Bayesian Networks

CISC 5800
Professor Daniel Leeds

Approaches to learning/classification

For classification, find highest probability class given features

- P(x_1,…,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

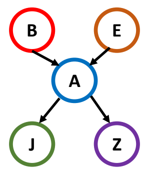
B – Burglar
E – Earthquake
A – Alarm goes off
J – Jill is called
Z – Zack is called

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)

5

Bayes nets: conditional independence

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

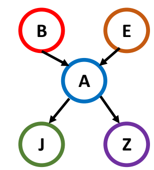
In Bayes nets, some variables depend on other variables:

- 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



6

Probability review

Conditional Probabilities:

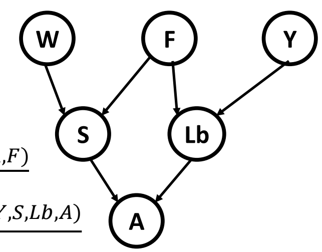
- $P(A|B) = \frac{P(A,B)}{P(B)}$

Marginal Probability

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

8

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



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

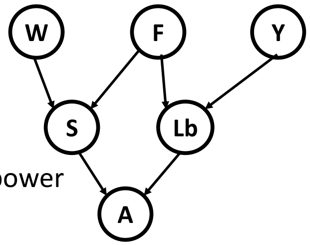
$P(S,Lb,A|F) = \frac{P(S,Lb,A,F)}{P(F)}$

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

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

10

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



Moving variables out of irrelevant summation loops saves computation power

P(S,Lb,A|F)

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

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

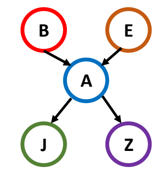
11

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



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

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

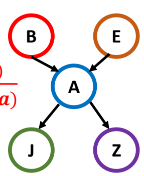
13

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



$$P(E=yes|A=a,J=j,Z=z)$$

$$= \frac{\sum_B P(E=yes)P(B)P(A=a|E=yes,B)P(J=j|A=a)P(Z=z|A=a)}{\sum_E \sum_B P(E=yes)P(B)P(A=a|E=yes,B)P(J=j|A=a)P(Z=z|A=a)}$$

$$= \frac{\cancel{P(J=j|A=a)P(Z=z|A=a)}P(E=yes) \sum_B P(B)P(A=a|E=yes,B)}{\cancel{P(J=j|A=a)P(Z=z|A=a)} \sum_E \sum_B P(E)P(B)P(A=a|E,B)}$$

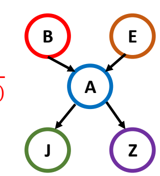
14

Variable elimination

Pull constant terms outside the sigma-sum loop

Cancel out constants appearing in both numerator and denominator

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



$$P(E=yes|A=a,J=j,Z=z)$$

$$= \frac{\sum_B P(E=yes)P(B)P(A=a|E=yes,B)P(J=j|A=a)P(Z=z|A=a)}{\sum_E \sum_B P(E=yes)P(B)P(A=a|E=yes,B)P(J=j|A=a)P(Z=z|A=a)}$$

$$= \frac{\cancel{P(J=j|A=a)P(Z=z|A=a)}P(E=yes) \sum_B P(B)P(A=a|E=yes,B)}{\cancel{P(J=j|A=a)P(Z=z|A=a)} \sum_E \sum_B P(E)P(B)P(A=a|E,B)}$$

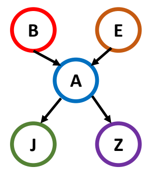
15

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



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

Also useful when there are holes in your data

EM example

Missing data in training set:
- E=yes, J=yes, Z=no
- Unknown: class B (burglary), feature A (alarm)

- Estimate A with a “random” guess
- Loop
  - Estimate B=argmax_B P(B | E=yes, J=yes, Z=no, A=$A_{estimate}$)
  - Estimate A=argmax_A P(A | E=yes, J=yes, Z=no, B=$B_{estimate}$)

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]
- E.g., % bat^i = #{$wordsInArticle^i$==bat}/#{$wordsInArticle^i$}

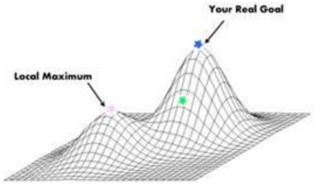
Logistic regression classifier

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

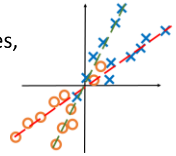
Types of learning

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