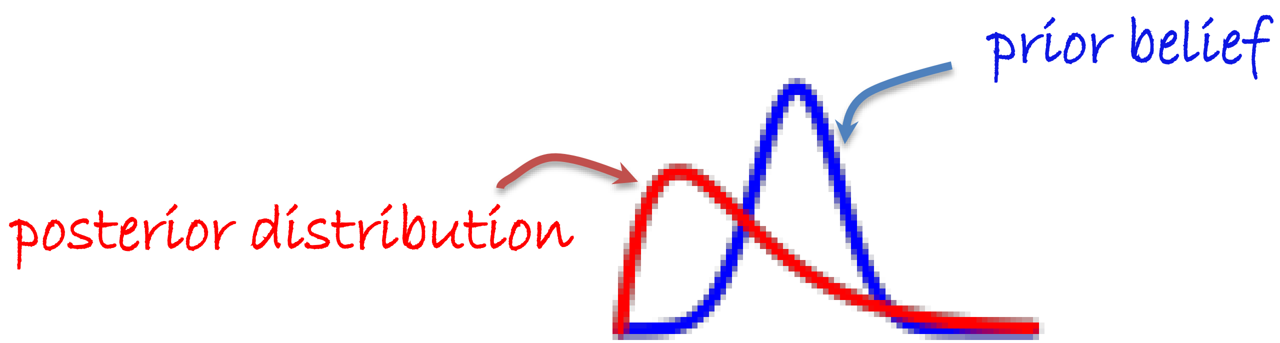
Bayesian
Reasoning, Inference, and Estimation



Nagiza F. Samatova, samatova@csc.ncsu.edu
Professor, Department of Computer Science
North Carolina State University

Senior Scientist, Computer Science & Mathematics Division
Oak Ridge National Laboratory

Oak Ridge National Laboratory

NC STATE UNIVERSITY
Department of Computer Science

Applications of Bayesian Inference

- Bayesian Classification:
  - Classify e-mails (Spam, NotSpam)
  - Classify sentiments
  - Classify Documents (web pages, news articles)
- Bayesian Parameter Estimation
- Bayesian Regression
- Reasoning under uncertainty

Application: Naive Bayes Classifier in R

There are at least two packages for Naive Bayes on CRAN:

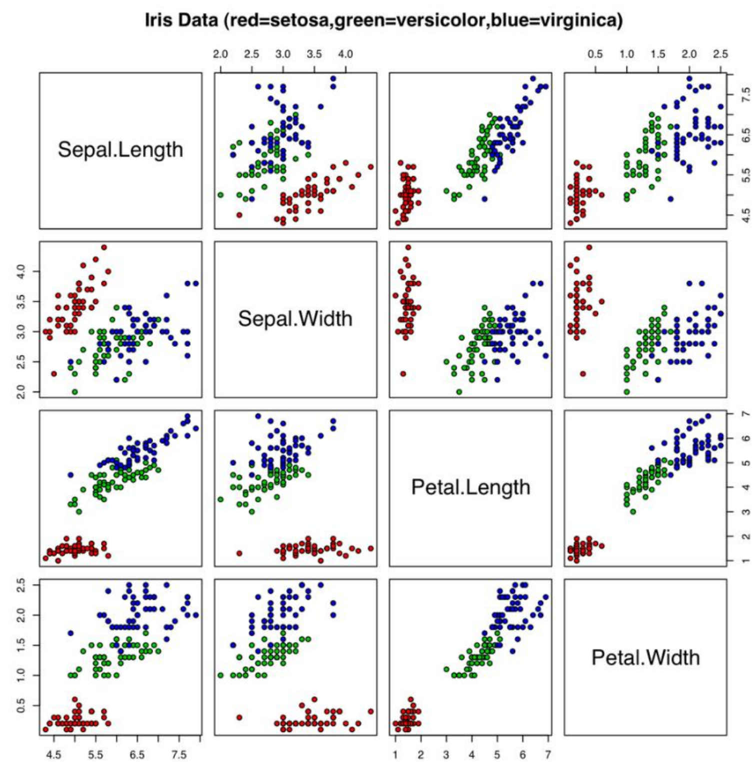
- **e1071**
- **klaR**

```
install.packages ('e1071', dependencies = TRUE)
library (class)
library (e1071)

pairs(iris[1:4], main = "Iris Data (red=setosa,
green=versicolor, blue=virginica)", pch = 21, bg =
c("red", "green3", "blue")[unclass(iris$Species)])

classifier <- naiveBayes (iris[, 1:4], iris[, 5])
table (predict (classifier, iris[, -5]), iris[, 5])
```

	setosa	versicolor	virginica
setosa	50	0	0
versicolor	0	47	3
virginica	0	3	47



Application: Bayesian Belief Network Classifier

There is at least one packages for BBN on R's CRAN:

- **bnlearn (see examples in the provided documentation)**

Bayes Theorem

$\boldsymbol{M}$: **Married** $\boldsymbol{K}$: **has Kids**

What is the probability $\boldsymbol{P(M \text{ and } K)}$?

$$
\begin{aligned}
\boldsymbol{P(M \text{ and } K)} &= \\
&= \boldsymbol{P(M)P(K|M)} \\
&= \boldsymbol{P(K)P(M|K)}
\end{aligned}
$$

$$\boxed{\boldsymbol{P(K|M) = \frac{P(K)P(M|K)}{P(M)}}}$$

Diachronic Interpretation of Bayes Thm

H: Hypothesis
E: Evidence

prior beliefs before seeing the evidence

likelihood of observing the evidence if H is correct

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

posterior probability

likelihood of the evidence under any circumstances; normalizing *constant*

Diachronic means through time:

- P(H | E): What is the probability of my hypothesis given that I have seen some new evidence, or
- **if you see some new evidence, then you can update your belief in your hypothesis**

In Bayes World: Posterior ~ Prior * Likelihood

H: Hypothesis

E: Evidence

prior beliefs before seeing the evidence

likelihood of observing the evidence if H is correct

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

posterior probability

likelihood of the evidence under any circumstances

$$P(disease \mid symptoms) = \frac{P(disease)\, P(symptoms \mid disease)}{P(symptoms)}$$

$$\sim P(disease)\, P(symptoms \mid disease)$$

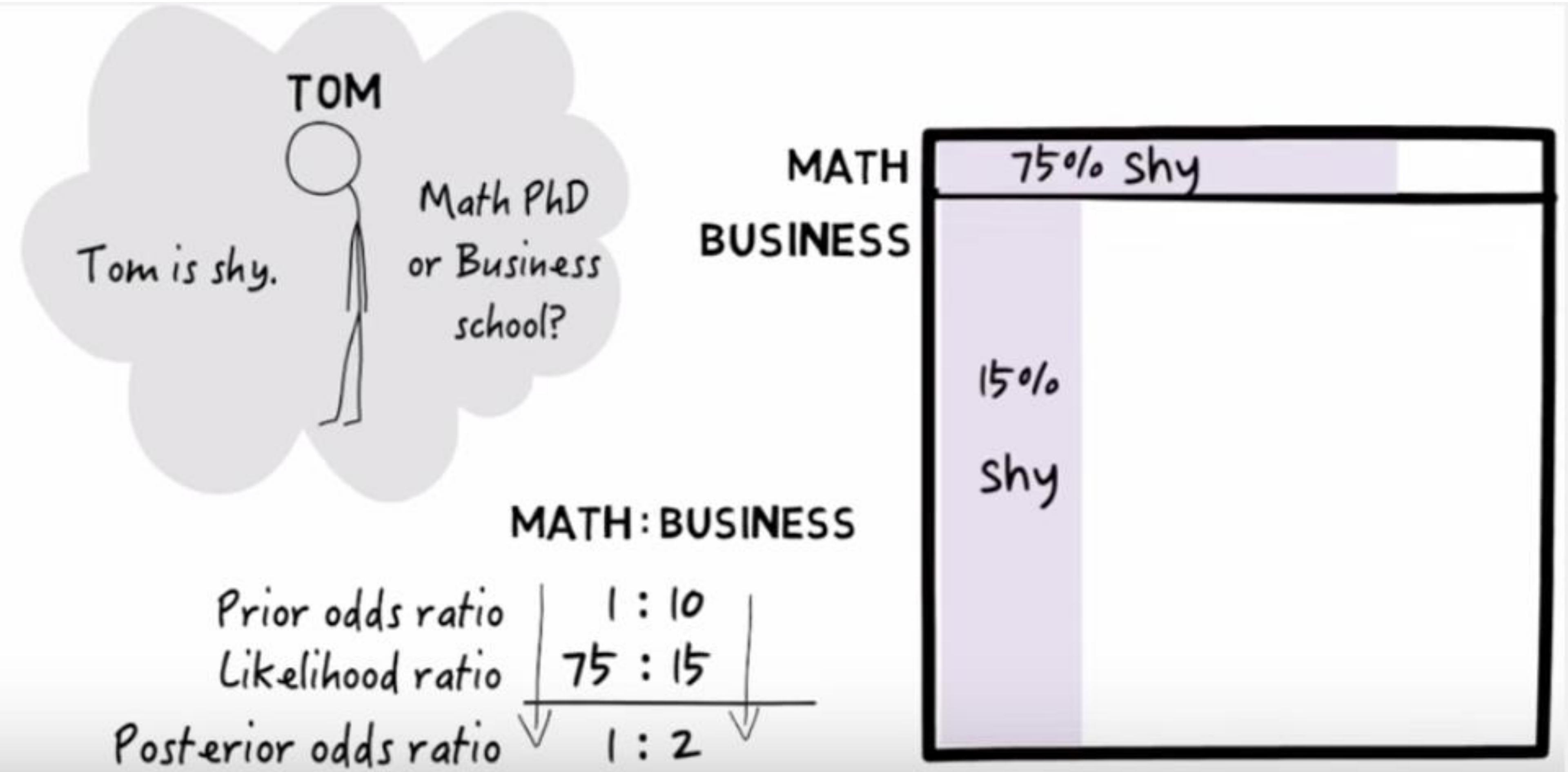
Example of Bayes Theorem

- **Given:**
  - *Prior* probability (***belief***) of any patient having meningitis is 1/50,000, **P (M)**
  - A doctor knows that meningitis causes stiff neck 50% of the time, **P (S | M)** (***likelihood***)
  - Prior probability of any patient having stiff neck is 1/20, **P(S)**
- **If a patient has stiff neck, what's the probability he/she has meningitis, P(M | S), i.e. *posterior probability*?**

$$P(M \mid S) = \frac{P(S \mid M) P(M)}{P(S)} = \frac{0.5 \times 1/50000}{1/20} = 0.0002$$

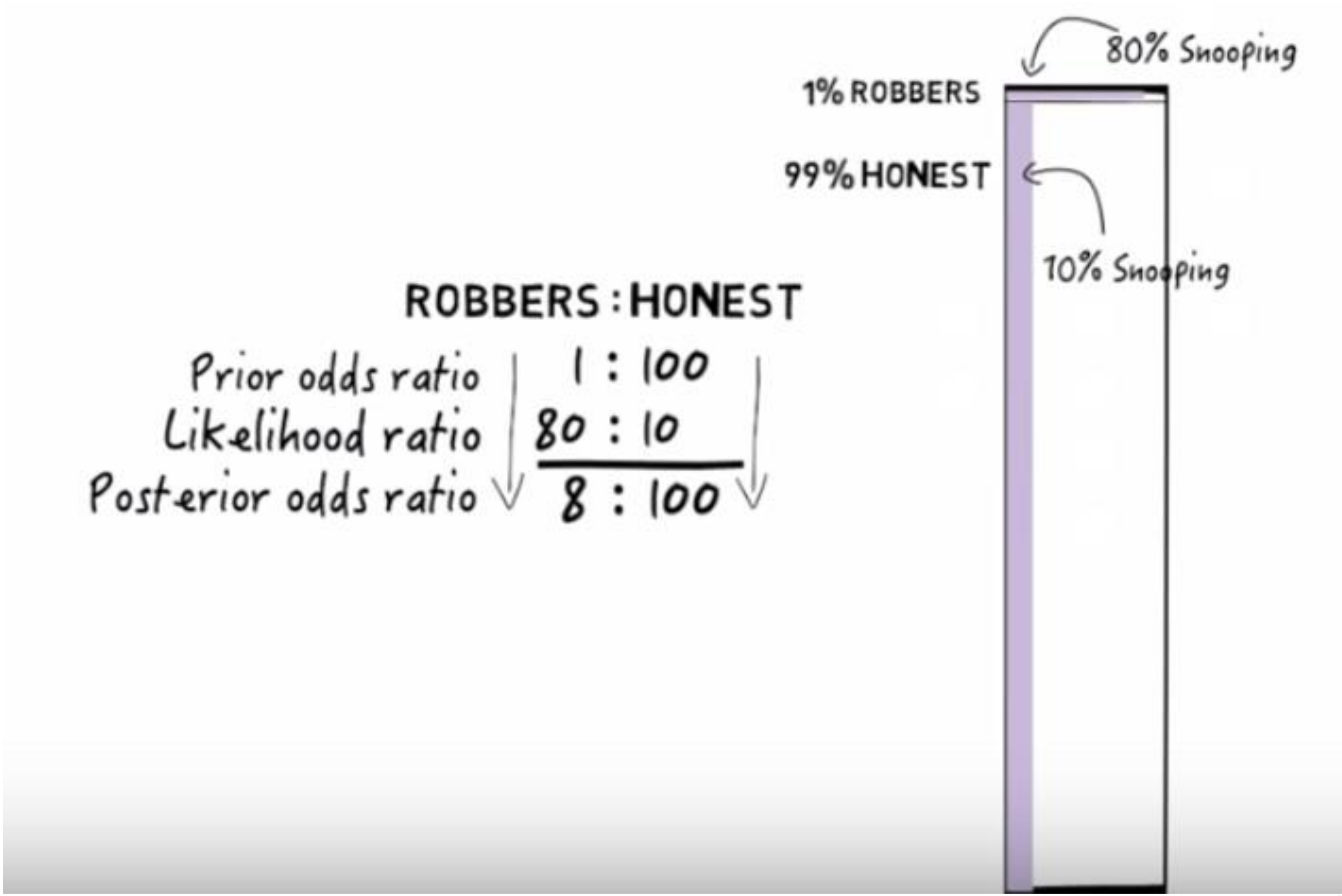
$$\boxed{P(disease \mid symptoms) = \frac{P(disease)\, P(symptoms \mid disease)}{P(symptoms)}}$$

Example: What is the posterior probability of a Shy Tom to be a mathematician?



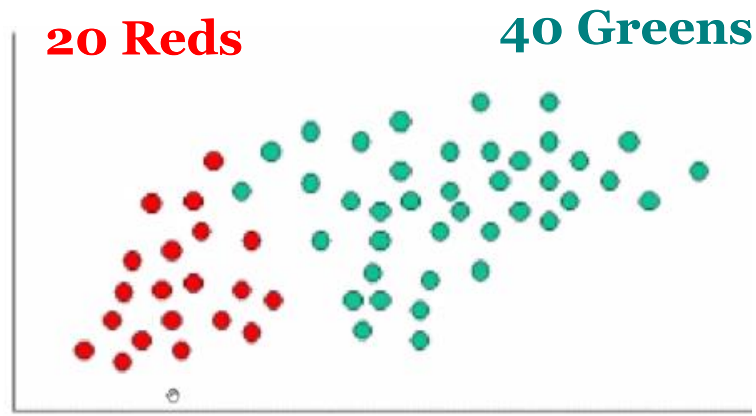
Src: Julia Galef

Example: What is Posterior Probability of the Snooping Repairman to be Honest?



Src: Julia Galef

Example: Prior Belief, P(H)



There are twice as many **GREENs** as **REDs** → the ***prior probability*** (***belief***) of **GREEN**:

$$P\,(\,H = \text{Green}\,) = \frac{40}{60} = \frac{4}{6} = \frac{2}{3}$$

$$P\,(\,H = \text{Red}\,) = \frac{20}{60} = \frac{2}{6} = \frac{1}{3}$$

Example: Likelihood, P(E | H)

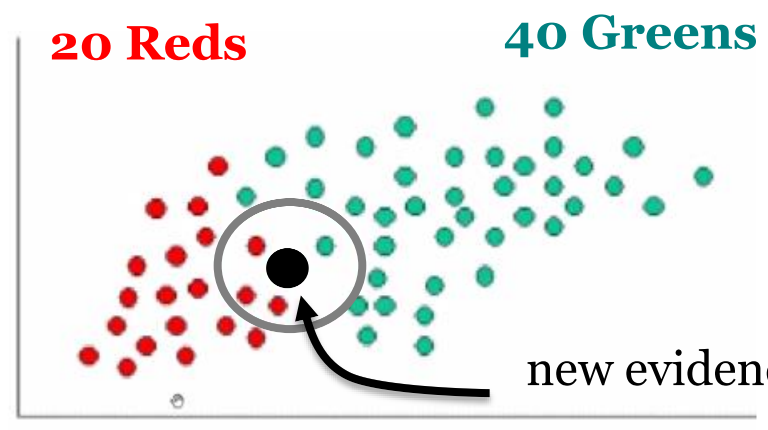
The ***likelihood*** of the new Evidence given H is correct:

$$\mathbf{P\ (E \mid H = Green) = \frac{1}{40}}$$

$$\mathbf{P\ (E \mid H = Green) = \frac{3}{20}}$$

Src: Ethm Alpaydin

Example: Posterior, P(H | E)



new evidence, E in the vicinity

The **posterior** probability:

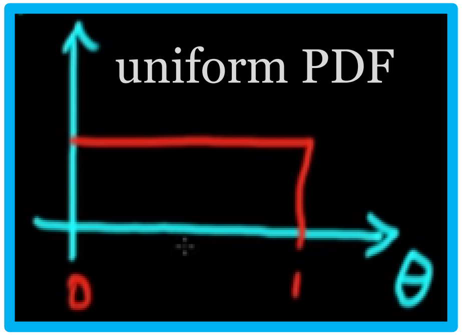
$$\mathbf{P\ (H = Green \mid E)} \sim \frac{2}{3} * \frac{1}{40}$$

$$\mathbf{P\ (H = Red \mid E)} \sim \frac{1}{3} * \frac{3}{20}$$

Model-based View on Bayesian Inference

prior beliefs about model parameters: pre-experimental knowledge of parameter values

likelihood of obtaining this data given our choice of θ

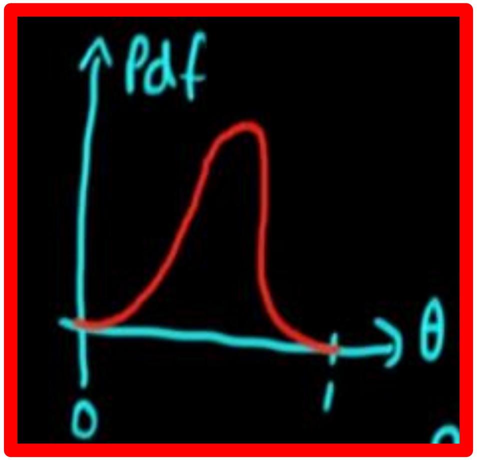


$$P(\theta \mid data) = \frac{P(\theta)\, P(data \mid \theta)}{P(data)}$$

posterior distribution

likelihood of the evidence under any circumstances

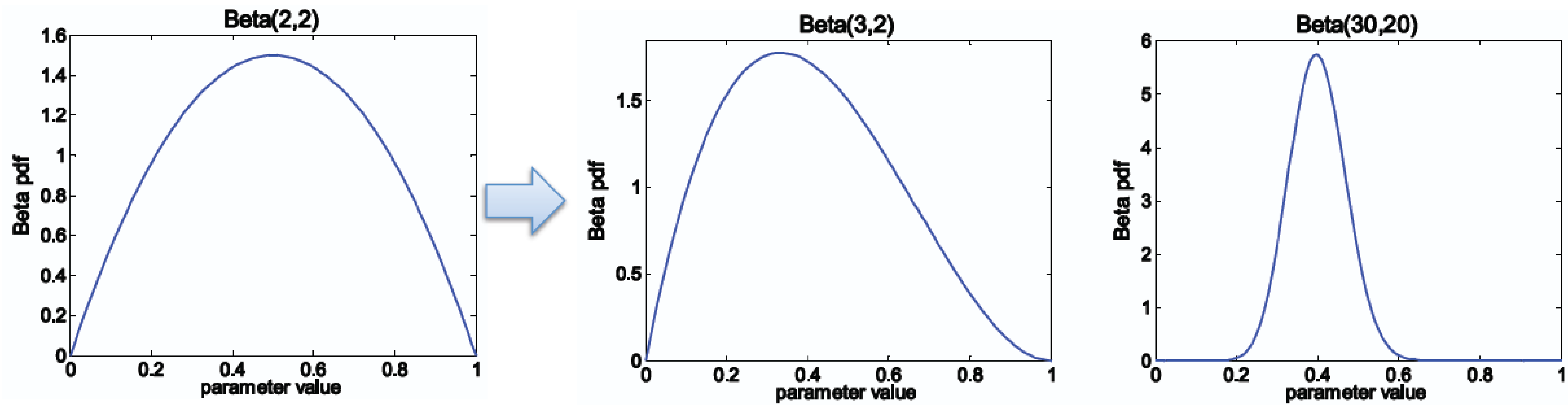
probability density function (PDF)



As the amount of data that you collect increases, then the priors plays less and less in terms of determining the posterior

Diachronic Evolution of Beta Conjugate Prior

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



As $n = \alpha_H + \alpha_T$ increases

As we get more samples, effect of prior is “washed out”

"In Bayesian probability theory, if the posterior distributions $p(\theta|x)$ are in the same family as the prior probability distribution $p(\theta)$, the prior and posterior are then called **conjugate distributions,** and the prior is called a **conjugate prior** for the likelihood function." (Wikipedia)

Frequentist vs. **Bayesian** $\equiv$ **MLE** vs. **MAP**

- Maximum Likelihood estimation (MLE)

  Choose value that maximizes the probability of observed data

$$\widehat{\theta}_{MLE} = \arg\max_{\theta} P(D|\theta)$$

- Maximum *a posteriori* (MAP) estimation

  Choose value that is most probable given observed data and prior belief

$$\begin{aligned}\widehat{\theta}_{MAP} &= \arg\max_{\theta} P(\theta|D) \\ &= \arg\max_{\theta} P(D|\theta)P(\theta)\end{aligned}$$

Informally, Frequentist vs. Bayesian

Frequentist: Sampling is infinite and decision rules can be sharp. Data are a repeatable random sample - there is a frequency. Underlying **parameters are fixed** i.e. they remain constant during this repeatable sampling process.

Bayesian: Unknown quantities are treated probabilistically and the state of the world can always be updated. Data are observed from the realized sample. Parameters are unknown and described probabilistically. It is the data which are fixed.

Src: http://stats.stackexchange.com/questions/22/bayesian-and-frequentist-reasoning-in-plain-english

or funny.., Frequentist vs. Bayesian

A Bayesian is one who, vaguely expecting a horse, and catching a glimpse of a donkey, strongly believes he has seen a mule.

From this site:
http://www2.isye.gatech.edu/~brani/isyebayes/jokes.htm

What if there are multiple attributes?

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

- ***Prior Belief*** **based on prior Class distribution:**

  Prior P(C) = N_c/N

  – e.g., Prior P(No) = 7/10,
  Prior P(Yes) = 3/10

- ***Likelihood*** **for each attribute:**

  $$P(A_i \mid C_j) = |A_{ij}| / N_c$$

  – where $|A_{ij}|$ is number of instances having attribute A_i and belong to class C_j

  – Examples:

  P (Status=Married | No) = 4/7

  P(Refund=Yes | Yes) = 0

How to combine individual likelihoods to get posterior probability?

Bayesian Classifiers

- **Consider each feature/attribute and class label as random variables**

- **Given a record with attributes (A_1, A_2,…,A_n)**
  - Goal is to predict class C
  - Specifically, we want to find the value of C (e.g., YES or NO) that maximizes posterior probability P (C | A_1, A_2,…,A_n)

- **Can we estimate posterior probability**

$$P\ (C\ |\ A_1,\ A_2,\ldots,A_n\ )$$

directly from the data?

Bayesian Classifiers

- **Approach:**
  - compute the ***posterior*** probability $P(C \mid A_1, A_2, \ldots, A_n)$ for all values of C using the **Bayes theorem**

$$\boxed{P(C \mid A_1, A_2, \ldots, A_n) = \frac{P(C)\, P(A_1, A_2, \ldots, A_n \mid C)}{P(A_1, A_2, \ldots, A_n)}}$$

  - Choose value of C that maximizes
    $\mathbf{P(C \mid A_1, A_2, \ldots, A_n)}$

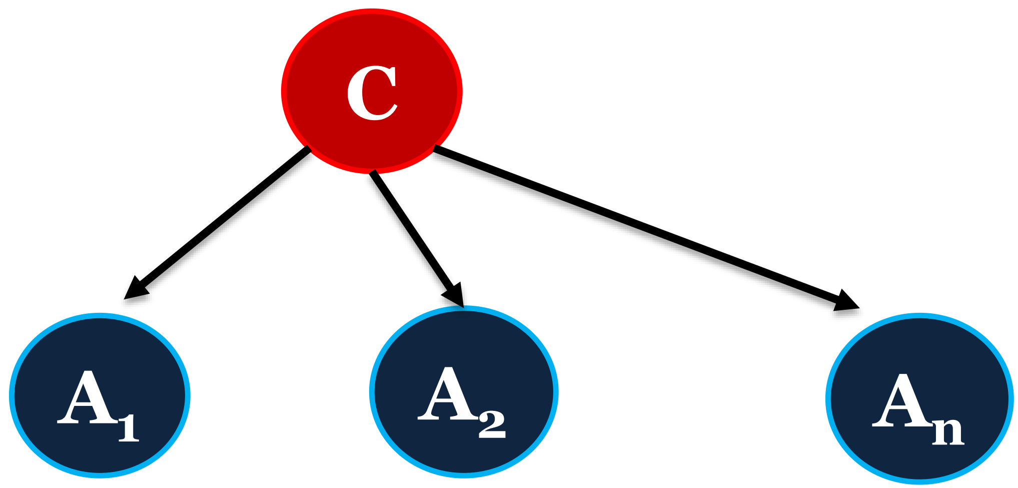
  - Equivalent to choosing value of C that maximizes
    $\mathbf{P(A_1, A_2, \ldots, A_n \mid C)} * \mathbf{P(C)}$

- **How to estimate $P(A_1, A_2, \ldots, A_n \mid C)$?**

Naïve Bayes Classifier: Product of Independent Likelihoods

Assume independence among attributes A_i when class is given:

$$P(A_1, A_2, \ldots, A_n \mid C) = P(A_1 \mid C)\, P(A_2 \mid C) \ldots P(A_n \mid C)$$



estimate $P(A_i \mid C_j)$ for all A_i and $C=C_j$

Posterior Probability in Naïve Bayes Classifier

Assuming independence among attributes A_i when class is given, posterior probability of new point/evidence to belong to class C_j:

$$P(C=C_j \mid A_1, A_2, \ldots, A_n) \sim P(C_j)\, P(A_1, A_2, \ldots, A_n \mid C)$$

$$\sim P(C_j)\, P(A_1 \mid C_j)\; P(A_2 \mid C_j) \ldots\; P(A_n \mid C_j)$$

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

$P(\text{Evade} = \text{No}) = 7/10$

$P(\text{Status=Married} \mid \text{No}) = 4/7$
$P(\text{Refund=Yes} \mid \text{No}) = 3/7$
$P(\text{Income} > 100K \mid \text{No}) = 4/7$

$P(\text{Evade} = \text{No} \mid \text{Status=Married}, \text{Refund=Yes}, \text{Income} > 100K)$
$= 7/10 * 4/7 * 3/7 * 4/7$

Example of Naïve Bayes Classifier

	Name	Give Birth	Can Fly	Live in Water	Have Legs	Class
1	human	yes	no	no	yes	mammals
2	python	no	no	no	no	non-mammals
3	salmon	no	no	yes	no	non-mammals
4	whale	yes	no	yes	no	mammals
5	frog	no	no	sometimes	yes	non-mammals
6	komodo	no	no	no	yes	non-mammals
7	bat	yes	yes	no	yes	mammals
8	pigeon	no	yes	no	yes	non-mammals
9	cat	yes	no	no	yes	mammals
10	leopard shark	yes	no	yes	no	non-mammals
11	turtle	no	no	sometimes	yes	non-mammals
12	penguin	no	no	sometimes	yes	non-mammals
13	porcupine	yes	no	no	yes	mammals
14	eel	no	no	yes	no	non-mammals
15	salamander	no	no	sometimes	yes	non-mammals
16	gila monster	no	no	no	yes	non-mammals
17	platypus	no	no	no	yes	mammals
18	owl	no	yes	no	yes	non-mammals
19	dolphin	yes	no	yes	no	mammals
20	eagle	no	yes	no	yes	non-mammals

A: attributes

M: mammals

N: non-mammals

$$P(A|M) = \frac{6}{7} \times \frac{6}{7} \times \frac{2}{7} \times \frac{2}{7} = 0.06$$

$$P(A|N) = \frac{1}{13} \times \frac{10}{13} \times \frac{3}{13} \times \frac{4}{13} = 0.0042$$

$$P(A|M)P(M) = 0.06 \times \frac{7}{20} = 0.021$$

$$P(A|N)P(N) = 0.004 \times \frac{13}{20} = 0.0027$$

Give Birth	Can Fly	Live in Water	Have Legs	Class
yes	no	yes	no	?

P(A|M)P(M) > P(A|N)P(N)

=> Mammals

Exercise: Play Tennis?

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
Day1	Sunny	Hot	High	Weak	*No*
Day2	Sunny	Hot	High	Strong	*No*
Day3	Overcast	Hot	High	Weak	Yes
Day4	Rain	Mild	High	Weak	Yes
Day5	Rain	Cool	Normal	Weak	Yes
Day6	Rain	Cool	Normal	Strong	*No*
Day7	Overcast	Cool	Normal	Strong	Yes
Day8	Sunny	Mild	High	Weak	*No*
Day9	Sunny	Cool	Normal	Weak	Yes
Day10	Rain	Mild	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes
Day12	Overcast	Mild	High	Strong	Yes
Day13	Overcast	Hot	Normal	Weak	Yes
Day14	Rain	Mild	High	Strong	*No*

Question: For the day <sunny, cool, high, strong>, what's the play prediction?

Naive Bayes is often used in Text Mining

Baseline: Naive Bayes Text Classification

- $P(\mathbf{X}|Y)$ is huge!!!
  - Article at least 1000 words, $\mathbf{X}=\{X_1,\ldots,X_{1000}\}$
  - X_i represents i^{th} word in document, i.e., the domain of X_i is entire vocabulary, e.g., Webster Dictionary (or more), 10,000 words, etc.

- NB assumption helps a lot!!!
  - $P(X_i=x_i|Y=y)$ is just the probability of observing word x_i at the i^{th} position in a document on topic y

$$h_{NB}(\mathbf{x}) = \arg\max_y P(y) \prod_{i=1}^{LengthDoc} P(x_i|y)$$

Bag of Words Model

- Typical additional assumption – **Position in document doesn't matter**: $P(X_i=x_i|Y=y) = P(X_k=x_i|Y=y)$
  - "Bag of words" model – order of words on the page ignored
  - Sounds really silly, but often works very well!

$$\prod_{i=1}^{LengthDoc} P(x_i|y) = \prod_{w=1}^{W} P(w|y)^{count_w}$$

Text → Bag of Words


the world of
TOTAL

▸ All About The Company
Global Activities
Corporate Structure
TOTAL's Story
Upstream Strategy
Downstream Strategy
Chemicals Strategy
TOTAL Foundation
Homepage

all about the **company**

Our energy exploration, production, and distribution operations span the globe, with activities in more than 100 countries.

At TOTAL, we draw our greatest strength from our fast-growing oil and gas reserves. Our strategic emphasis on natural gas provides a strong position in a rapidly expanding market.

Our expanding refining and marketing operations in Asia and the Mediterranean Rim complement already solid positions in Europe, Africa, and the U.S.

Our growing specialty chemicals sector adds balance and profit to the core energy business.

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

NB Classifier for Bag of Words

- Learning phase: using multiple training documents
  - Class Prior P(Y)
  - $P(X_i|Y)$
- Test phase:
  - For each test document, use naïve Bayes decision rule:

$$
\begin{aligned}
h_{NB}(\mathbf{x}) &= \arg\max_y P(y) \prod_{i=1}^{LengthDoc} P(x_i|y) \\
&= \arg\max_y P(y) \prod_{w=1}^{W} P(w|y)^{count_w}
\end{aligned}
$$

Twenty News Groups Results

89% Naive Bayes Classification Accuracy of which news group a document belongs to.

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	

How to Estimate Probabilities from Continuous Data?

- **For continuous attributes:**
  - Discretize the range into bins
    - one binary attribute per bin
    - violates independence assumption
  - Two-way split: (A < v) or (A > v)
    - choose only one of the two splits as new attribute
  - Probability density estimation:
    - Test that an attribute follows a certain distribution (e.g., normal distribution) for **each Class label separately**
    - Use data to estimate parameters of distribution (e.g., mean and standard deviation)
    - Once probability distribution is known, can use it to estimate the conditional probability $P(A_i \mid c)$

Normal Distribution Test

Use a **Shapiro-Wilk** test & some qqplots

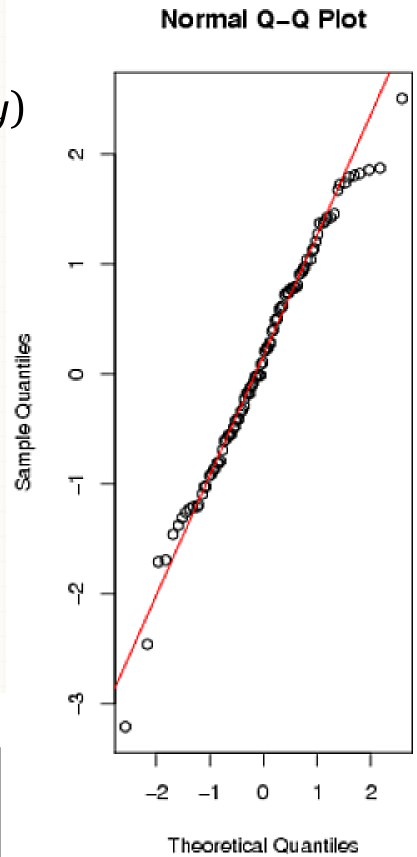
(*when the p-value is lower than 0.05, you can conclude that the sample deviates from normality*)

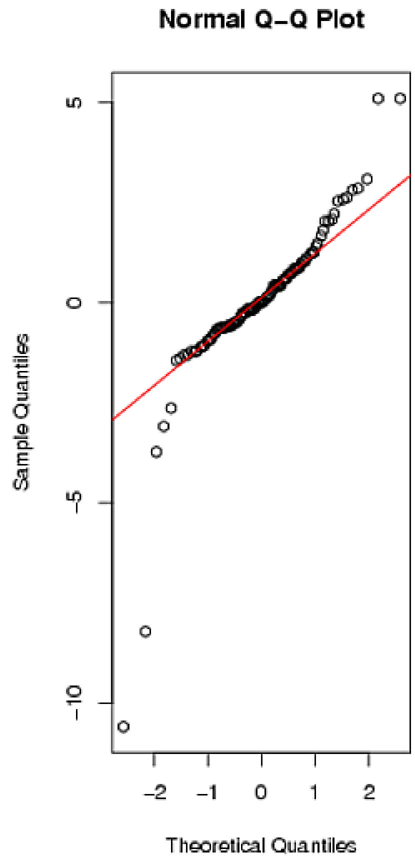
```
## Generate two data sets
## First Normal, second from a t-distribution
words1 = rnorm(100); words2 = rt(100, df=3)

## Have a look at the densities
plot(density(words1));plot(density(words2))

## Perform the test
shapiro.test(words1); shapiro.test(words2)

## Plot using a qqplot
qqnorm(words1);qqline(words1, col = 2)
qqnorm(words2);qqline(words2, col = 2)
```



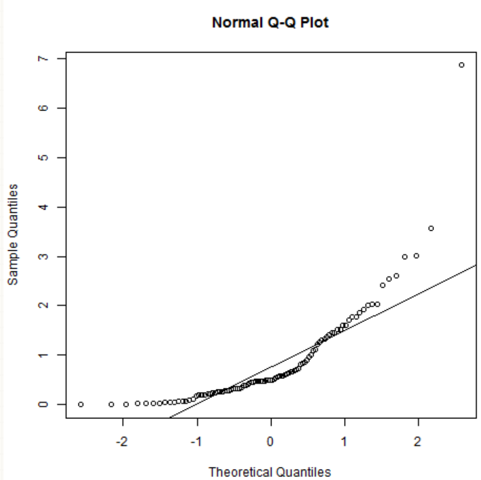


Non-normal (Gamma) distribution

```
y <- rgamma(100, 1)
```

The QQ-normal plot:

```
qqnorm(y); qqline(y)
```



Checking the validity of the assumption of normality in R

```
library(moments)
library(nortest)
library(e1071)

set.seed(777)
x <- rnorm(250,10,1)

# skewness and kurtosis, they should be around (0,3)
skewness(x)
kurtosis(x)

# Shapiro-Wilks test
shapiro.test(x)

# Kolmogorov-Smirnov test
ks.test(x,"pnorm",mean(x),sqrt(var(x)))

# Anderson-Darling test
ad.test(x)

# qq-plot: you should observe a good fit of the straight line
qqnorm(x)
qqline(x)

# p-plot: you should observe a good fit of the straight line
probplot(x, qdist=qnorm)

# fitted normal density
f.den <- function(t) dnorm(t,mean(x),sqrt(var(x)))
curve(f.den,xlim=c(6,14))
hist(x,prob=T,add=T)
```

How to Estimate Probabilities from Continuous Data?

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

- **Normal distribution:**

$$P(A_i \mid c_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}^2}} e^{-\frac{(A_i - \mu_{ij})^2}{2\sigma_{ij}^2}}$$

  – One for each (A_i, c_i) pair

- **For (Income, Class=No):**
  – If Class=No
    - sample mean = 110
    - sample variance = 2975

$$P(Income = 120 \mid No) = \frac{1}{\sqrt{2\pi}(54.54)} e^{-\frac{(120-110)^2}{2(2975)}} = 0.0072$$

Example of Naïve Bayes Classifier

Given a Test Record:

$$X = (\text{Refund} = \text{No}, \text{Married}, \text{Income} = 120\text{K})$$

naive Bayes Classifier:

P(Refund=Yes|No) = 3/7
P(Refund=No|No) = 4/7
P(Refund=Yes|Yes) = 0
P(Refund=No|Yes) = 1
P(Marital Status=Single|No) = 2/7
P(Marital Status=Divorced|No)=1/7
P(Marital Status=Married|No) = 4/7
P(Marital Status=Single|Yes) = 2/7
P(Marital Status=Divorced|Yes)=1/7
P(Marital Status=Married|Yes) = 0

For taxable income:

If class=No:	sample mean=110
	sample variance=2975
If class=Yes:	sample mean=90
	sample variance=25

- P(X|Class=No) = P(Refund=No|Class=No) $\times$ P(Married| Class=No) $\times$ P(Income=120K| Class=No) = 4/7 $\times$ 4/7 $\times$ 0.0072 = 0.0024
- P(X|Class=Yes) = P(Refund=No| Class=Yes) $\times$ P(Married| Class=Yes) $\times$ P(Income=120K| Class=Yes) = 1 $\times$ 0 $\times$ 1.2 $\times$ 10^{-9} = 0

Since P(X|No)P(No) > P(X|Yes)P(Yes)

Therefore P(No|X) > P(Yes|X)

=> Class = No

Naïve Bayes Classifier

- **If one of the conditional probabilities (likelihoods) is zero, then the entire expression becomes zero**
- **Probability estimation:**

$$\text{Original}: P(A_i \mid C) = \frac{N_{ic}}{N_c}$$

$$\text{Laplace}: P(A_i \mid C) = \frac{N_{ic} + 1}{N_c + c}$$

$$\text{m - estimate}: P(A_i \mid C) = \frac{N_{ic} + mp}{N_c + m}$$

c: number of classes

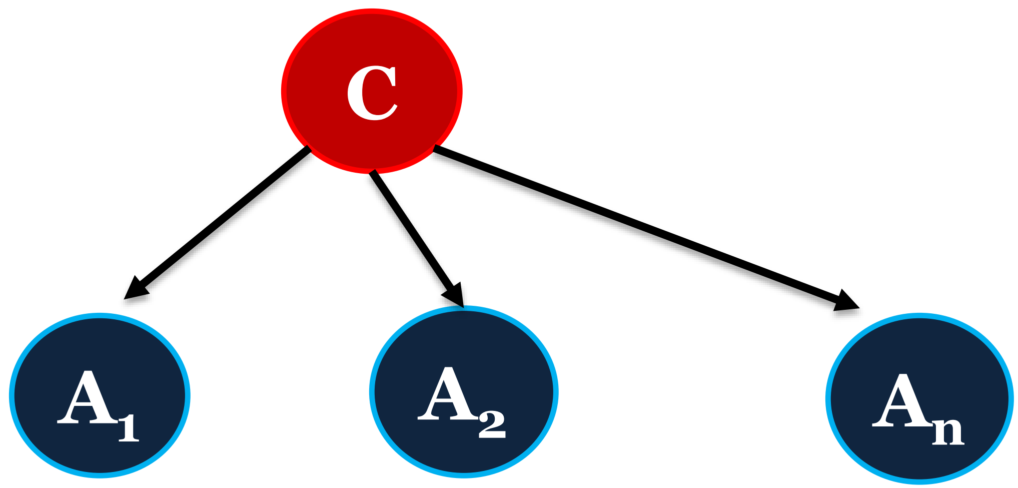
p: prior probability

m: parameter

Naïve Bayes Classifier: Product of Independent Likelihoods

Assume independence among attributes A_i when class is given:

$$P(A_1, A_2, \ldots, A_n \mid C) = P(A_1 \mid C)\; P(A_2 \mid C) \ldots\; P(A_n \mid C)$$



estimate $P(A_i \mid C_j)$ for all A_i and $C=C_j$

Bayesian Belief Network (BBN)

- BBNs relax **independence** assumption about attributes
  - Model dependencies as a Directed Acyclic Graph (**DAG**): NO CYCLES
  - With **joint probability tables** over parents for each node
- Absence of a link/edge in the DAG implies conditional independence

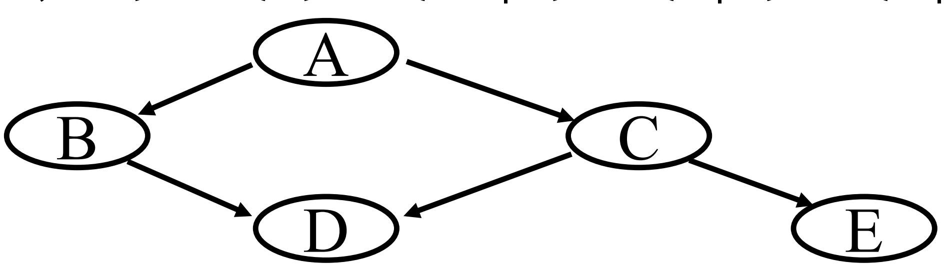
$$\mathbf{P\ (X_1, \ldots, X_n) = Product\ P(X_i \mid parents\ (X_i))}$$

- Probability A,B,C,D,E all present:

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

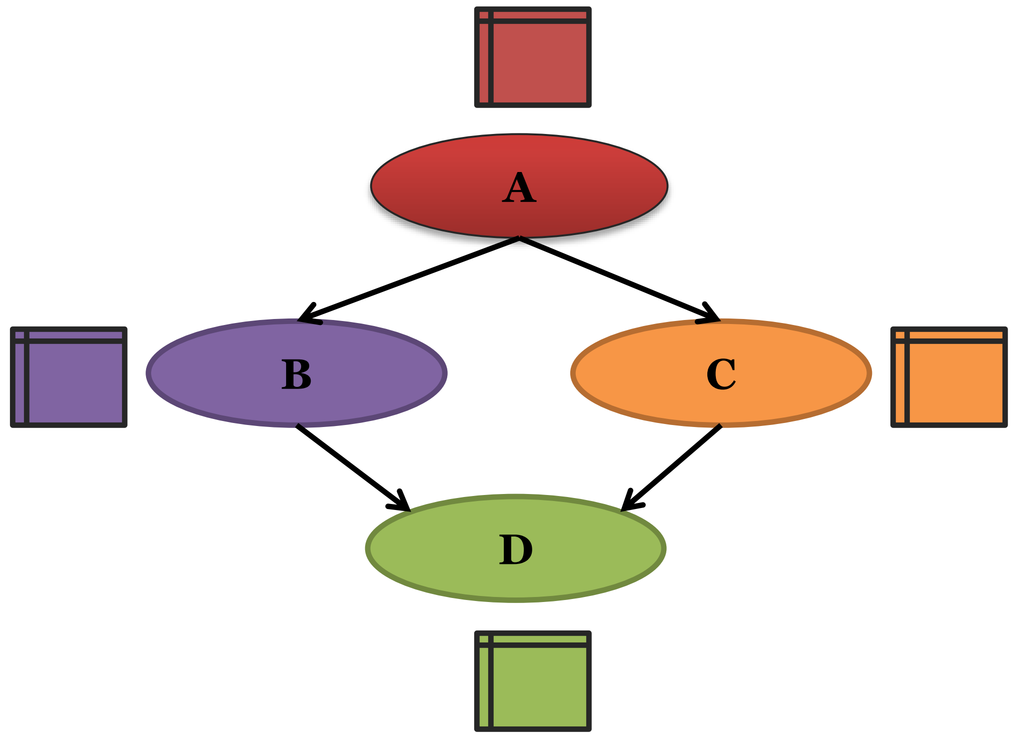
- Probability A,C,D present and B,E absent:

  P (A,¬B,C,D, ¬E) = P(A) * P(¬B|A) * P(C|A) * P(D|¬B,C) * P(¬E|C)



Bayesian Belief Network (BBN)

- Introduced by Pearl, 1985
- A BBN = <G, Θ>
- G is a directed acyclic graph, G = <V, E>
- Θ is a set of parameters
- A BBN encodes the joint probability distribution



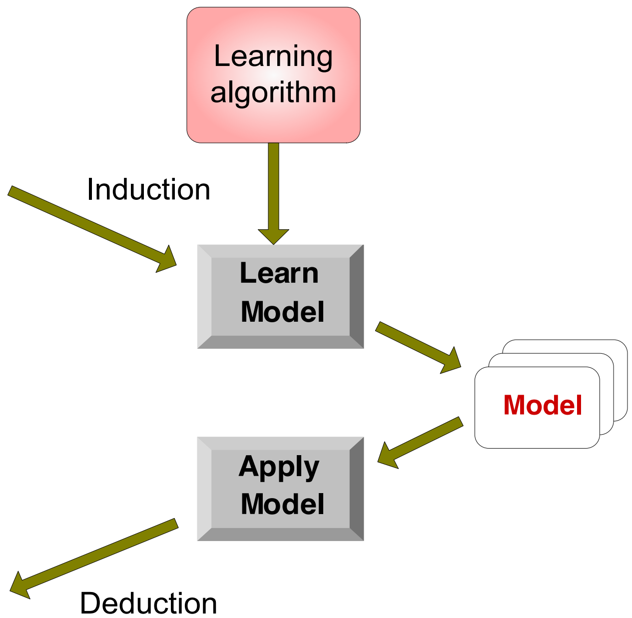
P(A,B,C,D) = P(A)P(B|A)P(C| A)P(D| B,C)

BBN as a Classifier

- As classifiers in a supervised learning
- To discover and represent dependency relationships among variables/features/ attributes in appl. domain

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

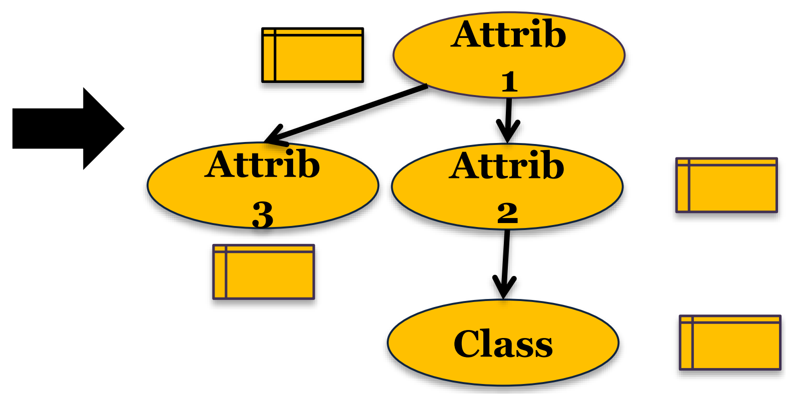
Training Set



Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Test Set

Fig. adopted from Kumar et al book



Model = Classifier

BBN is Powerful Learning Model

Characteristics

- Multiple hypotheses
- Probabilistic prediction
- Incorporation of prior knowledge
- Incomplete data
- Multinomial and continuous data
- Robust against overfitting
- Uncertainty

Application Domains

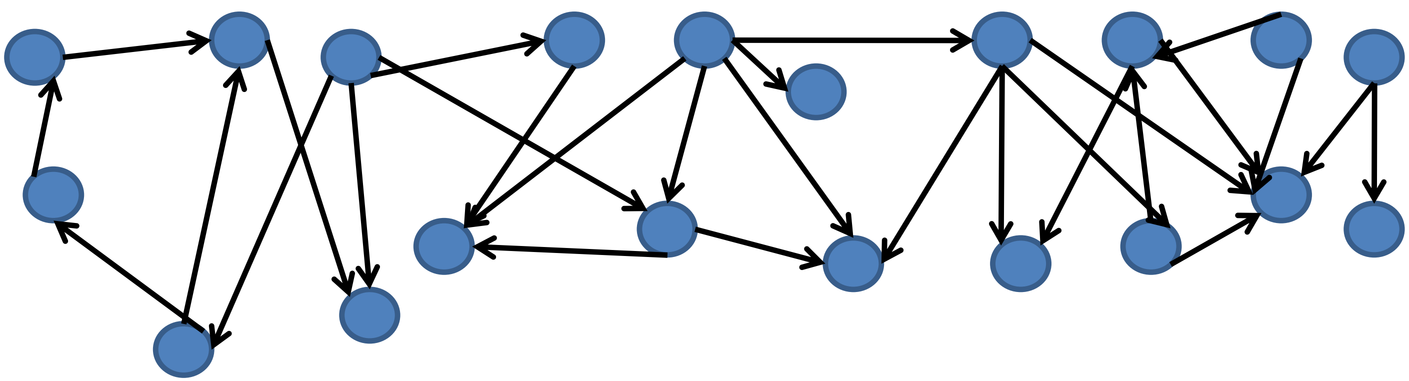
- Medical diagnostics
- Computational biology
- Bioinformatics
  - prediction of protein structure
  - gene regulatory network
- Image processing
- Gaming

The ability to incorporate prior knowledge into BBN learning allows the learner to combine knowledge from more than one sources to build a model. This is particularly beneficial in domains with already-existing domain experts like medical diagnostics. In term of machine learning, BBNs play an important role in learning under uncertainty due to its probabilistic nature.

Challenge: Learning a Structure of BBNs is Super-Exponential in Number of Features

- **Learn a BBN**
  1. Structure learning (G)
  2. Parameter learning (Θ)

Number of features	**Number of directed acyclic graphs**
4	**543**
6	**3781503**
10	**$O(10^{18})$**
N	**$O(2^{N^2})$**

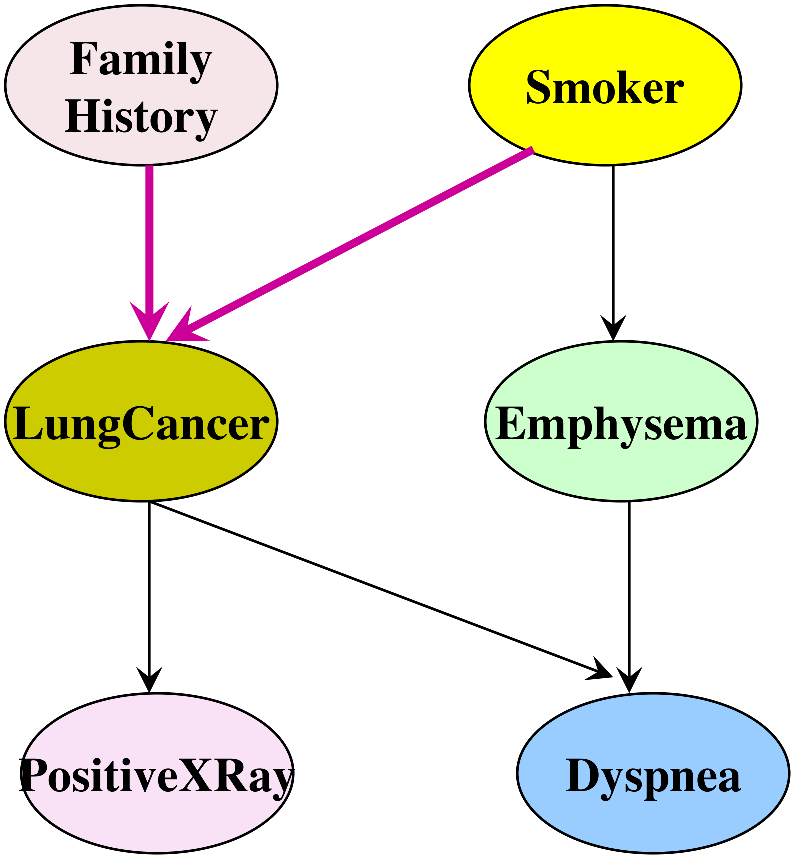


Bayesian Network Tables

- **Each node annotated with conditional probability table**
  - Probability of node values given values of parent nodes

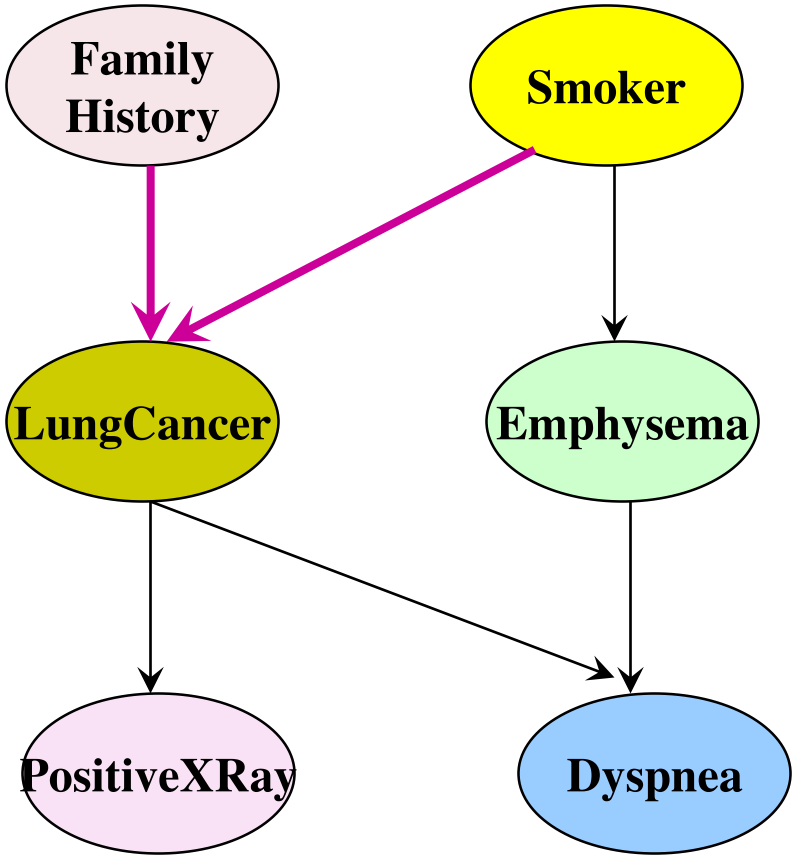
	(FH, S)	(FH, ~S)	(~FH, S)	(~FH, ~S)
LC	0.8	0.5	0.7	0.1
~LC	0.2	0.5	0.3	0.9

- **Conditional probability table for the variable LungCancer (LC)**



Bayesian networks

- Table of joint probability distribution has $2^6 = 64$ entries
- Bayesian network tables have 8 + 4 + 4 + 8 = 24 entries

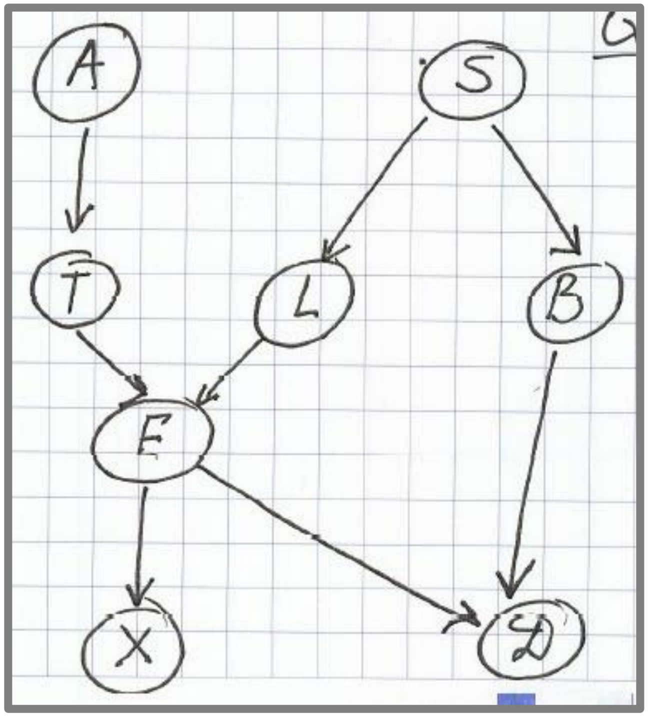


Ex: Modeling Dependencies as DAGs

Example 1 let's construct the graphical representation $G(V,E)$ for the following assumptions about variables:

1. A recent trip to Asia (A) increases the chances of tuberculosis (T)
2. Smoking is a risk factor for both lung cancer (L) and bronchitis (B)
3. The presence of either (E) tuberculosis or lung cancer can be detected by X-ray result (X), but X-ray alone can not distinguish between them
4. Dyspnoea (shortness of breath) (D) may be caused by bronchitis (B), or either (E) tuberculosis or lung cancer

Questions over Graphical Model



Q1: What is the joint probability distribution; or the probability that the patient has some combination of symptoms, test-results, and diseases?

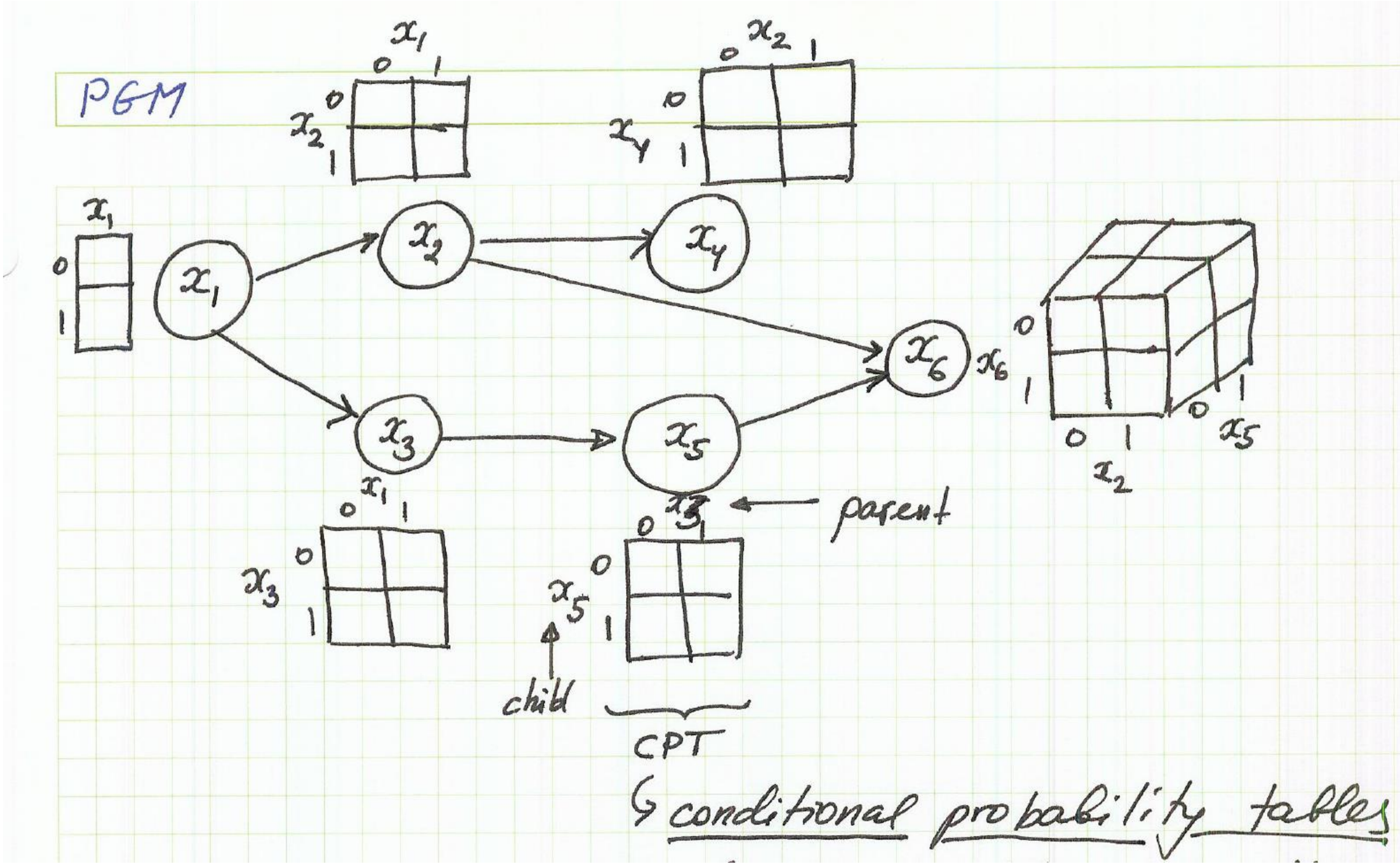
$$p(A,S,T,L,B,E,X,D) =$$
$$= p(A) \cdot p(S) \cdot p(T|A) \cdot p(L|S) \cdot p(B|S) \cdot$$
$$\cdot\, p(E|T,L) \cdot p(D|B,E) \cdot p(X|E)$$

Questions over Graphical Model

<u>Q2</u>: What is the <u>marginal</u> probability that a patient has shortness of breath?

$$P(D) = \sum_{A} \sum_{S} \sum_{T} \sum_{L} \sum_{B} \sum_{E} \sum_{X} P(A, S, T, L, B, E, X, D)$$

BBN is a Probabilistic Graphical Model (PGM)



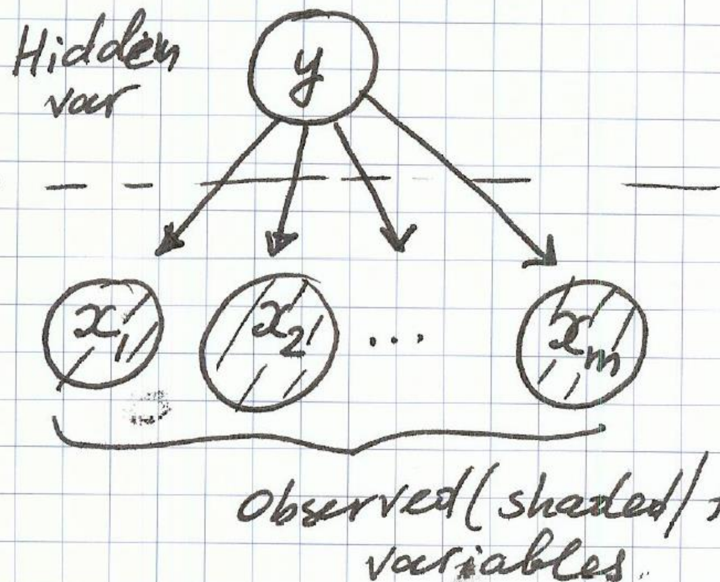
Naive Bayes Classifier Model

Input: ~~A~~ vector of observations

$$\vec{x} = (x_1, x_2, \dots, x_m)$$

Output: A predicted class variable y
(hidden / latent variable)

$$p(y \mid \vec{x}) = ?$$



(A) Naïve Bayes Network

Assumptions: All input variables x_i are cond. independent of each other

Application: email classification

Naive Bayes: Prediction

By Bayes theorem/law/rule:

$$p(y \mid \vec{x}) = \frac{p(y) \cdot p(\vec{x} \mid y)}{p(\vec{x})}$$

$p(\vec{x})$ is just a normalization constant

$$\boxed{p(y \mid \vec{x}) \sim p(y) \cdot p(\vec{x} \mid y) = p(\vec{x}, y)}$$

$$\boxed{p(y \mid \vec{x}) \sim p(\vec{x}, y) = p(y) \cdot \prod_{i=1}^{m} p(x_i \mid y)}$$

Bayesian Belief Network Model

Assumptions: A DAG describes cond. indep. among var

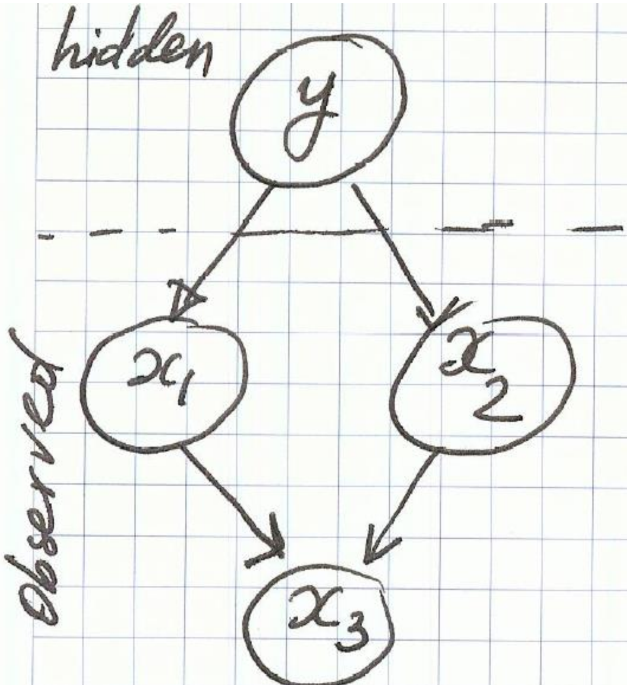
and Root is a hidden variable, y

$$p(y|\vec{x}) \sim p(y, \vec{x}) =$$

$$= p(y) \cdot \prod_{i=1}^{m} p(x_i | \underbrace{pa_i}_{\text{parents of } x_i})$$

Example:

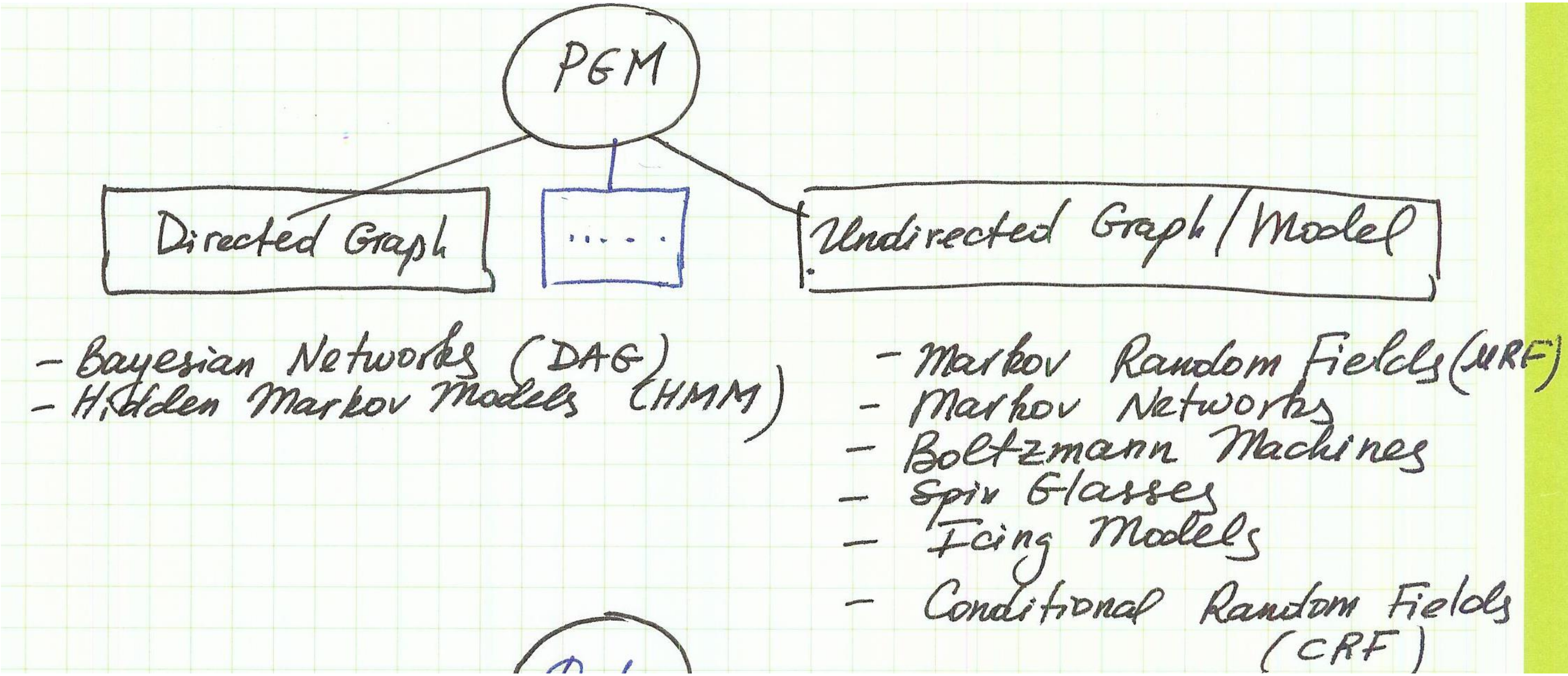
BBN: Prediction



$$p(y \mid x_1, x_2, x_3) = p(y) \cdot p(x_1 \mid y) \cdot p(x_2 \mid y) \cdot p(x_3 \mid x_1, x_2)$$

BBN Predicts: <u>Single class variable</u>
<u>that is hidden/latent</u>

PGM: Probabilistic Graphical Models



(CRF)

Data

● Hidden / latent variables, filled black nodes
$\vec{y} = (y_1, y_2, .., y_n)$

○ Observed vars, white nodes
$\vec{x} = (x_1, x_2, .., x_n)$

$\vec{y} = y_1$ single class	$\vec{y} = y_1 \to y_2 \to .. \to y_n$ sequence of var's	$\vec{y} = (y_1, y_2, .. y_n)$ a vector of var's
Bayesian Networks	HMM's	
	CRF's	MRF's
Disease diagnostics from symptoms	POS - part of speech NER - named entity	Image denoising

HMM: Hidden Markov Model

Input: an observed sequence

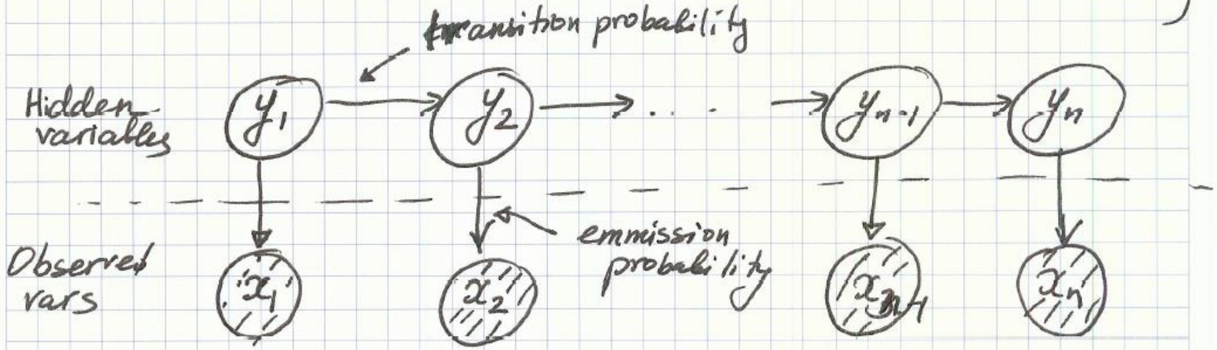
$\vec{x} = (x_1, x_2, .., x_n)$ (emission symbols)

Output: A predicted sequence of hidden/latent class labels

$\vec{y} = (y_1, y_2, .., y_n)$ (states)

Example:
- Part of speech tagging
- Named entity recognition

Input: sentences w/ words characterized by features
Output: Tags (PERSON, LOCATION, ORGANIZATION, OTHER)



$$p(\vec{y}, \vec{x}) = p(y_1) \cdot \prod_{n=2}^{N} p(y_n | y_{n-1}) p(x_n | y_n)$$

Summary: Bayesian Inference

- **Posterior probability is estimated as the product of our prior belief from prior experience/observations and the likelihood of new evidence/data if our prior belief is true**
- **Naive Bayes Classifier: assumes independence between attributes**
- **Bayesian Belief Networks (BBNs) relax this independence assumption:**
  - Model dependencies as a **DAG**, Directed Acyclic Graph
  - Learning the structure of the DAG is very expensive; better bring domain knowledge
- **Both Naive and BBN are examples of more general Probabilistic Graphical Models (PGMs):**
  - Hidden and observed attributes
  - Model dependencies as a graph (DAG or even undirected)
- **Robust to isolated noise points**
- **Handle missing values by ignoring the instance during probability estimate calculations**
- **Robust to irrelevant attributes**

Acknowledgements

- **Authors of Data Mining Book:**
  - Michael Steinbach, Vipin Kumar, etc
- **Text Classification with Naive Bayes slides**
  - Dr. Min Chi, NCSU