FUDAN SOE SUMMER SCHOOL INTRODUCTION TO AI

Probabilistic Reasoning and Decision Making

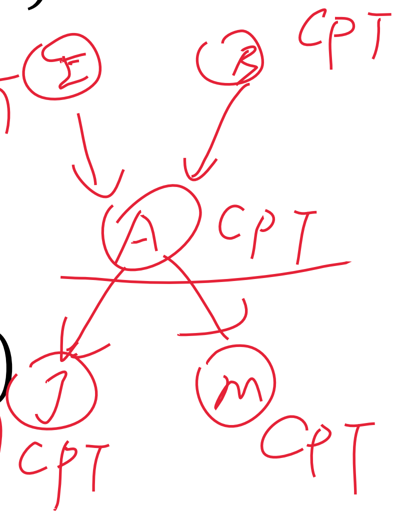
Day 8 – Maximum Likelihood Learning

Estimating Parameters of a Bayes Net

<u>Parameters to estimate (CPTs of the network):</u> $P(X_i = x | Pa(X_i) = \pi)$

<u>Data set</u>: $\{(x_1{}^{(t)}, x_2{}^{(t)}, \dots, x_n{}^{(t)})\}_{t=1}^{T}$

flav $\quad p = \dfrac{\#\text{ of cherries}}{T}$



$$\mathcal{L} = \log P(data) \overset{iid}{=} \log \prod_{t=1}^{T} P\left(X_1 = x_1{}^{(t)}, X_2 = x_2{}^{(t)}, \dots, X_n = x_n{}^{(t)}\right)$$

$$= \sum_{t=1}^{T} \log\left(P\left(X_1 = x_1^{(t)}, X_2 = x_2^{(t)}, \dots, X_n = x_n^{(t)}\right)\right)$$

$$P\left(X_1 = x_1{}^{(t)}, X_2 = x_2{}^{(t)}, \dots, X_n = x_n{}^{(t)}\right) = \prod_{i=1}^{n} P\left(X_i = x_i^{(t)} \mid Pa(X_i) = Pa(x_i)^{(t)}\right)$$

$$L = \sum_{t=1}^{T} \sum_{i=1}^{n} \log P\left(X_i = x_i^{(t)} \mid Pa(X_i) = Pa(x_i)^{(t)}\right)$$

sum over all samples

sum over all variables

E, B	P(A=1 | E, B)
0 1	—
1 0	—
0 0	—
1 1	—

Estimating Parameters of a Bayes Net

<u>Parameters to estimate (CPTs of the network):</u> $P(X_i = x|Pa(X_i) = \pi)$

<u>Data set</u>: $\{(x_1{}^{(t)}, x_2{}^{(t)}, \dots, x_n{}^{(t)})\}_{t=1}^{T}$

Log-likelihood $$\mathcal{L} = \sum_{i=1}^{n} \boxed{\sum_{t=1}^{T} \log P\left(x_i{}^{(t)} | pa_i{}^{(t)}\right)}$$

Where do we go from here?

Reminder: Single node Bayes Net example:

$$\mathcal{L} = \sum_{t=1}^{T} \log P\left(X = x^{(t)}\right)$$

How did we simplify this expression?

p $\quad$ $1-p$

n_c $\quad$ $T \quad N_j$

$$\frac{dL}{dp} \stackrel{?}{=} 0, \qquad p = \frac{N_c}{N_c + N_j} = \frac{N_c}{T}$$

Estimating Parameters of a Bayes Net

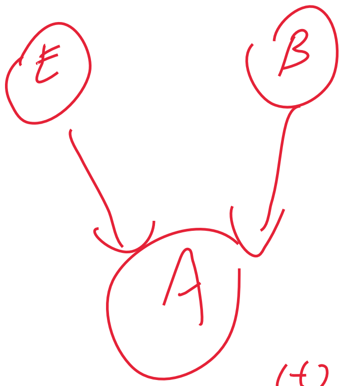
Parameters to estimate (CPTs of the network): $P(X_i = x | Pa(X_i) = \pi)$

$P(A=0 | E, B= \sim)$

Data set: $\{(x_1{}^{(t)}, x_2{}^{(t)}, \dots, x_n{}^{(t)})\}_{t=1}^{T}$

Log-likelihood $\mathcal{L} = \sum_{i=1}^{n} \sum_{t=1}^{T} \log P(x_i{}^{(t)} | pa_i{}^{(t)})$

Where do we go from here? (example)



data

E	B	A
0	0	0
0	0	1
1	0	0
1	0	0
1	1	0
⋮	⋮	⋮

T rows

define: $\text{Count}(X_i = x_i^{(t)}, Pa(X_i) = Pa_i^{(t)})$

$\text{Count}(A=0, Pa(A)=00) = 1$ $\quad$ $\text{Count}(A=1, Pa(A)=11) = 0$

$\text{Count}(A=0, Pa(A)=11) = 1$

Let $\text{count}(X_i = x_i^{(t)}, Pa(X_i) = Pa_i^{(t)})$ be # of samples where X_i is $x_i^{(t)}$, and parents of X_i is $Pa_i^{(t)}$ (π)

$$L = \sum_{i=1}^{n} \boxed{\sum_{t=1}^{T} \underbrace{\log P(X_i = x_i^{(t)} \mid Pa(X_i) = Pa_i^{(t)})}_{\text{finite \# of values}}}$$

$$= \underbrace{\sum_{i=1}^{n}}_{\text{sum over variables}} \underbrace{\sum_{x}}_{\text{over the values of } X_i} \underbrace{\sum_{\pi}}_{\text{over the value of } Pa(X_i)} \text{count}(X_i = x, Pa(X_i) = \pi) \log P(X_i = x \mid Pa(X_i) = \pi)$$

$T=6, \quad n=3$

$$\begin{array}{ccc} E & & B \\ & \searrow \quad \swarrow & \\ & A & \end{array} \quad \text{CPT}$$

t	E	B	A
1	0	0	0
2	0	0	0
3	0	1	1
4	0	1	0
5	1	1	1
6	1	1	1

$$L = \sum_{i=1}^{3} \sum_{t=1}^{6} \log P(X_i = x_i^{(t)} \mid Pa(X_i) = \pi_i)$$

$$= \sum_{t=1}^{6} \log P(E = e^{(t)}) + \sum_{t=1}^{6} \log P(B = b^{(t)})$$

$$+ \sum_{t=1}^{6} \log P(A = a^{(t)} \mid \underbrace{E, B}_{Pa(A)} = \pi_i)$$

$$\log P(A=0 \mid EB=00) + \log P(A=0 \mid EB=00) + \log P(A=1 \mid E,B=01)$$

$$+ \log P(A=0 \mid EB=01) + \log P(A=1 \mid EB=11) + \log P(A=1 \mid EB=11)$$

$$2 \log P(A=0 \mid EB=00) + \overbrace{\boxed{2}}^{\text{count}(A=1,\ EB=11)} \log P(A=1 \mid EB=11)$$

$$+ 1 \cdot \log P(A=1 \mid EB=01) + \log P(A=0 \mid EB=01)$$

Estimating Parameters of a Bayes Net

Parameters to estimate (CPTs of the network): $P(X_i = x | Pa(X_i) = \pi)$

Data set: $\{(x_1{}^{(t)}, x_2{}^{(t)}, \dots, x_n{}^{(t)})\}_{t=1}^{T}$

Log-likelihood

Explain this in your own words

$$\mathcal{L} = \sum_{i=1}^{n} \sum_{t=1}^{T} \log P(x_i{}^{(t)} | pa_i{}^{(t)})$$

$$= \sum_{i=1}^{n} \sum_{x} \sum_{\pi} \underbrace{\text{count}(X_i = x, pa_i = \pi)}_{\text{constant}} \log P(X_i = x | pa_i = \pi)$$

$$\frac{\partial L}{\partial P(X_i = x | Pa_i = \pi)} \stackrel{?}{=} 0$$

Can take the derivative to maximize. Which term are we optimizing over?

A. $\text{count}(X_i = x, pa_i = \pi)$
B. $P(X_i = x | pa_i = \pi)$
C. $X_i = x$
D. $pa_i = \pi$
E. t

ML Parameters for Bayes Nets:

if w/o parents

$$P_{ML}(X_i = x) = \frac{\text{Count}(X_i = x)}{T}$$

<u>Parameters to estimate (CPTs of the network):</u> $P(X_i = x | Pa(X_i) = \pi)$

if w/ parents

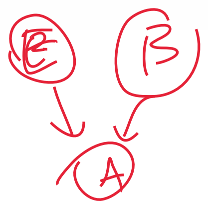
$$P(X_i = x | Pa(X_i) = \pi) = \frac{\text{Count}(X_i = x, Pa(X_i) = \pi)}{\text{Count}(Pa(X_i) = \pi)}$$

$$P_{ML}(X_i = x | pa_i = \pi) = \frac{\text{count}(X_i = x, pa_i = \pi)}{\sum_{x'} \text{count}(X_i = x', pa_i = \pi)}$$

all possible values for X_i

isolate a pattern for parents, find the ratio of a particular value against all possible values for X_i

t	E	B	A
1	0	0	0
2	0	0	0
3	0	1	1
4	0	1	0
5	1	1	1
6	1	1	1



Pa(A)

$$P(A=1 | EB = 00) = \frac{0}{2} = 0$$

$$P(A=1 | EB = 01) = \frac{1}{1+1} = \frac{1}{2} = .5$$

$$P(A=1 | EB = 11)$$

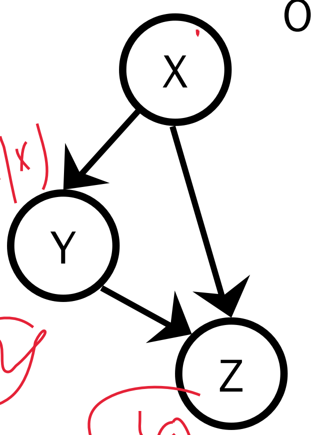
A: 0.5
B: 0
C: 1

ML Parameters for Bayes Net: Example

X | P(X)
1 | — (1)

T: 12 h: 3

$$P_{ML}(X_i = x | pa_i = \pi) = \frac{\text{count}(X_i = x, pa_i = \pi)}{\sum_{x'} \text{count}(X_i = x', pa_i = \pi)}$$



x | P(Y=1|x)
0 |
1 |
(2)

(4)

X, Y and Z are Boolean variables

X Y | P(Z=1|XY)
0 0 |
0 1 |
1 0 |
1 1 |

Observed data:

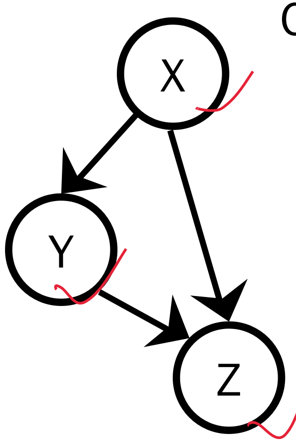
X	Y	Z
0	0	1
0	1	0
0	1	1
0	1	0
1	0	0
1	0	0
0	1	1
1	0	0
0	1	1
0	0	1
0	1	1
1	0	0

Which of the following is a parameter we would like to estimate?

A. P(X=1)
B. P(Y=1)
C. P(X=1|Y=1)
D. More than one of these
E. None of these

ML Parameters for Bayes Net: Example

$$P_{ML}(X_i = x | pa_i = \pi) = \frac{\text{count}(X_i = x, pa_i = \pi)}{\sum_{x'} \text{count}(X_i = x', pa_i = \pi)}$$



X, Y and Z are Boolean variables

Observed data:

X	Y	Z
0	0	1
0	1	0
0	1	1
0	1	0
1	0	0
1	0	0
0	1	1
1	0	0
0	1	1
0	0	1
0	1	1
1	0	0

Not including complements (e.g. P(X=1) and P(X=0)), how many different parameters are there to estimate?

A. 3
B. 4
C. 5
D. 7
E. more than 7

$P(X=1)\checkmark = \frac{\text{count}(X=1)}{T} = \frac{4}{12} = \frac{1}{3}$

$0 = \frac{0}{4} = P(Y=1|X=1)$ $\quad P(Y=1|X=0) = \frac{6}{8} = \frac{3}{4}$

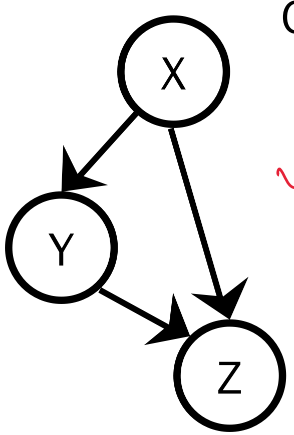
$1 = P(Z=1|X=0, Y=0)$, $P(Z=1|X=0, Y=1)$

$P(Z=1|X=1, Y=0)$, $\underline{P(Z=1|X=1, Y=1)}$ undefined

$= \frac{0}{\sim} = 0$ $\quad \frac{4}{6} = \frac{2}{3}$

ML Parameters for Bayes Net: Example

$$P_{ML}(X_i = x | pa_i = \pi) = \frac{\text{count}(X_i = x, pa_i = \pi)}{\sum_{x'} \text{count}(X_i = x', pa_i = \pi)}$$



X, Y and Z are Boolean variables

Observed data: T: 12

X	Y	Z
0	0	1
0	1	0
0	1	1
0	1	0
1	0	0
1	0	0
0	1	1
1	0	0
0	1	1
0	0	1
0	1	1
1	0	0

What is the ML estimate for P(Z=1|X=0, Y=0)?

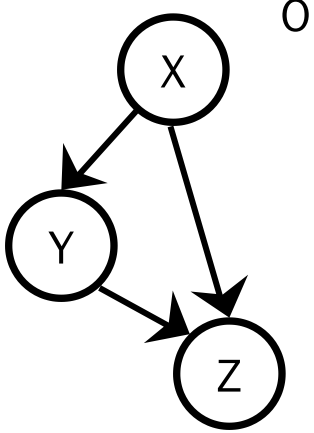
A. 0
B. 1/6
C. 1/2
D. 1
E. None of the above

$$= \frac{\text{Count}(Z=1, X=0, Y=0)}{\text{Count}(X=0, Y=0)} = \frac{2}{2}$$

$$= 1$$

ML Parameters for Bayes Net: Example

$$P_{ML}(X_i = x|pa_i = \pi) = \frac{\text{count}(X_i = x, pa_i = \pi)}{\sum_{x'} \text{count}(X_i = x', pa_i = \pi)}$$



X, Y and Z are Boolean variables

Observed data:

X	Y	Z
0	0	1
0	1	0
0	1	1
0	1	0
1	0	0
1	0	0
0	1	1
1	0	0
0	1	1
0	0	1
0	1	1
1	0	0

Which parameter has an undefined ML estimate?

A. P(X=1)
B. P(Y=1|X=0)
C. P(Z=1|X=0, Y=0)
D. P(Z=1|X=1, Y=1) $= \dfrac{\text{count}(Z=1, X=1, Y=1)}{\underbrace{\text{count}(X=1, Y=1)}_{0}}$ undefined
E. More than one of the above

Summary: Estimating Parameters of a Bayes Net

Parameters to estimate (CPTs of the network): $P(X_i = x|Pa(X_i) = \pi)$

Data set: $\{(x_1{}^{(t)}, x_2{}^{(t)}, \dots, x_n{}^{(t)})\}_{t=1}^{T}$

Log-likelihood

Explain this in your own words

$$\mathcal{L} = \sum_{i=1}^{n} \sum_{t=1}^{T} \log P\left(x_i{}^{(t)}|pa_i{}^{(t)}\right)$$

$$= \sum_{i=1}^{n} \sum_{x} \sum_{\pi} \text{count}(X_i = x, pa_i = \pi) \log P(X_i = x|pa_i = \pi)$$

Review: ML Parameters for Bayes Nets

<u>Parameters to estimate (CPTs of the network):</u> $P(X_i = x | Pa(X_i) = \pi)$

$$P_{ML}(X_i = x | pa_i = \pi) = \frac{\text{count}(X_i = x, pa_i = \pi)}{\sum_{x'} \text{count}(X_i = x', pa_i = \pi)}$$

$$= \frac{\text{count}(X_i = x, Pa_i = \pi)}{\text{count}(Pa_i = \pi)}$$ w/ parents

$$P_{ML}(X_i = x) = \frac{\text{count}(X_i = x)}{T}$$ w/o parents

Using the Indicator function I

$$I(x,y) = \begin{cases} 1 & \text{if } x == y \\ 0 & \text{otherwise} \end{cases}$$

- Which of the following correctly expresses the ML estimate $P_{ML}(X_i = x)$? $= \frac{\text{Count}(X_i = x)}{T} = \frac{1}{T}\underbrace{\left(\sum_{t=1}^{T} I(x_i^{(t)}, x)\right)}_{\text{Count}(X_i = x)}$

A. $I(x_i, x) \times T$

B. $\sum_{i=1}^{n} I(x_i, x)$

C. $\frac{1}{T}\sum_{t=1}^{T} I\left(x_i{}^{(t)}, x\right)$

D. $\frac{1}{T}\sum_{t=1}^{T} I\left(x_i{}^{(t)}, x\right) P\left(X_i = x_i{}^{(t)}\right)$

E. None of these

Suppose we have a belief network with nodes $X_1, \ldots, X_n$, and let $\mathrm{Pa}(X_i)$ denote the parents of node X_i.

A fully-observed dataset for this model can be written as

$$\text{data} = \{(x_1^{(t)}, x_2^{(t)}, \ldots, x_n^{(t)})\}_{t=1}^{T}$$

The log-likelihood of this model is $\mathcal{L} = \log P(\text{data})$. Show that the log-likelihood can be written as

$$\mathcal{L} = \sum_{i=1}^{n} \sum_{x} \sum_{\pi} \text{count}(X_i = x, \mathrm{Pa}(X_i) = \pi) \log P(X_i = x | \mathrm{Pa}(X_i) = \pi)$$

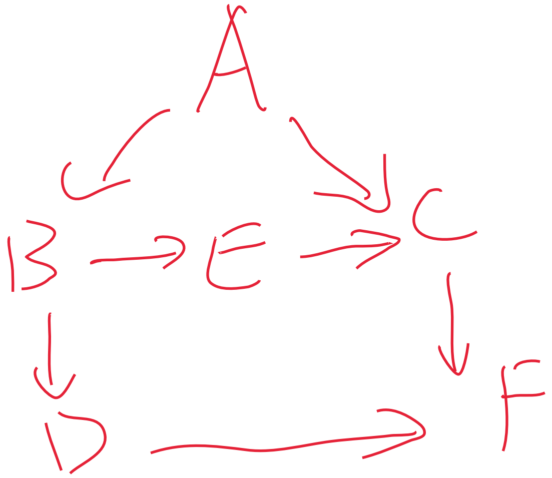
Let the graph be as follows:
Nodes: A, B, C, D, E, F
Edges: $A \to B$, $A \to C$, $B \to D$, $B \to E$, $E \to C$, $C \to F$, $D \to F$

Consider a complete data set of T i.i.d. examples: $\{(a_t, b_t, c_t, d_t, e_t, f_t)\}_{t=1}^{T}$. Compute the maximum likelihood estimates of the conditional probability tables (CPTs) shown below for this data set. Complete these CPT estimates in terms of the indication function denoted by $I(a, a_t) = 1$ if $a = a_t$ and 0 otherwise.

(a) $P(A = a)$

$$= \frac{1}{T} \sum_{t=1}^{T} \underbrace{I(a_t, a)}$$

↳ for each sample, is $a_t == a$??



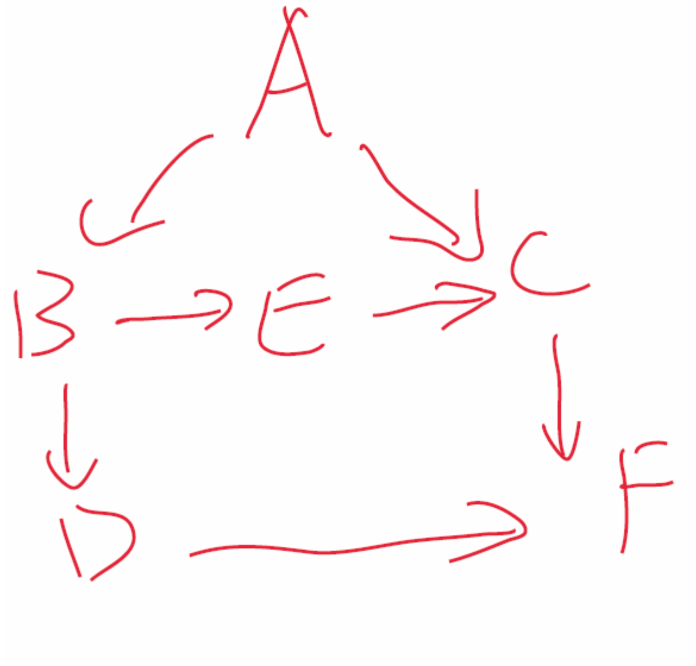
Let the graph be as follows:
Nodes: A, B, C, D, E, F
Edges: $A \to B$, $A \to C$, $B \to D$, $B \to E$, $E \to C$, $C \to F$, $D \to F$

Consider a complete data set of T i.i.d. examples: $\{(a_t, b_t, c_t, d_t, e_t, f_t)\}_{t=1}^{T}$. Compute the maximum likelihood estimates of the conditional probability tables (CPTs) shown below for this data set. Complete these CPT estimates in terms of the indication function denoted by $I(a, a_t) = 1$ if $a = a_t$ and 0 otherwise.

(b) $P(B = b|A = a)$

$$= \frac{\text{count}(A=a, B=b)}{\text{count}(A=a)}$$

$$= \frac{\sum_{t=1}^{T} I(a_t, a) \cdot I(b_t = b)}{\sum_{t=1}^{T} I(a_t, a)}$$

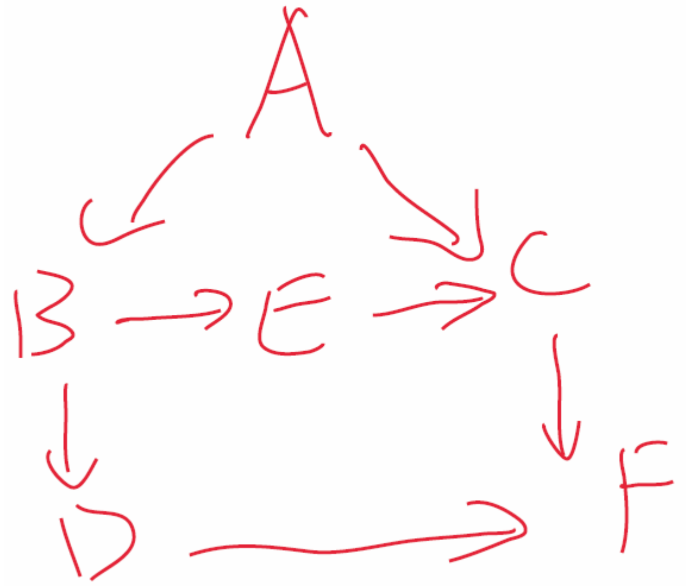


Let the graph be as follows:
Nodes: A, B, C, D, E, F
Edges: $A \to B$, $A \to C$, $B \to D$, $B \to E$, $E \to C$, $C \to F$, $D \to F$

Consider a complete data set of T i.i.d. examples: $\{(a_t, b_t, c_t, d_t, e_t, f_t)\}_{t=1}^{T}$. Compute the maximum likelihood estimates of the conditional probability tables (CPTs) shown below for this data set. Complete these CPT estimates in terms of the indication function denoted by $I(a, a_t) = 1$ if $a = a_t$ and 0 otherwise.

(c) $P(C = c | A = a, E = e)$

$$= \frac{\sum_{t=1}^{T} I(a_t, a)\, I(e_t, e)\, I(c_t, c)}{\sum_{t=1}^{T} I(a_t, a)\, I(e_t, e)}$$

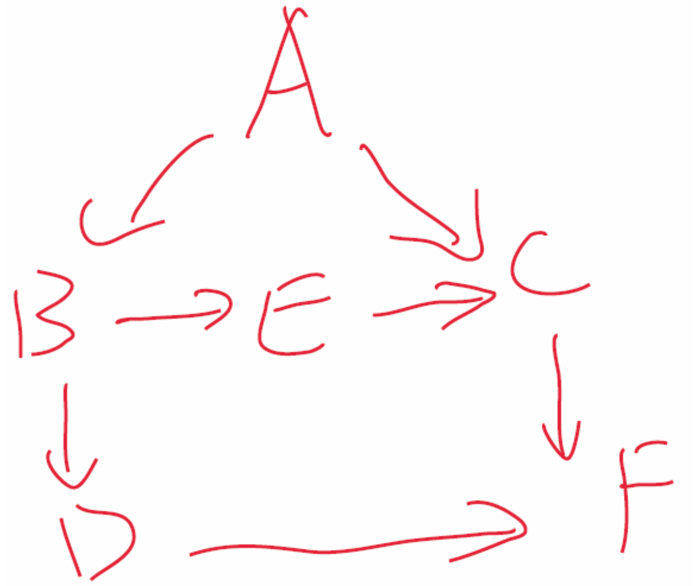


Let the graph be as follows:
Nodes: A, B, C, D, E, F
Edges: $A \to B$, $A \to C$, $B \to D$, $B \to E$, $E \to C$, $C \to F$, $D \to F$

Consider a complete data set of T i.i.d. examples: $\{(a_t, b_t, c_t, d_t, e_t, f_t)\}_{t=1}^{T}$. Compute the maximum likelihood estimates of the conditional probability tables (CPTs) shown below for this data set. Complete these CPT estimates in terms of the indication function denoted by $I(a, a_t) = 1$ if $a = a_t$ and 0 otherwise.

(d) $P(D = d | B = b)$

$$= \frac{\sum_t I(d_t, d)\, I(b_t, b)}{\sum_t I(b_t, b)}$$

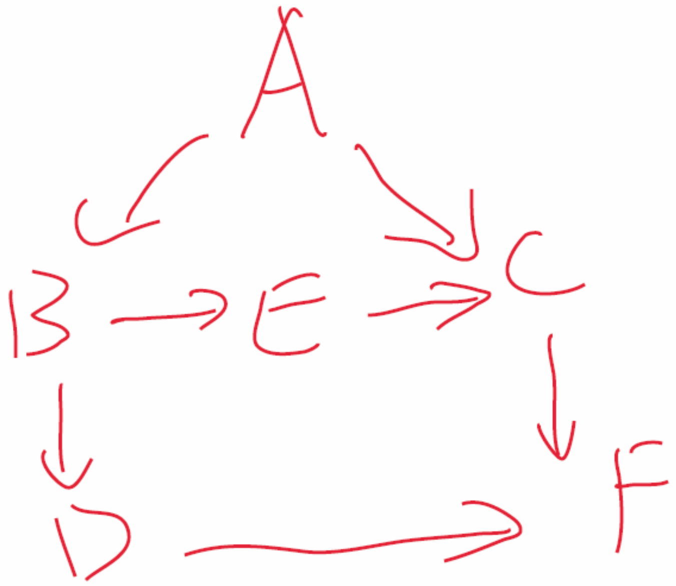


Let the graph be as follows:
Nodes: A, B, C, D, E, F
Edges: $A \to B$, $A \to C$, $B \to D$, $B \to E$, $E \to C$, $C \to F$, $D \to F$

Consider a complete data set of T i.i.d. examples: $\{(a_t, b_t, c_t, d_t, e_t, f_t)\}_{t=1}^{T}$. Compute the maximum likelihood estimates of the conditional probability tables (CPTs) shown below for this data set. Complete these CPT estimates in terms of the indication function denoted by $I(a, a_t) = 1$ if $a = a_t$ and 0 otherwise.

(e) $P(E = e|B = b)$

$$= \frac{\sum_t I(b_t, b) I(e_t, e)}{\sum_t I(b_t, b)}$$



Let the graph be as follows:
Nodes: A, B, C, D, E, F
Edges: $A \to B$, $A \to C$, $B \to D$, $B \to E$, $E \to C$, $C \to F$, $D \to F$

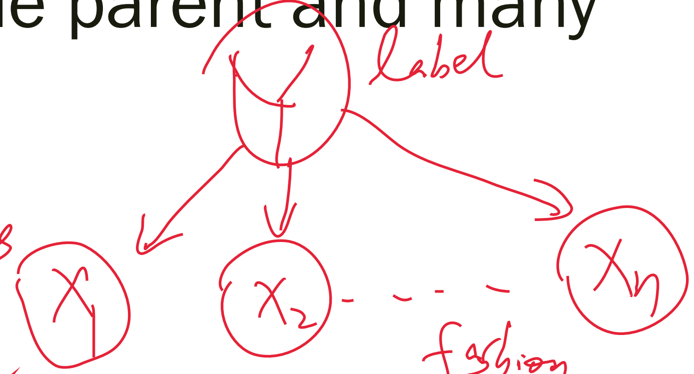
Consider a complete data set of T i.i.d. examples: $\{(a_t, b_t, c_t, d_t, e_t, f_t)\}_{t=1}^{T}$. Compute the maximum likelihood estimates of the conditional probability tables (CPTs) shown below for this data set. Complete these CPT estimates in terms of the indication function denoted by $I(a, a_t) = 1$ if $a = a_t$ and 0 otherwise.

(f) $P(F = f | C = c, D = d)$

$$= \frac{\sum_t I(f_t, f)\, I(c_t, c)\, I(d_t, d)}{\sum_t I(c_t, c)\, I(d_t, d)}$$

Naïve Bayes and Markov Models: Two kinds of Bayes Nets

- A Naïve Bayes model is a Bayes net with a single parent and many children.
- Example: Document Classification



sport

$Y \in \{1, 2, 3, \ldots, k\}$ Where 1=sports, 2=fashion, 3=politics, ..., etc.

politics fashion

$X_i \in \{0, 1\}$ for i=1…n, where 0 means ith word in dictionary does not appear in document and 1 means it does

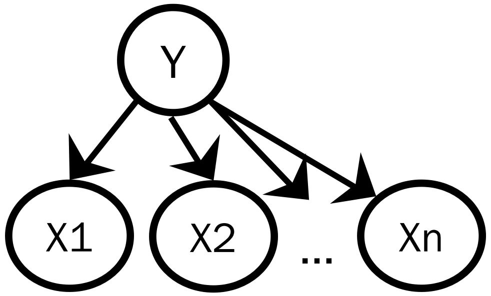
new vector

label ?

X_1 X_2 X_n

vector

Naïve Bayes: Learning and Classification



Learning:

$$P(Y = y) = \frac{\sum_t 1(y_t, y)}{T}$$

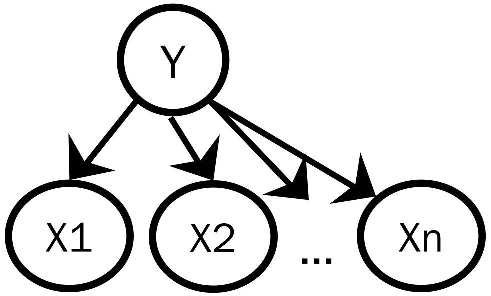
$$P(X_i = 1|Y = y) = \frac{\sum_t 1(y_t, y)\, 1(x_t, 1)}{\sum_t 1(y_t, y)}$$

Classification

$$P(Y = y \mid X_1 = x_1^{(new)}, X_2 = x_2^{(new)}, \ldots, X_n = x_n^{(new)})$$

$$= \frac{P(Y = y, X_1 = x_1^{(n)}, X_2 = x_2^{(n)}, \ldots, X_n = x_n^{(new)})}{\sum_{y'} P(Y = y', X_1 = x_1^{(n)}, X_2 = x_2^{(n)}, \ldots, X_n = x^{(n)})}$$

Naïve Bayes: Learning and Classification



Learning:

$P(Y = y)$: Proportion of documents labeled as category y

$P(X_i = 1|Y = y)$: Proportion of category y documents where word X_i occurs.

Classification

$$\frac{P(Y = y)\prod_{i=1}^{n} P(X_i = x_i|Y = y)}{\sum_{y'} P(Y = y')\prod_{i=1}^{n} P(X_i = x_i|Y = y')}$$

y politics : .5

fashion : .2

sports : .1

:

Strengths and weaknesses?

label : best probability

Markov Models: Sequential models where each element depends on only elements before it

Example: Language

Let W_i denote the i^{th} word in a sentence. $P(w_1, w_2, w_3, \dots, w_L) = \prod_{i=1}^{n} P(w_i | w_1, \dots, w_{i-1})$

<u>Two simplifying assumptions:</u>

1. Finite Context

2. Position Invariance