Follow-ups to HMMs
Graphical Models
Semi-supervised learning

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(rumble|Y=dog)=$N(\mu_{dog},\sigma_{dog})$
- Learn/use probability look-up table for each combination of features:

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

Joint probability over N features

Problem with learning table with N features:
- If all dependent, exponential number of model parameters

Naïve Bayes – all independent
- Linear number of model parameters

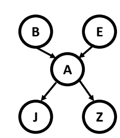
Burglar breaks in	Alarm goes off	Jill gets a call	Zack gets a call	P(A,J,Z|B)
Y	Y	Y	Y	0.3
Y	Y	Y	N	0.03
Y	Y	N	Y	0.03
Y	Y	N	N	0.06
		⋮		

What if not all features are independent?

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:
  - 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

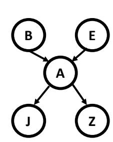


Example evaluation of Bayes nets

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

- Use joint probabilities to find more probable variable value
- Compute P(E=yes|A,J,Z), P(E=no|A,J,Z)

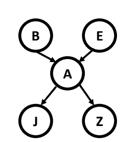
$$P(E|A,J,Z) = \frac{P(E,A,J,Z)}{P(A,J,Z)} = \frac{\sum_B P(E,B,A,J,Z)}{\sum_E \sum_B P(E,B,A,J,Z)}$$
$$= \frac{\sum_B P(E)P(B)P(A|E,B)P(J|A)P(Z|A)}{\sum_E \sum_B P(E)P(B)P(A|E,B)P(J|A)P(Z|A)}$$



Conditional independence

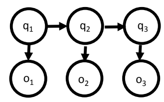
B – Burglar
E – Earthquake
A – Alarm goes off
J – Jill is called
Z – Zack is called

- 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: "$J \perp Z|A$"



HMM: a kind-of example of Bayes nets

- $P(q_1,q_2,q_3,o_1,o_2,o_3)$ =



7

Back to 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

8

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

9

Document classification example

Two classes: {farm, zoo}
- 5 labeled zoo articles, 5 labeled jungle articles
- 100 unlabeled training articles

Features: [% bat, % elephant, % monkey, % snake, % lion, %penguin]
- E.g., % bat^i = #{$wordsInArticle^i$==bat}/#{$wordsInArticle^i$}

Logistic regression classifier

10

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

11