Probabilistic Classifiers

Master in Data Science and Advanced Analytics
BA and DS

Roberto Henriques

NOVA IMS INFORMATION MANAGEMENT SCHOOL

Bayesian Classification: Why?

- A statistical classifier
  - performs *probabilistic prediction, i.e.,* predicts class membership probabilities
- Foundation
  - Based on Bayes' Theorem
- Performance
  - A simple Bayesian classifier, *naïve Bayesian classifier*, has comparable performance with decision tree and selected neural network classifiers
- Incremental
  - Each training example can incrementally increase/decrease the probability that a hypothesis is correct — prior knowledge can be combined with observed data
- Standard
  - Even when Bayesian methods are computationally intractable, they can provide a standard of optimal decision making against which other methods can be measured

Bayes’s Theorem

- Probability of A and B occurring

$P(A \text{ and } B)$

$= P(A) * P(B \quad)$

$=P(A) * P(B|A)$

$=P(B) * P(A|B)$

Then

$P(A|B)= P(A) * P(B|A)/ P(B)$

Diacronic Bayes’s Theorem

- Probability of an hypothesis h given that I have seen some evidence E
- If we see some new evidence , we can update the believe on that hypothesis based on:

$$P(H|E) = \frac{P(H) * P(E|H)}{P(E)}$$

- P(H) – what we believed before we saw the evidence
- P(E|H) - the likelihood of seen that evidence if the hypothesis was correct
- P(E) - the likelihood of that evidence under any circumstances

Example

- Suppose we have two bowls of cookies. Bowl 1 has 10 chocolate cookies and 30 butter cookies while bowl 2 has 20 cookies of each.
- I just picked one cookie at random from one random bowl. The cookie turns out to be a butter cookie. How probable was that I took it from bowl 1?

Cookies

- H1 cookie from bowl1
- H2 cookie from bowl2

- $P(H_1) = P(H_2)$ =1/2

- $P(E|H_1)$ = 3/4
- $P(E|H_2)$ = 1/2
- $P(E)$ = 50/80 = 5/8

- $P(H_1|E) = \frac{P(H_1)*P(E|H_1)}{P(E)}$

- =$\frac{\frac{1}{2}*\frac{3}{4}}{\frac{5}{8}} = 0.6$

P(E)

- If we cannot do it by observation:

P(E) = P(H1)*P(E|H1) + P(H2)*P(E|H2)

$$P(E) = \sum P(H_i) * P(E|H_i)$$

Bayesian Theorem: Basics

- Let **X** be a data sample
  - X is considered an "*evidence*" with n attributes and class label unknown
  - X is a customer described by age and salary
- Let h be a *hypothesis* that X belongs to class C
  - Assume C is the class of **computer buyers**
- We want to determine P(h |**X**)
  - the **probability that the hypothesis holds given the observed data sample X**
  - Or the probability that X belongs to the class C given that we know the attributes of X

$$P(h|\mathbf{X}) = \frac{P(\mathbf{X}|h)P(h)}{P(\mathbf{X})}$$

$$Posterior = \frac{Likelihood \times Prior}{Evidence}$$

Bayesian Theorem: Basics

- P(**X**|h) (*likelihood*), the probability of observing the sample **X**, given that the hypothesis holds
  - Probability that a customer **X,** is 35 years old and has €4000 of income, given that we know he will buy a computer

- P(h) (*prior probability of h*), the initial probability
  - E.g., probability of any given customer to buy a computer, regardless of age and income

- P(**X**): marginal probability that sample data is observed
  - Probability that a person from our set of customers, has 35 years old and €4000 of income

Classification: Choosing Hypotheses

- Maximum Likelihood (maximize the likelihood)

$$h_{ML} = \arg\max_{h \epsilon H} P(D|h)$$

- Maximum a posteriori (maximize the posterior)
  - Usefull observation: it does not depend on the denominator P(D)

$$h_{MAP} = \arg\max_{h \epsilon H} P(h|D) = \arg\max_{h \epsilon H} P(D|h) \,.\, P(h)$$

Classification by Maximum A Posteriori

1. Let *D* be a training set of tuples and their associated class labels
   - Each tuple is represented by an n-Dim attribute vector **X** = $(x_1, x_2, \ldots, x_n)$
   - *n* measurements were made in the tuple from *n* attributes A_1 to A_n
2. Suppose there are *m* classes $C_1, C_2, \ldots, C_m$
3. Classification is to derive the maximum posteriori, i.e, the maximal $P(C_i|\mathbf{X})$
4. This can be derived from Bayes' theorem, as

$$P(C_i|\mathbf{X}) = \frac{P(\mathbf{X}|C_i)P(C_i)}{P(\mathbf{X})}$$

Since $P(\mathbf{X})$ is constant for all classes, then we only need to maximize:

$$P(C_i|\mathbf{X}) = P(\mathbf{X}|C_i)P(C_i)$$

Example: Covid 19 test

- A patient takes a test with two possible results (+, -), and the result comes back positive. It is known that the test returns
  - A correct positive result in 98% of the cases (TP)
  - A correct negative result in 97% of the cases (TN)
  - Only 0.008% of the population has the disease

- Q1. What is the probability that the patient is infected?
- Q2. What is the probability the patient is not infected?
- Q3. Final diagnosis?

Example: Covid 19 test

- $\mathrm{P}(\mathrm{covid})=0.008$ $\mathrm{P}(\neg\mathrm{covid})=0.992$
- $\mathrm{P}(+|\mathrm{covid})=0.98$ $\mathrm{P}(-|\mathrm{covid})=0.02$
- $\mathrm{P}(+|\neg\mathrm{covid})=0.03$ $\mathrm{P}(-|\neg\mathrm{covid})=0.97$

Using Bayes:

$$\mathrm{P}(\mathrm{covid}|+)=\mathrm{P}(+|\mathrm{covid}).\mathrm{P}(\mathrm{covid})/\mathrm{P}(+)$$
$$=0.98 \times 0.008 \div \mathrm{P}(+)=0.0784/\mathrm{P}(+)$$

$$\mathrm{P}(\neg\mathrm{covid}|+)=\mathrm{P}(+|\neg\mathrm{covid}).\mathrm{P}(\neg\mathrm{covid})/\mathrm{P}(+)$$
$$=0.03 \times 0.992 \div \mathrm{P}(+)=\mathbf{0.0298} \div P(+)$$

Another example: tests are rare!

- Patient arriving to the hospital is sneezing, can we diagnose for Covid-19 or just cold?
- Let’s assume that:
  - $\mathrm{P(sneezing)=0.3\ (\%\ of\ patients\ come\ in\ sneezing)}$
  - $\mathrm{P(Cold)=0.25}$
  - $\mathrm{P(Covid-19)=0.008}$

$$\mathrm{P(Cold|sneezing)=P(sneezing|Cold)\ x\ P(cold)\ /\ P(sneezing)}$$
$$\mathrm{P(Cold|sneezing)=1x0.25/0.3=0.83}$$
$$\mathrm{P(Covid|sneezing)=P(sneezing|Covid)\ x\ P(Covid)\ /\ P(Covid)}$$
$$\mathrm{P(Covid|sneezing)=1x0.008/0.3=0.0266}$$

Assuming
- probability that the patient sneezes because he has a cold = 100%
- probability the patient sneezes because he has covid = 100%

What if, you also measure fever from those patients?

- For two inputs, we must compute additional probabilities and conditional probabilities, including the following:
  - $\mathrm{P}(\text{Sneezing and Fever}|\text{cold})$
  - $\mathrm{P}(\text{Sneezing and}\neg\text{Fever}|\text{cold})$
  - $\mathrm{P}(\neg\text{Sneezing and Fever}|\text{cold})$
  - $\mathrm{P}(\neg\text{Sneezing and}\neg\text{Fever}|\text{cold})$
  - $\mathrm{P}(\text{Sneezing and Fever}|\text{Covid19})$
  - $\mathrm{P}(\text{Sneezing and}\neg\text{Fever}|\text{Covid19})$
  - $\mathrm{P}(\neg\text{Sneezing and Fever}|\text{Covid19})$
  - $\mathrm{P}(\neg\text{Sneezing and}\neg\text{Fever}|\text{Covid19})$

Naïve Bayesian Classifier

- Since P(X) is constant for all classes

$$P(C_i|\mathbf{X}) = P(\mathbf{X}|C_i)P(C_i)$$

- If the classes prior probabilities are not know usually it is assumed that classes are equally likely

$$P(C_1) = P(C_2) = P(C_3) = \ldots = P(C_m)$$

Naïve Bayes Classifier

- Simplified assumption: attributes are conditionally independent (i.e., no dependence relation between attributes):

$$P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i) = P(x_1|C_1)P(x_2|C_1)\ldots P(x_n|C_i)$$

- This greatly reduces the computation cost: Only counts the class distribution
- $P(C_i) = \left|C_{i,D}\right|/D$ → # of tuples of C_i in D / D
- $P(x_k|C_i)$ → # of tuples in C_i having value x_k for Ak divided by |Ci, D|

Example: play tennis or not?

Outlook	Temperature	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Overcast	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

Outlook	Temp.	Humidity	Windy	Play
Sunny	Cool	High	True	?

Weather data example

Outlook	Temp.	Humidity	Windy	Play
Sunny	Cool	High	True	?

New Evidence E

$$
\begin{aligned}
P(yes \mid E) = {} & P(Outlook = Sunny \mid yes) \\
& \times P(Temperature = Cool \mid yes) \\
& \times P(Humidity = High \mid yes) \\
& \times P(Windy = True \mid yes) \\
& \times \frac{P(yes)}{P(E)}
\end{aligned}
$$

Probability of class “yes”

Probabilities for weather data

- A new day:

Outlook			Temperature			Humidity			Windy			Play	
	Yes	*No*		*Yes*	*No*		*Yes*	*No*		*Yes*	*No*	*Yes*	*No*
Sunny	2	3	Hot	2	2	High	3	4	False	6	2	9	5
Overcast	4	0	Mild	4	2	Normal	6	1	True	3	3		
Rainy	3	2	Cool	3	1								
Sunny	2/9	3/5	Hot	2/9	2/5	High	3/9	4/5	False	6/9	2/5	9/14	5/14
Overcast	4/9	0/5	Mild	4/9	2/5	Normal	6/9	1/5	True	3/9	3/5		
Rainy	3/9	2/5	Cool	3/9	1/5								

Outlook	Temp.	Humidity	Windy	Play
Sunny	Cool	High	True	?

Likelihood of the two classes

For “yes” = 2/9 × 3/9 × 3/9 × 3/9 × 9/14 = 0.0053

For “no” = 3/5 × 1/5 × 4/5 × 3/5 × 5/14 = 0.0206

Conversion into a probability by normalization:

P(“yes”) = 0.0053 / (0.0053 + 0.0206) = 0.205

P(“no”) = 0.0206 / (0.0053 + 0.0206) = 0.795

The “zero-frequency problem”

- What if an attribute value doesn’t occur with every class value?
  (e.g. “Outlook= overcast” for class “no”)
  - Probability will be zero!
  - *A posteriori* probability will also be zero!
    (No matter how likely the other values are!)
- Remedy: add 1 to the count for every attribute value-class combination (*Laplace estimator)*
- Result: probabilities will never be zero!
  (also: stabilizes probability estimates)

Outlook	Temp.	Humidity	Windy	Play
Overcast	Cool	High	True	?

$$P(Outlook = Overcast \mid no) = 0$$

$$P(no \mid E) = 0$$

Outlook		
	Yes	*No*
Sunny	~~2~~ 3	~~3~~ 4
Overcast	~~4~~ 5	~~0~~ 1
Rainy	~~3~~ 4	~~2~~ 3
Sunny	3/12	4/8
Overcast	5/12	1/8
Rainy	4/12	3/8

Missing values

- Training: instance is not included in frequency count for attribute value-class combination
- Classification: attribute will be omitted from calculation
- Example:

Outlook	Temp.	Humidity	Windy	Play
?	Cool	High	True	?

Likelihood of “yes” = 3/9 × 3/9 × 3/9 × 9/14 = 0.0238

Likelihood of “no” = 1/5 × 4/5 × 3/5 × 5/14 = 0.0343

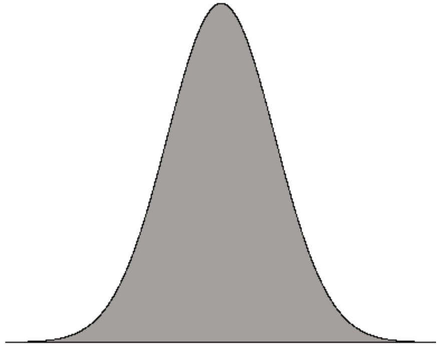
P(“yes”) = 0.0238 / (0.0238 + 0.0343) = 41%

P(“no”) = 0.0343 / (0.0238 + 0.0343) = 59%

Numeric attributes

- Usual assumption: attributes have a normal or Gaussian probability distribution (given the class)
- The probability density function for the normal distribution is defined by two parameters:
  - Sample mean μ

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$



  - Standard deviation σ

$$\sigma = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \mu)^2$$

  - Then the prob. density function f(x) is

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

Pdf = function that describes the relative likelihood for a random variable to take on a given value

Statistics for weather data

- Example density value:

Outlook			Temperature		Humidity		Windy			Play	
	Yes	No	Yes	No	Yes	No		Yes	No	Yes	No
Sunny	2	3	64, 68, 69, 70, 72, …	65, 71, 72, 80, 85, …	65, 70, 70, 75, 80, …	70, 85, 90, 91, 95, …	False	6	2	9	5
Overcast	4	0									
Rainy	3	2					True	3	3		
Sunny	2/9	3/5	μ =73	μ =75	μ =79	μ =86	False	6/9	2/5	9/14	5/14
Overcast	4/9	0/5	σ =6.2	σ =7.9	σ =10.2	σ =9.7	True	3/9	3/5		
Rainy	3/9	2/5	f=0.0340	f=0.0221	f=0.0291	f=0.0380					

$$f(temperature = 66 \mid yes) = \frac{1}{\sqrt{2\pi}6.2} e^{-\frac{(66-73)^2}{2*6.2^2}} = 0.0340$$

Classifying a new day

- A new day:

Outlook	Temp.	Humidity	Windy	Play
Sunny	66	90	true	?

Likelihood of “yes” = 2/9 × 0.0340 × 0.0221 × 3/9 × 9/14 = 0.000036

Likelihood of “no” = 3/5 × 0.0291 × 0.0380 × 3/5 × 5/14 = 0.000136

P(“yes”) = 0.000036 / (0.000036 + 0. 000136) = 20.9%

P(“no”) = 0.000136 / (0.000036 + 0. 000136) = 79.1%

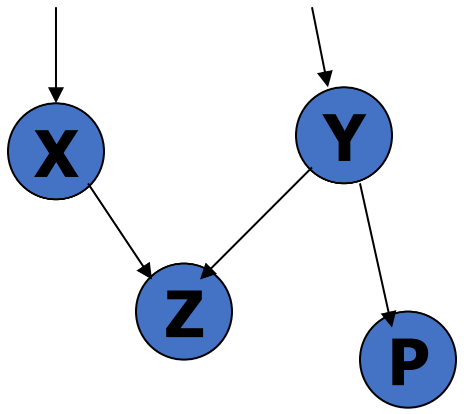
Naïve Bayes: discussion

- Advantages
  - Easy to implement
  - Good results obtained in most of the cases
- Disadvantages
  - Assumption: class conditional independence, therefore loss of accuracy
  - Practically, dependencies exist among variables
- How to deal with these dependencies? Bayesian Belief Networks

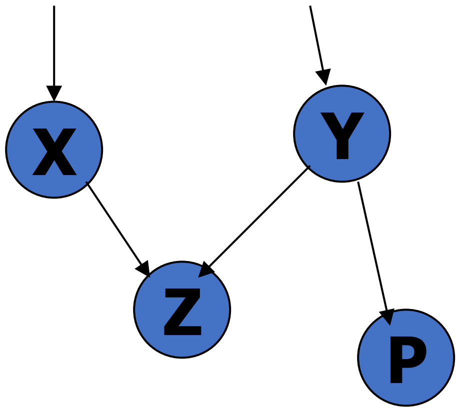
Bayesian Belief Networks

- **Bayesian belief networks** allows *class conditional independencies* between *subsets* of variables
- Composed by two components
  1. a *directed acyclic graph* (called a structure)
  2. a set of *conditional probability tables* (CPTs)

- A (*directed acyclic*) graphical model of *causal influence* relationships
  - Represents <u>dependency</u> among the variables
  - Gives a specification of joint probability distribution



Bayesian Belief Networks



- nodes represent variables,

- arcs represent the (directed) dependence among the variables.

- Node X is a parent or immediate predecessor of Node Z, and Node Z is a descendant of Node X, if there exists a directed arc from X to Z.

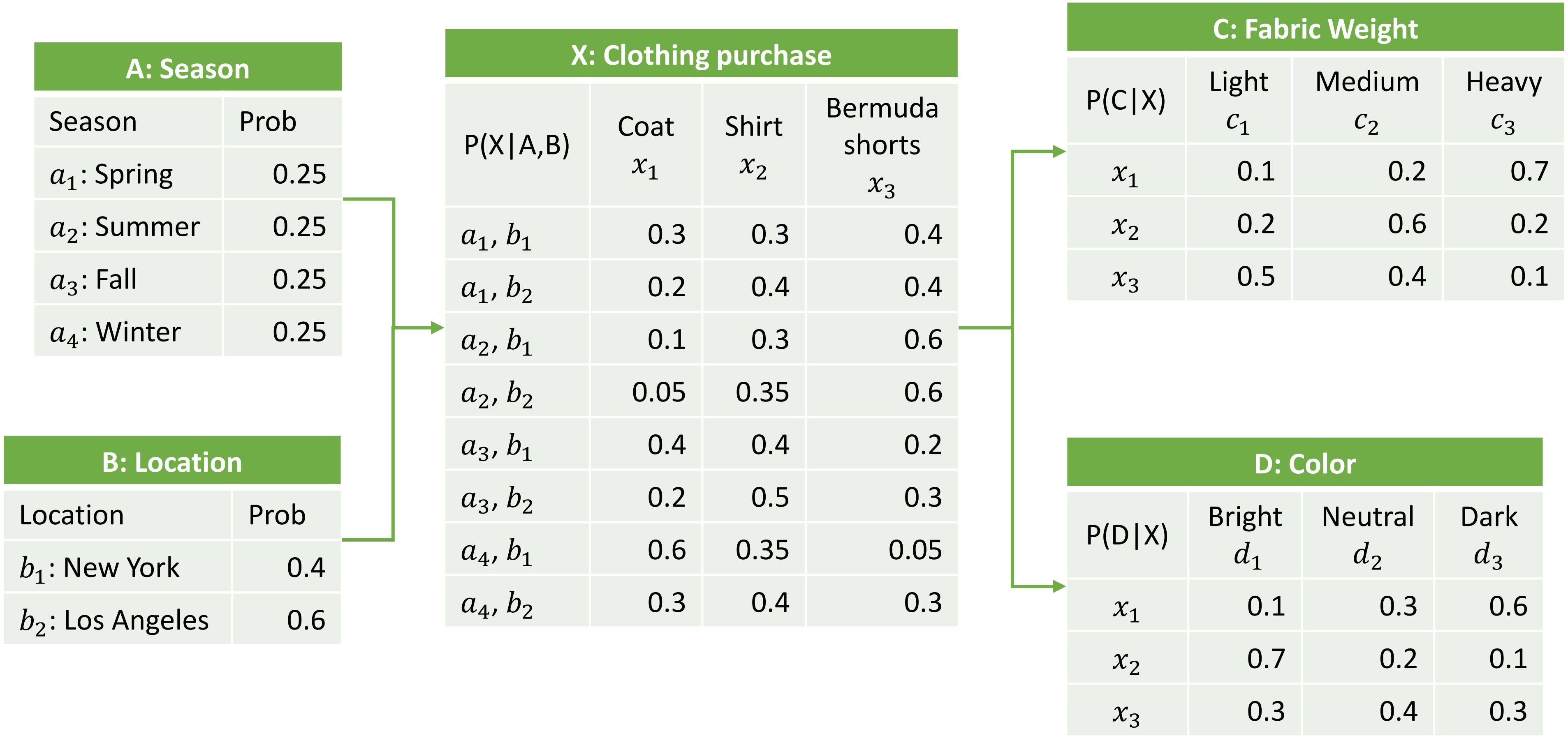
- The intrinsic relationship among the variables in a Bayesian network is as follows:

  Each variable in a Bayesian network is conditionally independent of its non-descendants in the network, given its parents.

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_m = x_m) = \prod_{i=1}^{m} P(\, X_i = x_i | parents(X_i))$$

Clothing retail example

- Suppose a retailer has 2 outlets (NY and LA)
- Producing sales in all four seasons
- Retailer is interested in probabilities related to coats, shirts and bermuda shorts



A: Season

Season	Prob
a_1: Spring	0.25
a_2: Summer	0.25
a_3: Fall	0.25
a_4: Winter	0.25

B: Location

Location	Prob
b_1: New York	0.4
b_2: Los Angeles	0.6

X: Clothing purchase

P(X|A,B)	Coat x_1	Shirt x_2	Bermuda shorts x_3
a_1, b_1	0.3	0.3	0.4
a_1, b_2	0.2	0.4	0.4
a_2, b_1	0.1	0.3	0.6
a_2, b_2	0.05	0.35	0.6
a_3, b_1	0.4	0.4	0.2
a_3, b_2	0.2	0.5	0.3
a_4, b_1	0.6	0.35	0.05
a_4, b_2	0.3	0.4	0.3

C: Fabric Weight

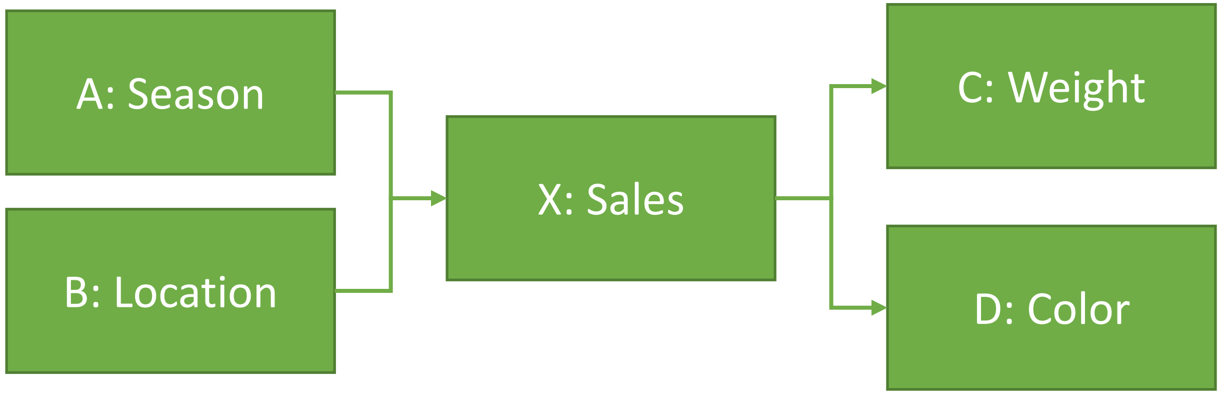
P(C|X)	Light c_1	Medium c_2	Heavy c_3
x_1	0.1	0.2	0.7
x_2	0.2	0.6	0.2
x_3	0.5	0.4	0.1

D: Color

P(D|X)	Bright d_1	Neutral d_2	Dark d_3
x_1	0.1	0.3	0.6
x_2	0.7	0.2	0.1
x_3	0.3	0.4	0.3

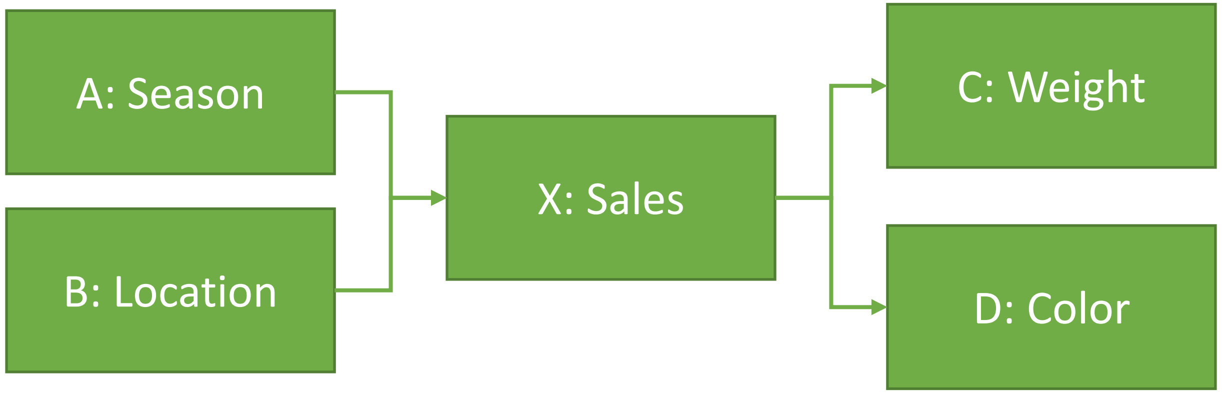
Clothing retail example

- To build a Bayesian network, there are two main considerations:
  1. What is the dependence relationship among the variables of interest?
  2. What are the associated "local" probabilities?



Clothing retail example

1. What is the dependence relationship among the variables of interest?
    1. season of the year does not depend on any of the other variables, and so we place the node for the variable *season* at the top of the Bayes network
    2. *location* does not depend on the other variables, and is therefore placed at the top of the network
    3. As the fabric weight and the color of the clothing is not known until the article is purchased, the node for the variable *clothing purchase* is inserted next into the network, with arcs to each of the *fabric weight* and *color* nodes



Clothing retail example

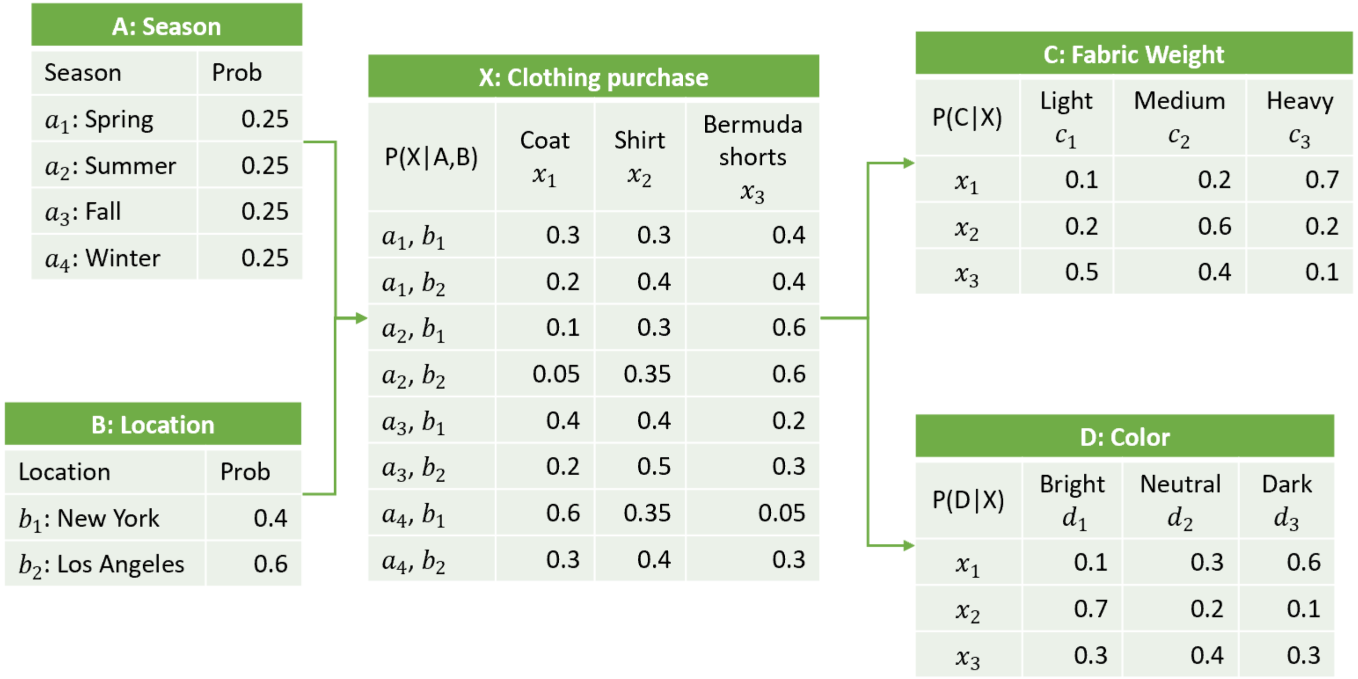
- What are the associated “local” probabilities?
  - The probabilities in the *season* node table indicate that clothing sales for this retail establishment are uniform throughout the four seasons
  - The probabilities in the *location* node probability table show that 60% of sales are generated from their Los Angeles store, and 40% from their New York store.
  - Don’t need supply conditional probabilities, because the nodes are at the top

A: Season	
Season	Prob
a_1: Spring	0.25
a_2: Summer	0.25
a_3: Fall	0.25
a_4: Winter	0.25

B: Location	
Location	Prob
b_1: New York	0.4
b_2: Los Angeles	0.6

Clothing retail example

- Probabilities of purchasing articles of clothing from the New York store in winter: probabilities of purchasing a coat, a business shirt, and Bermuda shorts are 0.60, 0.35, 0.05
- Given a particular item of clothing, the probabilities then need to be specified for *fabric weight* and *color:*
  - warm coat will have probabilities for being of light, medium, or heavy fabric or 0.10, 0.20, and 0.70
  - A business shirt will have probabilities of having bright, neutral, or dark color of 0.70, 0.20, and 0.10



A: Season

Season	Prob
a_1: Spring	0.25
a_2: Summer	0.25
a_3: Fall	0.25
a_4: Winter	0.25

B: Location

Location	Prob
b_1: New York	0.4
b_2: Los Angeles	0.6

X: Clothing purchase

P(X|A,B)	Coat x_1	Shirt x_2	Bermuda shorts x_3
a_1, b_1	0.3	0.3	0.4
a_1, b_2	0.2	0.4	0.4
a_2, b_1	0.1	0.3	0.6
a_2, b_2	0.05	0.35	0.6
a_3, b_1	0.4	0.4	0.2
a_3, b_2	0.2	0.5	0.3
a_4, b_1	0.6	0.35	0.05
a_4, b_2	0.3	0.4	0.3

C: Fabric Weight

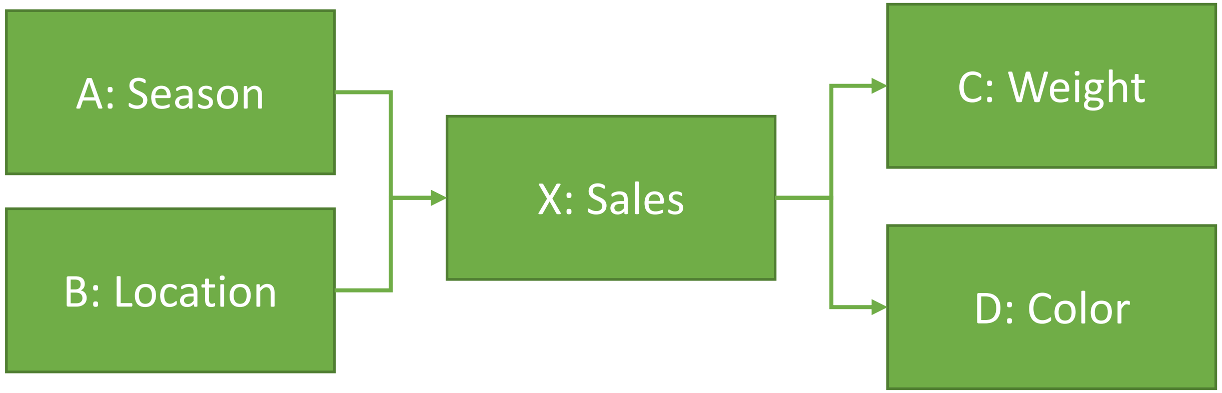
P(C|X)	Light c_1	Medium c_2	Heavy c_3
x_1	0.1	0.2	0.7
x_2	0.2	0.6	0.2
x_3	0.5	0.4	0.1

D: Color

P(D|X)	Bright d_1	Neutral d_2	Dark d_3
x_1	0.1	0.3	0.6
x_2	0.7	0.2	0.1
x_3	0.3	0.4	0.3

Relationships

- Fabric weight or color depends only on the item of clothing purchased, and not the location or season.
  - ***Color* is conditionally independent of *location***, given the article of clothing purchased.
- *Color* is conditionally independent of *season,* given clothing purchased.
- *Color* is conditionally independent of *fabric weight,* given clothing purchased.
- *Fabric weight* is conditionally independent of *color,* given clothing purchased.
- *Fabric weight* is conditionally independent of *location,* given clothing purchased.
- *Fabric weight* is conditionally independent of *season,* given clothing purchased.
- *Season* is conditionally independent of *location,* given its parents, **but** as *season* has no parents in the Bayes net, this means that ***season* and *location* are (unconditionally) independent**.



Using the bayesian network to find probabilities

- Find the probability that a given purchase involved **light-fabric**, **neutral-colored Bermuda shorts** was made in **New York** in the **winter**.

Using:

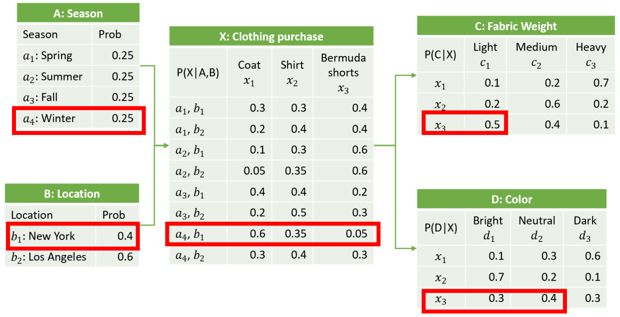
$$P(X_1 = x_1, X_2 = x_2, \ldots, X_m = x_m) = \prod_{i=1}^{m} P(\, X_i = x_i | parents(X_i))$$

$$P(A = a_4, B = b_1, C = c_1, D = d_2, X = x_3)$$

$$\boxed{= P(A = a_4)}\boxed{. P(B = b_1)}\boxed{. P(X = x_3 | A = a_4 \cap B = b_1)}\boxed{. P(C = c_1 | X = x_3)}\boxed{. P(D = d_2 | X = x_3)}$$

$$= \mathrm{P}(seas = wint)$$
$$\times P(Loc \ = NY)$$
$$\times \mathrm{P}(Cloth = shorts | seas = Wint\ and\ Loc = NY)$$
$$\times P(fab = light | cloth = shorts)$$
$$\times P(color = neutr | cloth = shorts)$$

$$= 0.25 \times 0.4 \times 0.05 \times 0.5 \times 0.4 = 0.001$$



A: Season

Season	Prob
a_1: Spring	0.25
a_2: Summer	0.25
a_3: Fall	0.25
a_4: Winter	0.25

B: Location

Location	Prob
b_1: New York	0.4
b_2: Los Angeles	0.6

X: Clothing purchase

P(X|A,B)	Coat x_1	Shirt x_2	Bermuda shorts x_3
a_1, b_1	0.3	0.3	0.4
a_1, b_2	0.2	0.4	0.4
a_2, b_1	0.1	0.3	0.6
a_2, b_2	0.05	0.35	0.6
a_3, b_1	0.4	0.4	0.2
a_3, b_2	0.2	0.5	0.3
a_4, b_1	0.6	0.35	0.05
a_4, b_2	0.3	0.4	0.3

C: Fabric Weight

P(C|X)	Light c_1	Medium c_2	Heavy c_3
x_1	0.1	0.2	0.7
x_2	0.2	0.6	0.2
x_3	0.5	0.4	0.1

D: Color

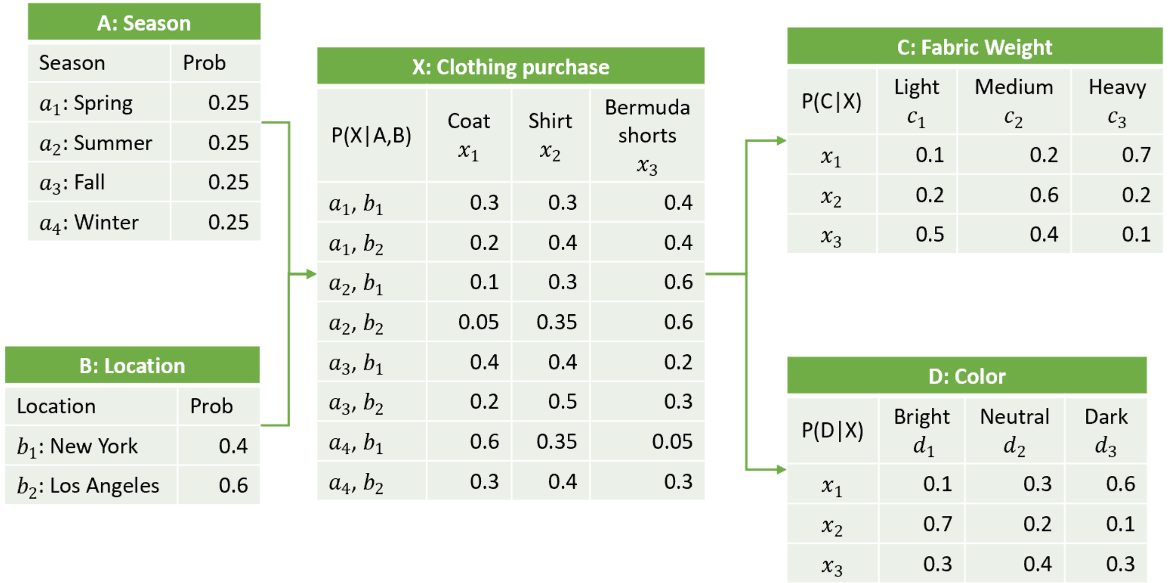
P(D|X)	Bright d_1	Neutral d_2	Dark d_3
x_1	0.1	0.3	0.6
x_2	0.7	0.2	0.1
x_3	0.3	0.4	0.3

Using the bayesian network to find probabilities

- prior probability of

a coat is as follows:

$$p(coat) = p(X = x1)$$
$$= p(X = x1 | A = a1 \cap B = b1) \cdot p(A = a1 \cap B = b1)$$
$$+ p(X = x1 | A = a1 \cap B = b2) \cdot p(A = a1 \cap B = b2)$$
$$+ p(X = x1 | A = a2 \cap B = b1) \cdot p(A = a2 \cap B = b1)$$
$$+ p(X = x1 | A = a2 \cap B = b2) \cdot p(A = a2 \cap B = b2)$$
$$+ p(X = x1 | A = a3 \cap B = b1) \cdot p(A = a3 \cap B = b1)$$
$$+ p(X = x1 | A = a3 \cap B = b2) \cdot p(A = a3 \cap B = b2)$$
$$+ p(X = x1 | A = a4 \cap B = b1) \cdot p(A = a4 \cap B = b1)$$
$$+ p(X = x1 | A = a4 \cap B = b2) \cdot p(A = a4 \cap B = b2)$$
$$= (0.30) \cdot (0.25*0.4) + (0.20) \cdot (0.15) + (0.10) \cdot (0.10) + (0.05) \cdot$$
$$(0.15)$$
$$+ (0.40) \cdot (0.10) + (0.20) \cdot (0.15) + (0.60) \cdot (0.10) + (0.30) \cdot (0.15)$$
$$= 0.2525$$



A: Season

Season	Prob
a_1: Spring	0.25
a_2: Summer	0.25
a_3: Fall	0.25
a_4: Winter	0.25

B: Location

Location	Prob
b_1: New York	0.4
b_2: Los Angeles	0.6

X: Clothing purchase

P(X|A,B)	Coat x_1	Shirt x_2	Bermuda shorts x_3
a_1, b_1	0.3	0.3	0.4
a_1, b_2	0.2	0.4	0.4
a_2, b_1	0.1	0.3	0.6
a_2, b_2	0.05	0.35	0.6
a_3, b_1	0.4	0.4	0.2
a_3, b_2	0.2	0.5	0.3
a_4, b_1	0.6	0.35	0.05
a_4, b_2	0.3	0.4	0.3

C: Fabric Weight

P(C|X)	Light c_1	Medium c_2	Heavy c_3
x_1	0.1	0.2	0.7
x_2	0.2	0.6	0.2
x_3	0.5	0.4	0.1

D: Color

P(D|X)	Bright d_1	Neutral d_2	Dark d_3
x_1	0.1	0.3	0.6
x_2	0.7	0.2	0.1
x_3	0.3	0.4	0.3

Note: *Season* and *location* are independent, so that $p(A \cap B) = p(A) \cdot p(B)$

How Are Bayesian Networks Constructed?

- **Subjective construction**: Identification of (direct) causal structure
  - People are quite good at identifying direct causes from a given set of variables & whether the set contains all relevant direct causes
  - Markovian assumption: Each variable becomes independent of its non-effects once its direct causes are known
  - HMM (Hidden Markov Model): often used to model dynamic systems whose states are not observable, yet their outputs are
- **Synthesis from other specifications**
  - E.g., from a formal system design: block diagrams & info flow
- **Learning from data**
  - E.g., from medical records or student admission record
  - Learn parameters give its structure or learn both structure and parms
  - Maximum likelihood principle: favors Bayesian networks that maximize the probability of observing the given data set

NOVA IMS INFORMATION MANAGEMENT SCHOOL