Bayesian Networks

CISC 5800
Professor Daniel Leeds

Approaches to learning/classification

For classification, find highest probability class given features

- $P(x_1,\ldots,x_n|y=?)$

Approaches:

- Learn/use function(s) for probability
  - P(light|Y=eclipse)=$N(\mu_{eclipse}, \sigma_{eclipse})$
- Learn/use probability look-up table for each combination of features:

$letter_1$	P($letter_1$ | word="duck")
"a"	0.001
"b"	0.010
"c"	0.005
"d"	0.950

2

Joint probability over N features

Problem with learning table with N features:

- If all dependent, exponential number of model parameters

Burglar breaks in	Alarm goes off	Jill gets call	Zack gets call	P(A,J,Z|B)
Y	Y	Y	Y	0.3
Y	Y	Y	N	0.03
Y	Y	N	Y	0.03
Y	Y	N	N	0.06
		⋮		

3

Joint probability over N features

Naïve Bayes – all independent

- Linear number of model parameters

What if only **some** features are independent?

Burglar breaks in	Alarm goes off	Jill gets call	Zack gets call	P(A,J,Z|B)
Y	Y	Y	Y	0.3
Y	Y	Y	N	0.03
Y	Y	N	Y	0.03
Y	Y	N	N	0.06
		⋮		

4

Bayes nets: conditional independence

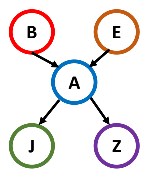
B – Burglar
E – Earthquake
A – Alarm goes off
J – Jill is called
Z – Zack is called

In Naïve Bayes: $P(x_1,x_2,x_3|y) = P(x_1|y)P(x_2|y)P(x_3|y)$

In Bayes nets, some variables depend on other variables:

Alarm depends on **B**urglar and **E**arthquake

Jill and **Z**ack calls each depend only on **A**larm



- P(B, E, A, J, Z) = P(B) P(E) P(A|B,E) P(J|A) P(Z|A)

Bayes nets: conditional independence

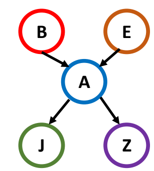
B – Burglar
E – Earthquake
A – Alarm goes off
J – Jill is called
Z – Zack is called

In Bayes nets, some variables depend on other variables:

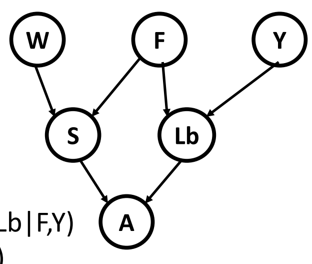
- P(B, E, A, J, Z) = P(B) P(E) P(A|B,E) P(J|A) P(Z|A)

In general for Bayes nets:

- $P(x_1,\ldots,x_n) = \prod_i P(x_i|Pa(x_i))$
- $Pa(x_i)$ are the "parents" of x_i – the variables x_i is conditioned on



Another Example: Health probabilities



F – Flu
S – Stress
Y – Age (years)
Lb – Weight
W – Weather
A – Activity

P(W,F,Y,S,Lb,A) =
P(W)P(F)P(Y)
P(S|W,F)P(Lb|F,Y)
P(A|S,Lb)

Probability review

Conditional Probabilities:

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

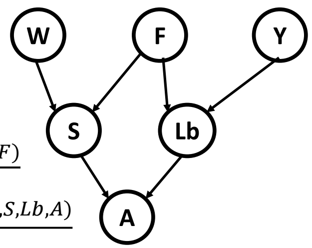
Marginalization:

- $P(A) = \sum_{b \in B} P(A, B = b)$

Variable elimination:

- $\frac{P(A)P(B)}{P(B)} = P(A)$

Health probabilities, find P(S,Lb,A | F)



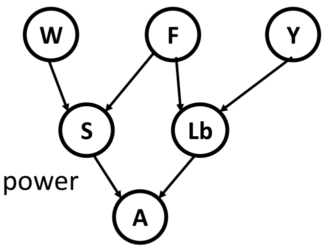
F – Flu
S – Stress
Y – Age (years)
Lb – Weight
W – Weather
A – Activity

P(S,Lb,A|F) $=\frac{P(S,Lb,A,F)}{P(F)}$

$$=\frac{\sum_{w\in W}\sum_{y\in Y}P(W,F,Y,S,Lb,A)}{P(F)}$$

$$=\frac{\sum_{w\in W}\sum_{y\in Y}\boldsymbol{P(F)P(W)P(Y)P(S|W,F)P(Lb|F,Y)P(A|S,Lb)}}{\boldsymbol{P(F)}}$$

Health probabilities, find P(S,Lb,A | F)



Moving variables out of irrelevant summation loops saves computation power

P(S,Lb,A|F)

$$=\frac{\sum_{w\in W}\sum_{y\in Y}\boldsymbol{P(F)P(W)P(Y)P(S|W,F)P(Lb|F,Y)P(A|S,Lb)}}{\boldsymbol{P(F)}}$$

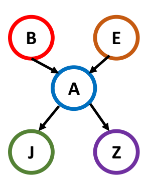
$$=\frac{\cancel{P(F)}P(A|S,Lb)\sum_{w\in W}P(W)P(S|W,F)\sum_{y\in Y}P(Y)P(Lb|F,Y)}{\cancel{P(F)}}$$

Example evaluation of Bayes nets

Use joint probabilities to find more probable class-variable value

Compute P(E=yes|A,J,Z), P(E=no|A,J,Z)

B – Burglar
E – Earthquake
A – Alarm goes off
J – Jill is called
Z – Zack is called



$$\boldsymbol{P(E|A,J,Z)=\frac{P(E,A,J,Z)}{P(A,J,Z)}=\frac{\sum_B P(E,B,A,J,Z)}{\sum_E\sum_B P(E,B,A,J,Z)}}$$

$$\boldsymbol{=\frac{\sum_b P(E)P(B=b)P(A|E,B=b)P(J|A)P(Z|A)}{\sum_e\sum_b P(E=e)P(B=b)P(A|E=e,B=b)P(J|A)P(Z|A)}}$$

Speeding up Bayes inference

D – Duck
F – Flies
Q – Quacks
S – Swims
T – Turtle
W – Water

P(D|S=yes,F=no):

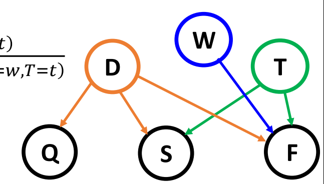
~118 operations total (add, multiply, divide lookup)

Multiply probabilities of listed variables and any conditioned variables

$$\frac{P(D,S=y,F=n)}{P(S=y,F=n)}=\frac{\sum_t\sum_w P(S=y|D,T=t)P(F=n|D,W=w,T=t)P(D)}{\sum_t\sum_w\sum_d P(S=y|D=d,T=t)P(F=n|D=d,W=w,T=t)P(D=d)}$$

Avoid repeat table look-ups

$$=\frac{P(D)\sum_t P(S=y|D,T=t)\sum_w P(F=n|D,W=w,T=t)}{\sum_d P(D=d)\sum_t P(S=y|D=d,T=t)\sum_w P(F=n|D=d,W=w,T=t)}$$



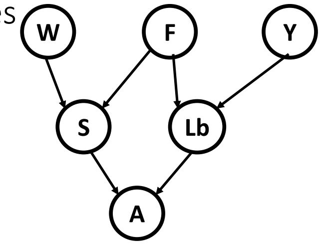
$$=\frac{f(D)}{\sum_d f(d)}$$

where

~53 operations total (add, multiply, divide lookup)

$\mathrm{f(d)}=P(D=d)\sum_t P(S=y|D=d,T=t)\sum_w P(F=n|D=d,W=w,T=t)$

Learning Bayes net probabilities



$$P(W = cold) = \frac{\#D\{W{=}cold\}}{|D|}$$

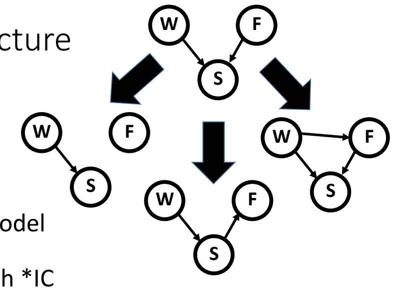
$$P(Lb = high|F = no, Y = young)$$
$$= \frac{\#D\{Lb{=}high \wedge F{=}no \wedge Y{=}Young\}}{\#D\{F{=}no \wedge Y{=}Young\}}$$

Or

$$P(Lb = high|F = no, Y = young):\ \mu_{Lb} = \frac{\sum_{i \in F=no,Y=young} Lb^i}{\#D\{F{=}no \wedge Y{=}Young\}}$$

14

Learning Bayes net structure



- Start with guessed structure
- Make one modification
- Learn probabilities for new model
- Evaluate net performance with *IC

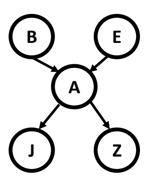
15

Conditional independence

If two variables x_i and x_j share same "parent," the x_i and x_j are independent given that parent

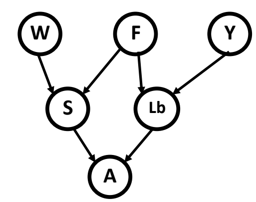
J and Z are independent given A: "$J \perp Z|A$"

B – Burglar
E – Earthquake
A – Alarm goes off
J – Jill is called
Z – Zack is called



16

Learning with partial data



- Initialize net probabilities
- Assign missing values based on net probabilities
- Update net probabilities

W	F	Y	S	Lb	A
Wind	Yes	Old	Low	Light	High
Snow	No	Mid	Mid	Mid	Low
?	No	Young	Mid	?	High
Rain	?	Old	?	Heavy	Mid
Sun	Yes	?	High	Light	?
Wind	No	Mid	Low	Mid	High

18

Expectation-Maximization

Problem: Uncertain of y^i (class), uncertain of θ^i (parameters)

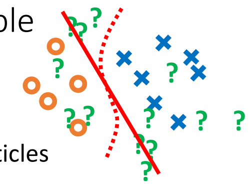
Solution: Guess y^i, deduce θ^i, re-compute y^i, re-compute θ^i ... etc.

OR: Guess θ^i, deduce y^i, re-compute θ^i, re-compute y^i

Will converge to a solution

- E step: Fill in expected values for missing variables
- M step: Regular MLE given known and filled-in variables

Document classification example



Two classes: {farm, zoo}

- 5 labeled zoo articles, 5 labeled farm articles
- 100 unlabeled training articles

Features: [% bat, % elephant, % monkey, % snake, % lion, %penguin]

Logistic regression classifier

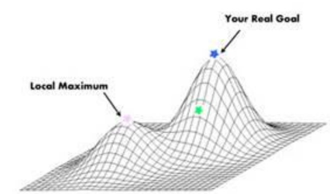
Merge knowledge from labeled and unlabeled data

Iterative learning

Learn **w** with labeled training data
Use classifier to assign labels to originally unlabeled training data
Learn **w** with known and newly-assigned labels
Use classifier to re-assign labels to originally unlabeled training data

Converges to a stable answer

Local vs global optimum



- EM increases probability at each step
- Reaches **local** maximum

To seek “global maximum”

- Re-start EM at different locations in label/parameter space

~~Same principle in logistic regression gradient ascent~~

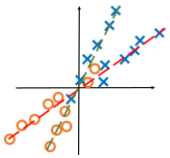
Types of learning

Supervised: each training data point has known features and class label

- Most examples so far

Unsupervised: each training data point has known features, but no class label

- ICA – each component meant to describe subset of data points



Semi-supervised: each train data point has known features, but only some have class labels

- Related to expectation maximization

24