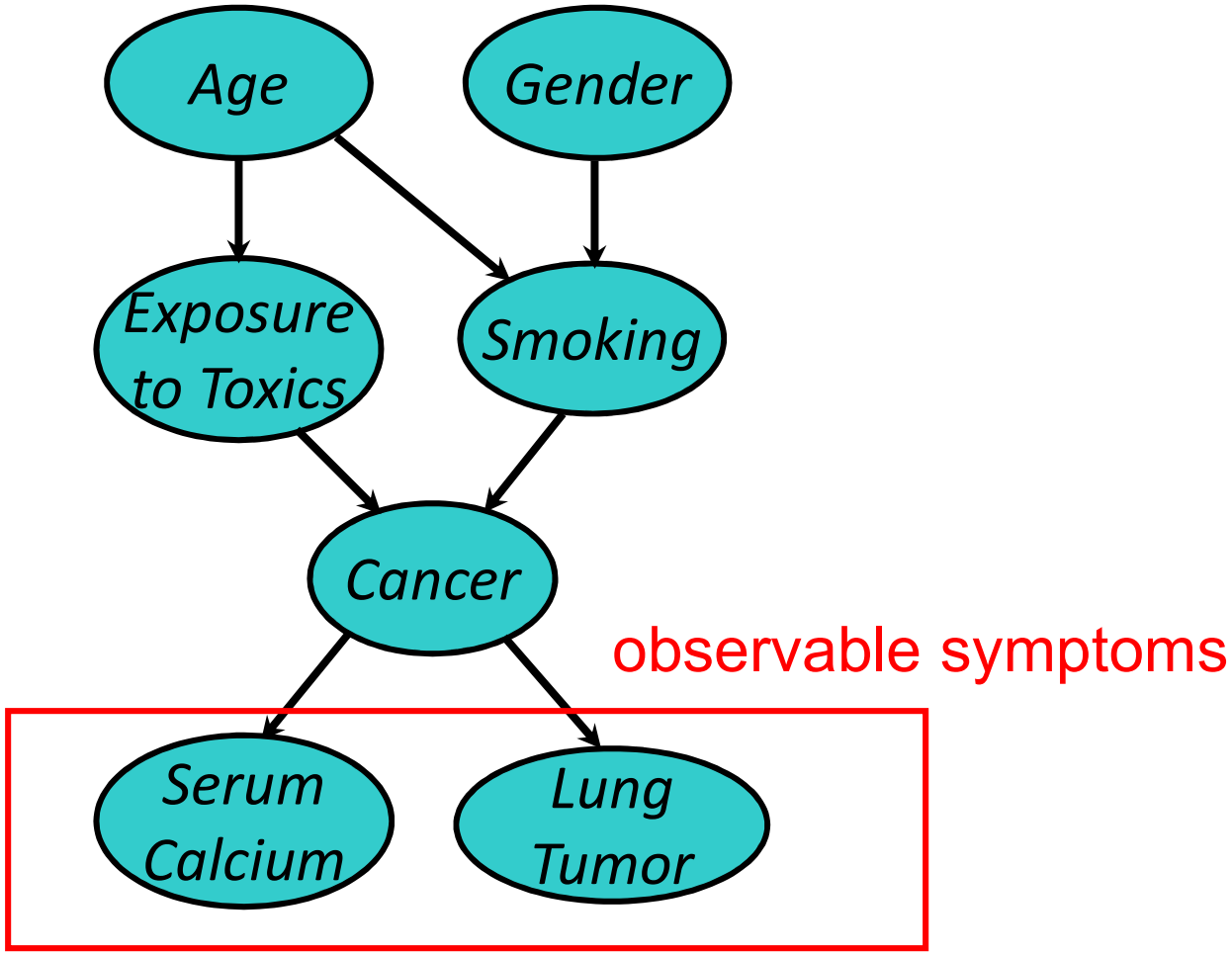
More Complex Bayesian Network



More Complex Bayesian Network

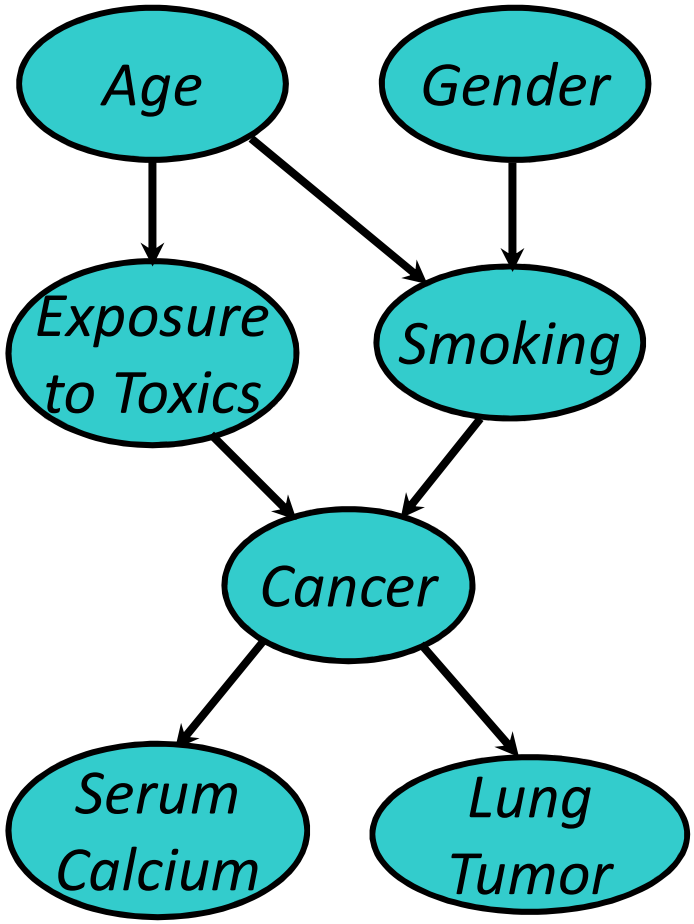


More Complex Bayesian Network



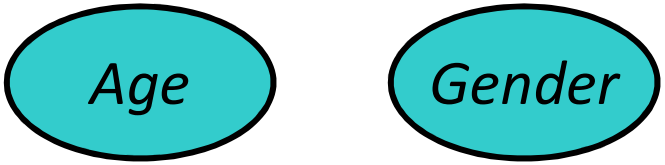
More Complex Bayesian Network

Can we predict likelihood of **lung tumor** given values of other 6 variables?



- Model has 7 variables
- Complete joint probability distribution will have 7 dimensions!
- Too much data required ☹
- BBN simplifies: a node has a CPT with data on itself & parents in graph

Independence



Age and *Gender* are independent.

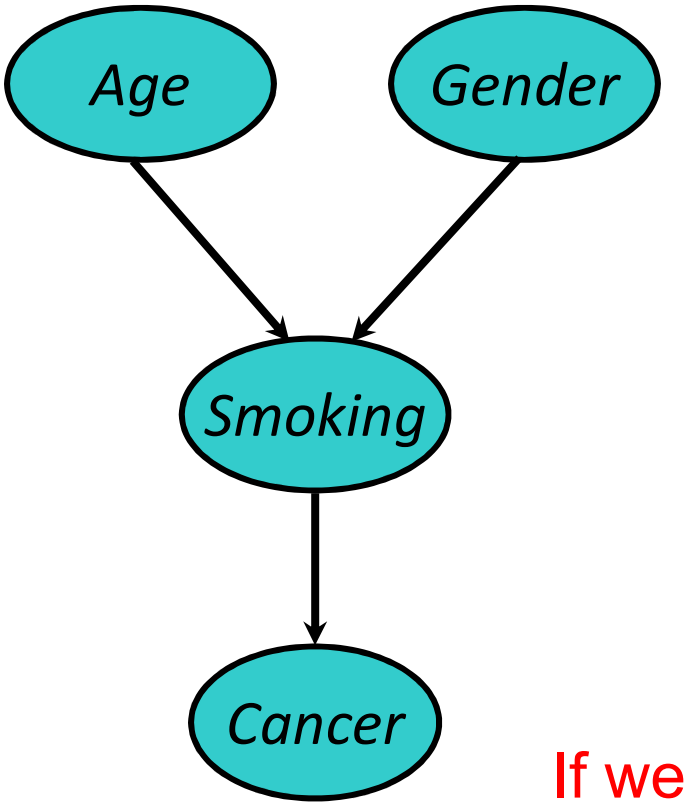
$P(A,G) = P(G) * P(A)$

There is no path between them in the graph

$P(A \mid G) = P(A)$
$P(G \mid A) = P(G)$

$P(A,G) = P(G|A)\ P(A) = P(G)P(A)$
$P(A,G) = P(A|G)\ P(G) = P(A)P(G)$

Conditional Independence

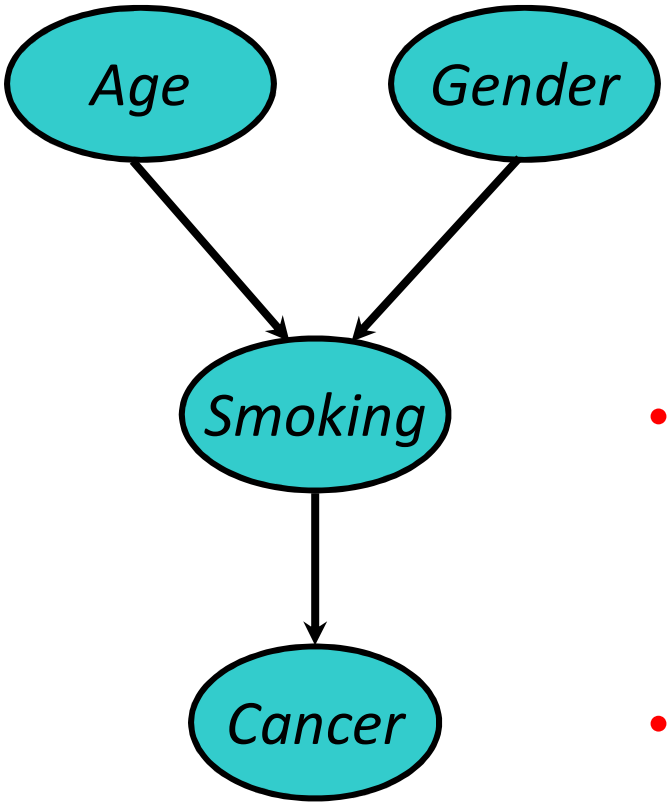


Cancer is independent of *Age* and *Gender* given *Smoking*

$$P(C \mid A,G,S) = P(C \mid S)$$

If we know value of smoking, no need to know values of age or gender

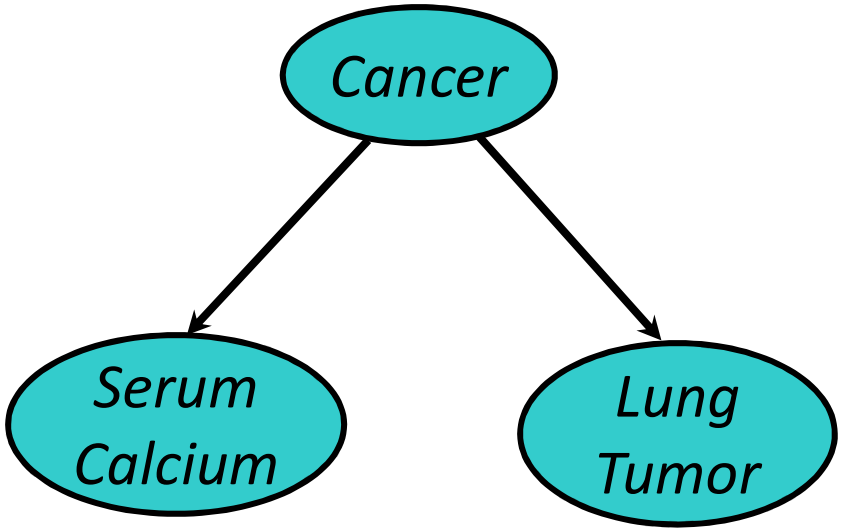
Conditional Independence



Cancer is independent of *Age* and *Gender* given *Smoking*

- Instead of one big CPT with 4 variables, we have two smaller CPTs with 3 and 2 variables
- If all variables binary: 12 models ($2^3 + 2^2$) rather than 16 (2^4)

Conditional Independence: Naïve Bayes



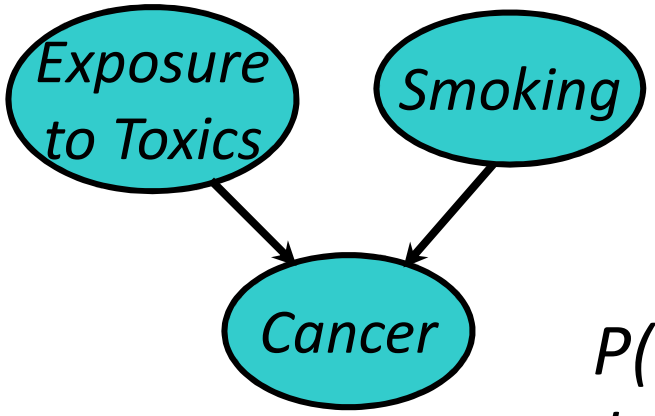
Serum Calcium and *Lung Tumor* are dependent

Serum Calcium is independent of *Lung Tumor*, given *Cancer*

$$P(L \mid SC,C) = P(L \mid C)$$
$$P(SC \mid L,C) = P(SC \mid C)$$

Naïve Bayes assumption: evidence (e.g., symptoms) independent given disease; easy to combine evidence

Explaining Away



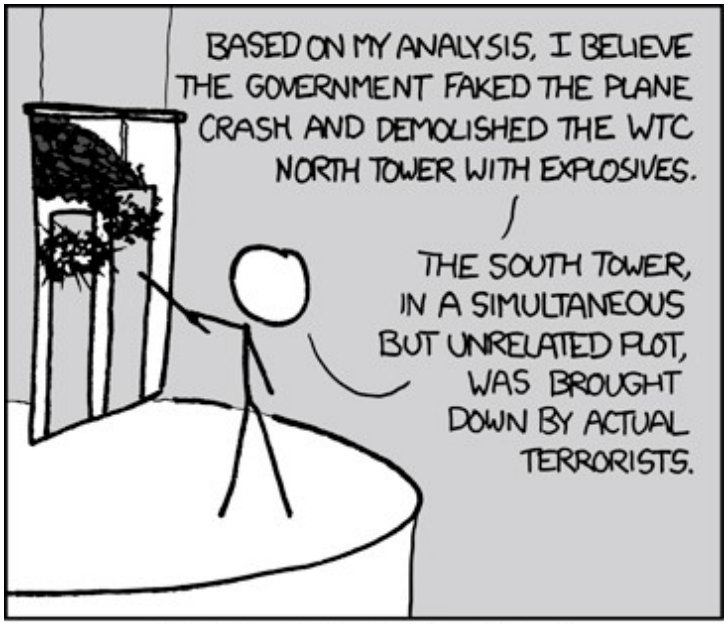
Exposure to Toxics and *Smoking* are independent

Exposure to Toxics is **dependent** on *Smoking*, given *Cancer*

P(E=heavy | C=malignant) > P(E=heavy | C=malignant, S=heavy)

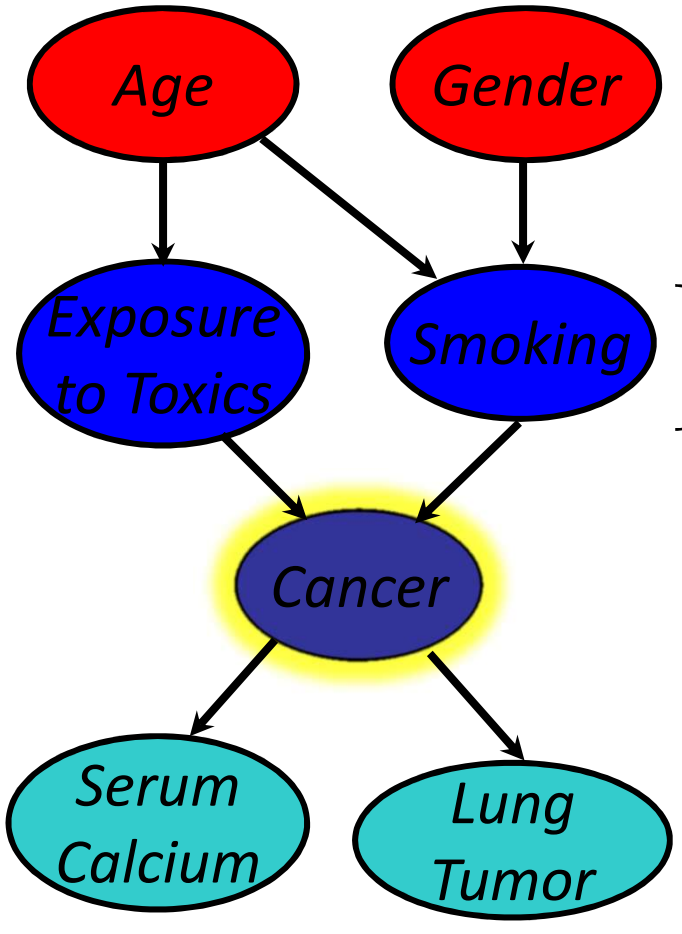
- *Explaining away:* reasoning pattern where confirmation of one cause reduces need to invoke alternatives
- Essence of Occam's Razor (prefer hypothesis with fewest assumptions)
- Relies on independence of causes

Explaining away



Conditional Independence

A variable (node) is conditionally independent of its non-descendants given its parents



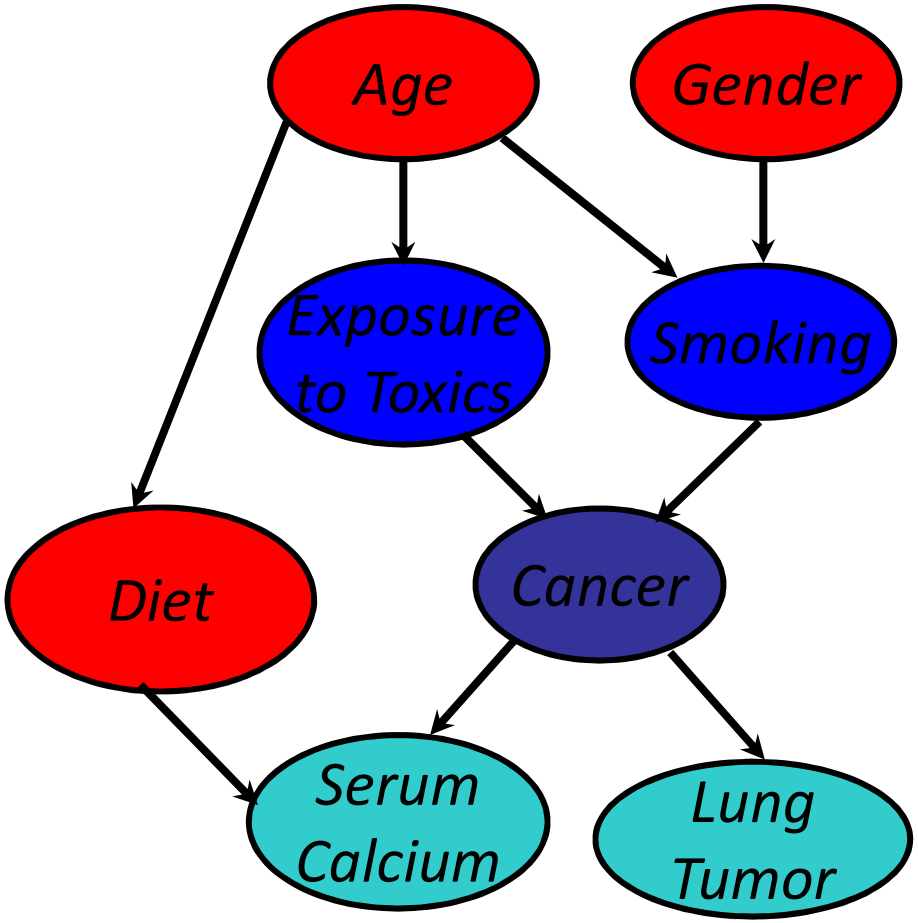
Non-Descendants

Parents

Descendants

Cancer is independent of *Age* and *Gender* given *Exposure to Toxics* and *Smoking*.

Another non-descendant



A variable is conditionally independent of its non-descendants given its parents

Cancer is independent of *Diet* given *Exposure to Toxics* and *Smoking*

BBN Construction

The **knowledge acquisition** process for a BBN involves three steps

KA1: Choosing appropriate variables

KA2: Deciding on the network structure

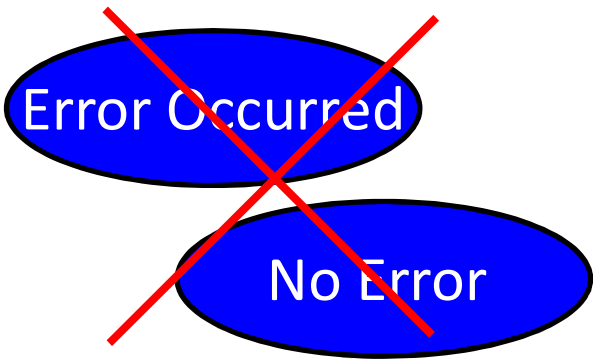
KA3: Obtaining data for the conditional probability tables

KA1: Choosing variables

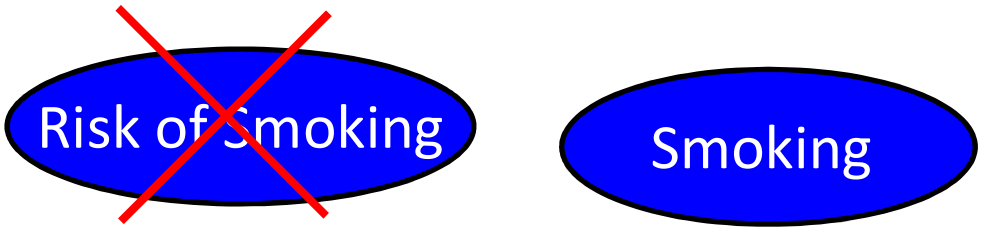
- Variable values: integers, reals or enumerations
- Variable should have collectively *exhaustive*, *mutually exclusive* values

$$x_1 \vee x_2 \vee x_3 \vee x_4$$

$$\neg(x_i \wedge x_j) \quad i \neq j$$



- They should be values, not probabilities

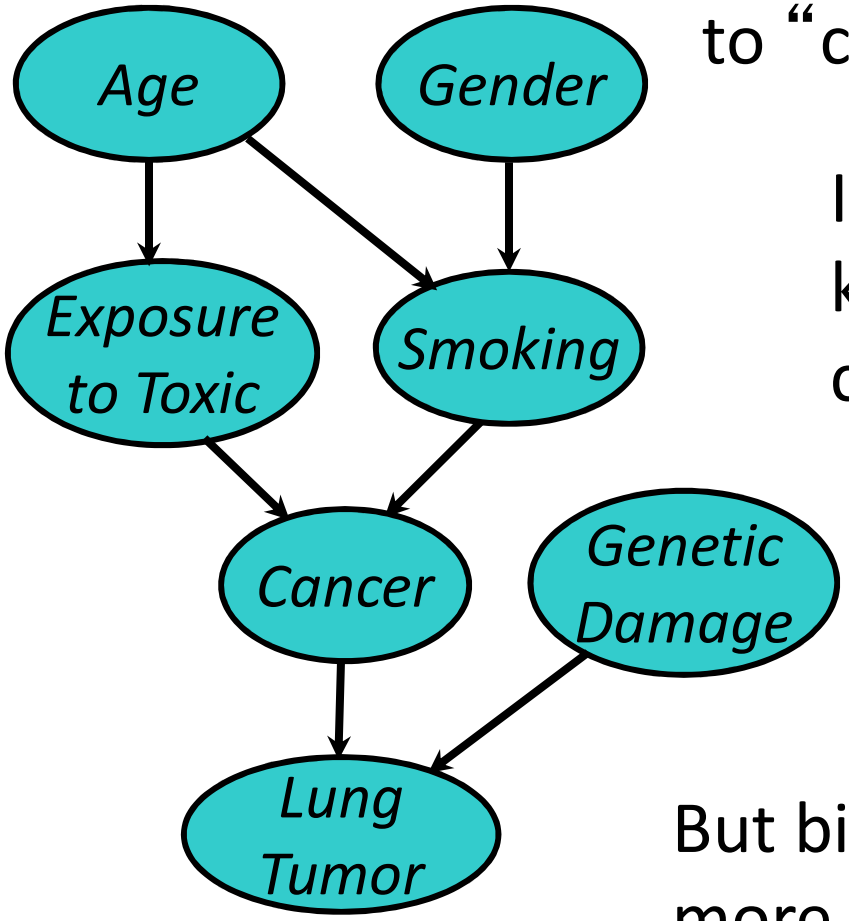


Heuristic: Knowable in Principle

Example of good variables

- Weather: {Sunny, Cloudy, Rain, Snow}
- Gasoline: $ per gallon {<1, 1-2, 2-3, 3-4, >4}
- Temperature: { $\geq 100^\circ$ F , < 100° F}
- User needs help on Excel Charts: {Yes, No}
- User's personality: {dominant, submissive}

KA2: Structuring



Network structure corresponding to "causality" is usually good.

Initially this uses designer's knowledge and intuitions but can be checked with data

May be better to add suspected links than to leave out

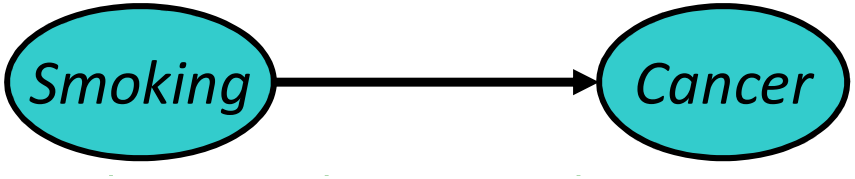
But bigger CPT tables mean more data collection

KA3: The Numbers

- For each variable we have a table of probability of its value for values of its **parents**
- For variables w/o parents, we have **prior probabilities**

$$S \in \{no, light, heavy\}$$

$$C \in \{none, benign, malignant\}$$

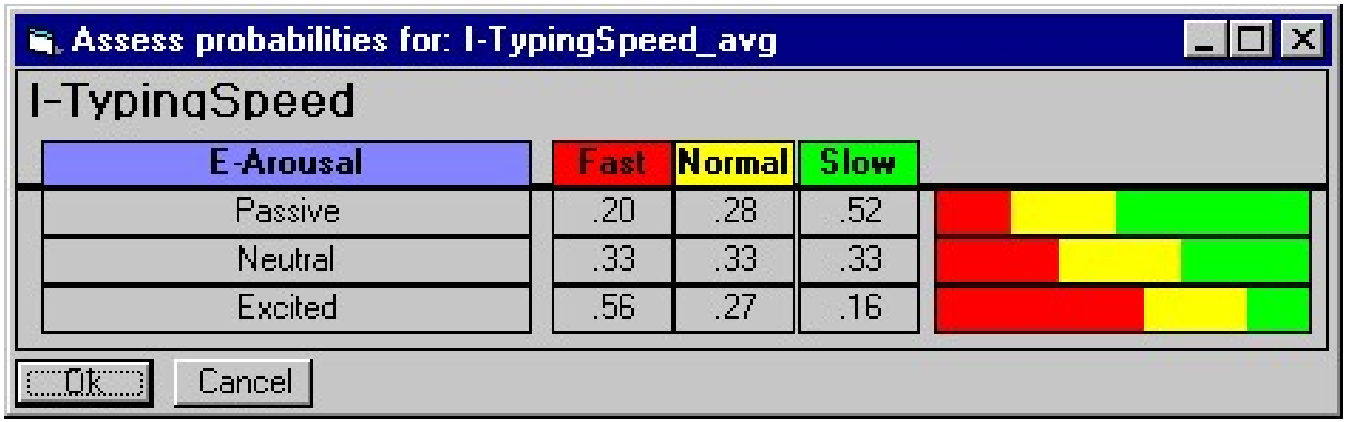


smoking priors	
no	0.80
light	0.15
heavy	0.05

	smoking		
cancer	**no**	**light**	**heavy**
none	0.96	0.88	0.60
benign	0.03	0.08	0.25
malignant	0.01	0.04	0.15

KA3: The numbers

- Second decimal usually doesn't matter
- Relative probabilities are important



- Zeros and ones are often enough
- Order of magnitude is typical: 10^{-9} vs 10^{-6}
- Sensitivity analysis can be used to decide accuracy needed

Three kinds of reasoning

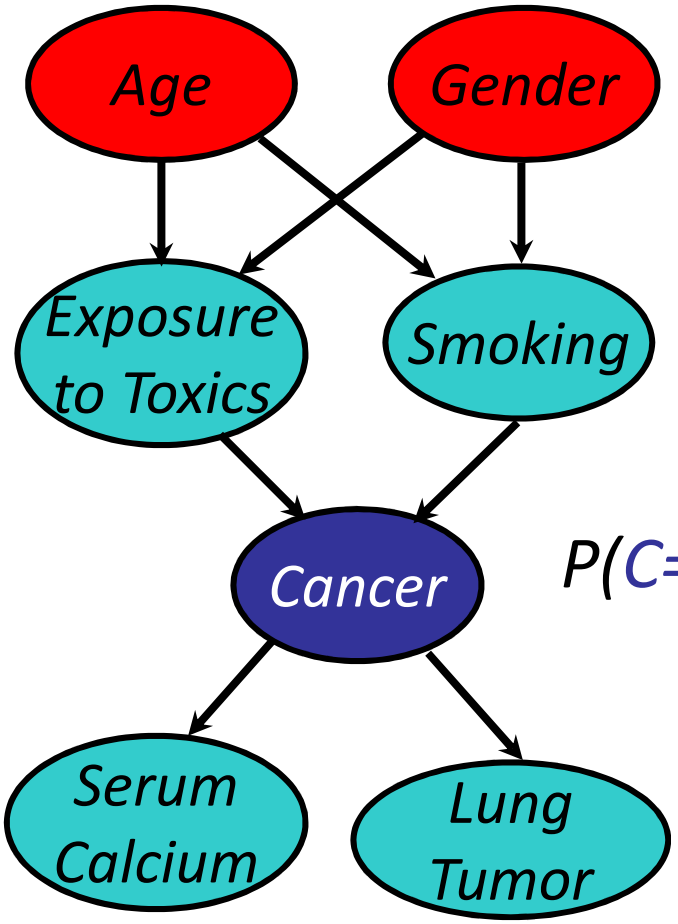
BBNs support three main kinds of reasoning:

- **Predicting** conditions given predispositions
- **Diagnosing** conditions given symptoms (and predisposing)
- **Explaining** a condition by one or more predispositions

To which we can add a fourth:

- **Deciding** on an action based on probabilities of the conditions

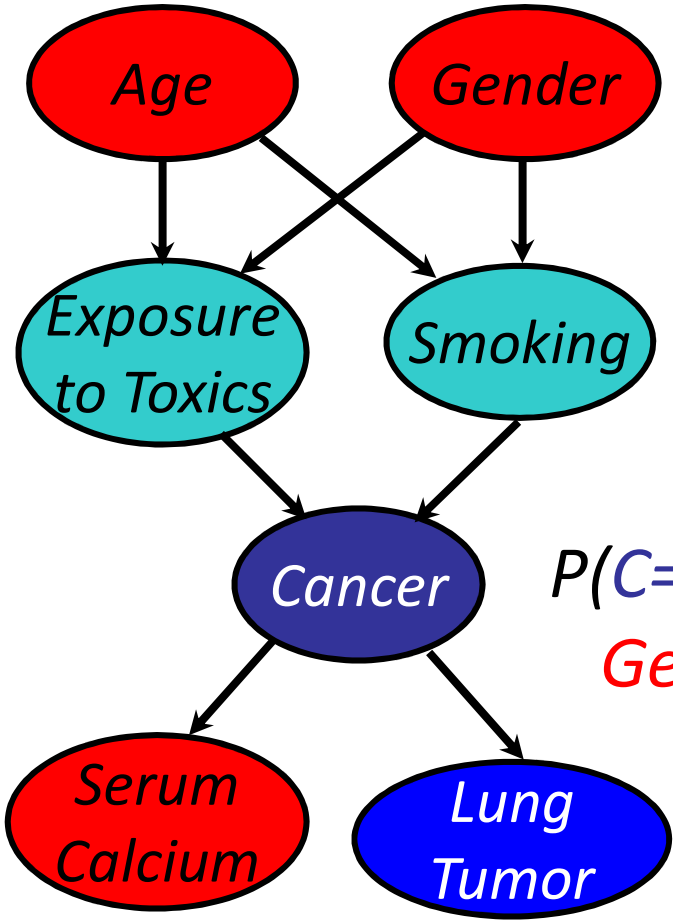
Predictive Inference



How likely are elderly males to get malignant cancer?

P(C=malignant | Age>60, Gender=male)

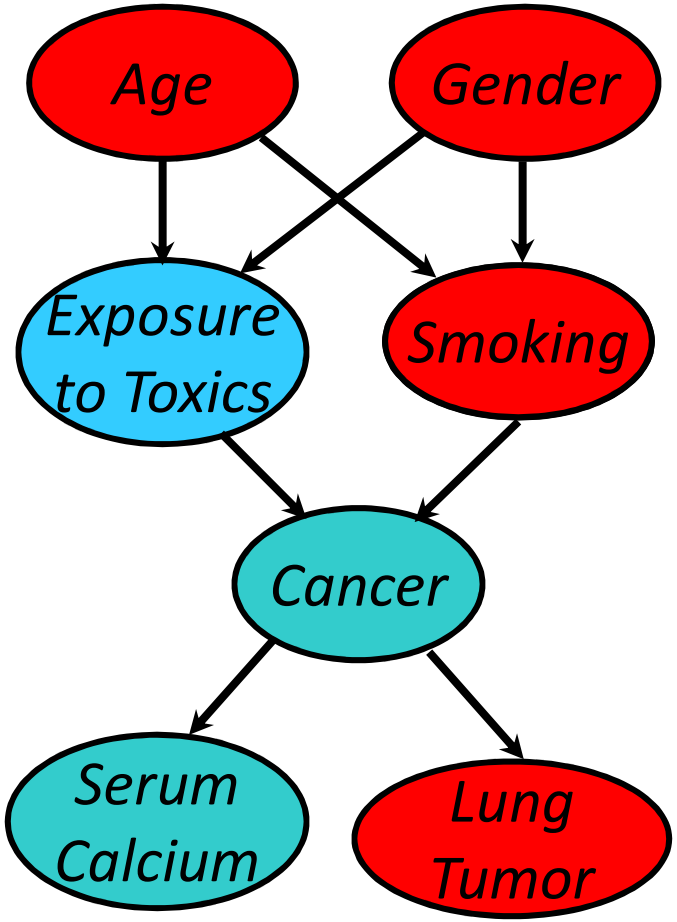
Predictive and diagnostic combined



How likely is an elderly male patient with high Serum Calcium to have malignant cancer?

P(C=malignant | Age>60, Gender= male, Serum Calcium = high)

Explaining away



- If we see a lung tumor, the probability of heavy smoking and of exposure to toxics both go up

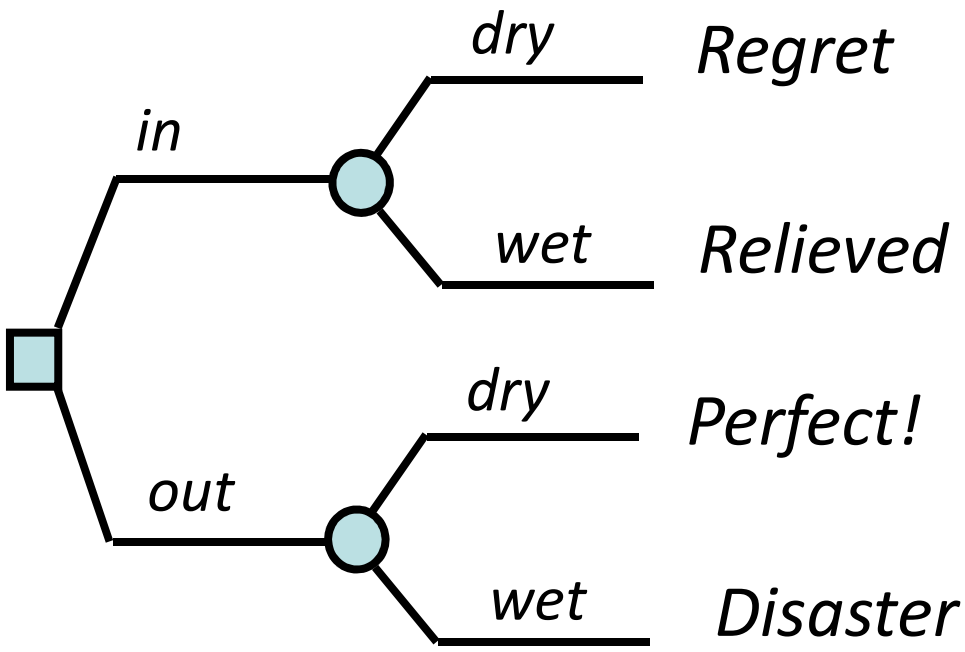
- If we then observe heavy smoking, the probability of exposure to toxics goes back down

Decision making

- A decision is a medical domain might be a choice of treatment (e.g., radiation or chemotherapy)
- Decisions should be made to maximize expected utility
- View decision making in terms of
  - Beliefs/Uncertainties
  - Alternatives/Decisions
  - Objectives/Utilities

Decision Problem

Should I have my party inside or outside?



Value Function

A numerical score over all possible states allows a BBN to be used to make decisions

Location?	Weather?	Value
in	dry	$50
in	wet	$60
out	dry	$100
out	wet	$0

Using $ for the value helps our intuition

Decision Making with BBNs

- Today's weather forecast might be either sunny, cloudy or rainy
- Should you take an umbrella when you leave?
- Your decision depends only on the forecast
  - The forecast "depends on" the actual weather
- Your satisfaction depends on your decision and the weather
  - Assign a utility to each of four situations: (rain|no rain) x (umbrella, no umbrella)

Decision Making with BBNs

- Extend BBN framework to include two new kinds of nodes: **decision** and **utility**
- **Decision** node computes the expected utility of a decision given its parent(s) (e.g., forecast) and a valuation
- **Utility** node computes utility value given its parents, e.g. a decision and weather
  - Assign utility to each situations: (rain|no rain) x (umbrella, no umbrella)
  - Utility value assigned to each is probably subjective