FUDAN SOE SUMMER SCHOOL INTRODUCTION TO AI

Probabilistic Reasoning and Decision Making

Day 8 – Maximum Likelihood Learning

Estimating Parameters of a Bayes Net

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

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

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

$$P\left(X_1 = x_1{}^{(t)}, X_2 = x_2{}^{(t)}, \ldots, X_n = x_n{}^{(t)}\right) =$$

Estimating Parameters of a Bayes Net

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

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

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

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?

Estimating Parameters of a Bayes Net

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

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

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

Where do we go from here?
(example)

Estimating Parameters of a Bayes Net

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

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

Log-likelihood

Explain this in your own words

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

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

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

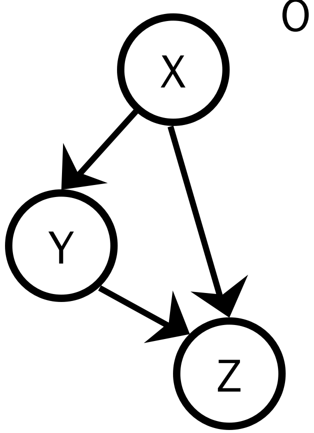
ML Parameters for Bayes Nets:

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

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

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:

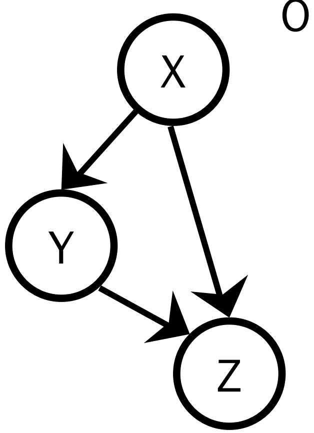
X	Y	Z
0	0	1
0	1	0
0	1	1
0	1	0
1	0	0
1	0	0
0	1	1
1	0	0
0	1	1
0	0	1
0	1	1
1	0	0

Which 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:

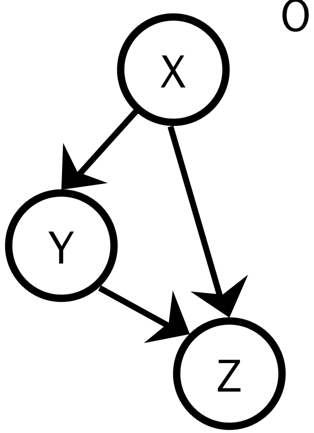
X	Y	Z
0	0	1
0	1	0
0	1	1
0	1	0
1	0	0
1	0	0
0	1	1
1	0	0
0	1	1
0	0	1
0	1	1
1	0	0

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

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:

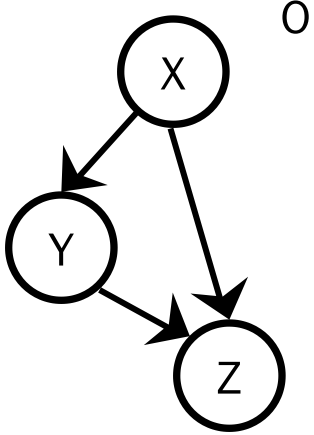
X	Y	Z
0	0	1
0	1	0
0	1	1
0	1	0
1	0	0
1	0	0
0	1	1
1	0	0
0	1	1
0	0	1
0	1	1
1	0	0

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

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

Log-likelihood

Explain this in your own words

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

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

Review: ML Parameters for Bayes Nets

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

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

Using the Indicator function I

- Which of the following correctly expresses the ML estimate $P_{ML}(X_i = x)$?

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: $A \rightarrow B$, $A \rightarrow C$, $B \rightarrow D$, $B \rightarrow E$, $E \rightarrow C$, $C \rightarrow F$, $D \rightarrow F$

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) $P(A = a)$

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

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

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

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

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

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.

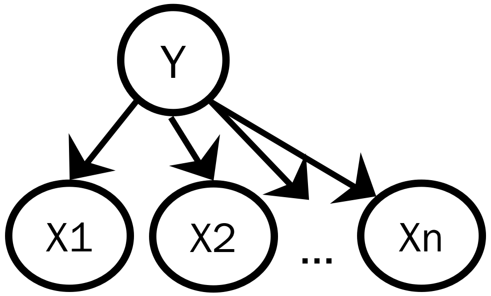
(f) $P(F = f | C = c, D = d)$

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

$Y \in \{1, 2, 3, \dots, k\}$ Where 1=sports, 2=fashion, 3=politics, ..., etc.
$X_i \in \{0, 1\}$ for i=1...n, where 0 means ith word in dictionary does not appear in document and 1 means it does

Naïve Bayes: Learning and Classification



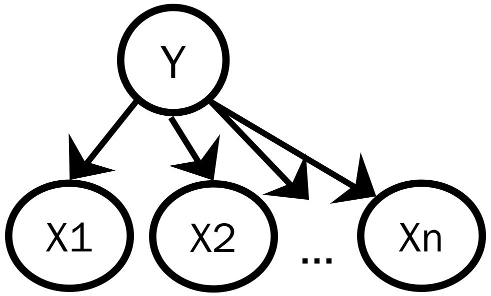
Learning:

$P(Y = y)$

$P(X_i = 1|Y = y)$

Classification

Naïve Bayes: Learning and Classification



<u>Learning:</u>

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

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

<u>Classification</u>

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

Strengths and weaknesses?

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. $P(w_1, w_2, w_3, \ldots, w_L) = \prod_{i=1}^{n} P(w_i | w_1, \ldots, w_{i-1})$

Two simplifying assumptions:

1. Finite Context

2. Position Invariance