Expectation Maximization

Introduction to
Artificial Intelligence
COS302
Michael L. Littman
Fall 2001

Administration

- Exams halfway graded. They assure me they will be working over Thanksgiving break.
- Project groups.
- Next week, synonyms via web.
- Week after, synonyms via wordnet. (See web site.)

Plan

- Connection between learning from data and finding a maximum likelihood (ML) model
- ML from complete data
- EM: ML with missing data
- EM for HMMs
- QED, PDQ, MOUSE

Learning from Data

- We want to learn a model with a set of parameter values M.
- We are given a set of data D.
- An approach: argmax_M Pr(M|D)
- This is the *maximum likelihood* model (ML).
- How relate to Pr(D|M)?

Super Simple Example

- Coin I and Coin II. (Weighted.)
- Pick a coin at random (uniform).
- Flip it 4 times.
- Repeat.

What are the parameters of the model?

Data

Coin I	Coin II
HHHT	TTTH
HTHH	THTT
HTTH	TTHT
THHH	HTHT
HHHH	HTTT

Probability of D Given M

p: Probability of H from Coin I

q: Probability of H from Coin II

Let's say h heads and t tails for Coin I. h' and t' for Coin II.

$Pr(D|M) = p^h (1-p)^t q^{h'} (1-q)^{t'}$

How maximize this quantity?

Maximizing p

$D_p(p^h (1-p)^t q^{h'} (1-q)^{t'}) = 0$

$D_p(p^h)(1-p)^t + p^h D_p((1-p)^t) = 0$

$h p^{h-1} (1-p)^t = p^h t(1-p)^{t-1}$

$h (1-p) = p t$

$h = p t + hp$

$h/(t+h) = p$

Duh…

Missing Data

HHHT	HTTH
TTTH	HTHH
THTT	HTTT
TTHT	HHHH
THHH	HTHT

Oh Boy, Now What!

If we knew the labels (which flips from which coin), we could find ML values for p and q.

What could we use to label?

p and q!

Computing Labels

p = ¾, q = 3/10

Pr(Coin I | HHTH)

= Pr(HHTH | Coin I) Pr(Coin I) / c

$= (3/4)^3(1/4)\ (1/2)/c = .052734375/c$

Pr(Coin II | HHTH)

= Pr(HHTH | Coin II) Pr(Coin II) / c

$= (3/10)^3(7/10)\ (1/2)/c = .00945/c$

Expected Labels

	I	II		I	II
HHHT	.85	.15	HTTH	.44	.56
TTTH	.10	.90	HTHH	.85	.15
THTT	.10	.90	HTTT	.10	.90
TTHT	.10	.90	HHHH	.98	.02
THHH	.85	.15	HTHT	.44	.56

Wait, I Have an Idea

Pick some model M_0

Expectation

• Compute expected labels via M_i

Maximization

• Compute ML model M_{i+1}

Repeat

Could This Work?

Expectation-Maximization (EM)

$Pr(D|M_i)$ will not decrease.

Sound familiar? Type of search.

Coin Example

Compute expected labels.

Compute counts of heads and tails (fractions).

Divide to get new probabilities.

p=.63 q=.42 $Pr(D|M)=9.95 \times 10^{-13}$

p=.42 q=.63 $Pr(D|M)=9.95 \times 10^{-13}$

p=.52 q=.52 $Pr(D|M)=9.56 \times 10^{-13}$

More General EM

Need to be able to compute probabilities: generative model

Need to tabulate counts to estimate ML model

Let's think this through with HMMs.

Recall HMM Model

N states, M observations

$\pi(s)$: prob. starting state is s

p(s,s'): prob. of s to s' transition

b(s, k): probability of obs k from s

$k_0\ k_1 \ldots k_l$: observation sequence

$\text{argmax}_{\pi, p, b} Pr(\pi, p, b \mid k_0\ k_1 \ldots k_l)$

ML in HMM

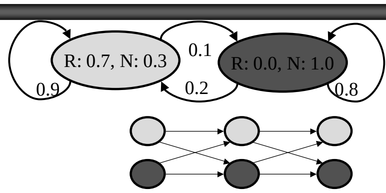
How estimate π, p, b?

What's the missing information?

$k_0\ k_1 \ldots k_l$

$s_0\ s_1 \ldots s_l$

$Pr(s_t=s|N\ N\ N)$



	N	N	N
UP			
DOWN			

Forward Procedure

- $\alpha(s,t)$: probability of seeing first t observations *and* ending up in state s: $Pr(k_0 \ldots k_t, s_t=s)$
- $\alpha(s,0) = \pi(s)\ b(k_0, s)$
- $\alpha(s,t) = sum_{s'}\ b(k_t,s)\ p(s,s')\ \alpha(s',t-1)$

Backward Procedure

- $\beta(s,t)$: probability of seeing observations from t to l given that we start in state s:
  $Pr(k_{t+1} \ldots k_l \mid s_t=s)$
- $\beta(s,l+1) = 1$
- $\beta(s,t)=sum_s\ p(s,s')\beta(s',t+1)\ b(k_{t+1},s)$

Combing α and β

- Want to know $Pr(s_t=s \mid k_0 \ldots k_l)$
- $= Pr(k_0 \ldots k_l, s_t=s\ ) / c$
- $= Pr(k_0 \ldots k_t\ k_{t+1} \ldots k_l, s_t=s\ ) / c$
- $= Pr(k_0 \ldots k_t, s_t=s\ )$
  $Pr(k_{t+1} \ldots k_l \mid k_0 \ldots k_t, s_t=s\ ) / c$
- $= Pr(k_0 \ldots k_t, s_t=s\ )\ Pr(k_{t+1} \ldots k_l \mid s_t=s\ ) / c$
- $= \alpha(s,t)\ \beta(s,t) / c$

EM For HMM

- Expectation: Forward-backward (Baum-Welch)
- Maximization: Use counts to reestimate parameters
- Repeat.
- Gets stuck, but still works well.

What to Learn

- Maximum Likelihood (counts)
- Expectation (expected counts)
- EM
- Forward-backward for HMMs

Homework 8 (due 11/28)

1. Write a program that decides if a pair of words are synonyms using the web. I'll send you the list, you send me the answers.
2. Recall the naïve Bayes model in which a class is chosen at random, then features are generated from the class. Consider a simple example with 2 classes with 3 binary features. Let's use EM to learn a naïve Bayes

(continued)

model. (a) What are the parameters of the model? (b) Imagine we are given data consisting of the two feature values for each sample from the model. We are not given the class label. Describe an "expectation" procedure to compute class labels for the data given a model. (c) How do you use this procedure to learn a maximum likelihood model for the data?

Homework 9 (due 12/5)

1. Write a program that decides if a pair of words are synonyms using wordnet. I'll send you the list, you send me the answers.
2. more soon