Expectation Maximization
Lecture 6: Sept 27, 2013

CS886-2 Natural Language Understanding

University of Waterloo

Text Categorization

- Algorithms
  - Naïve Bayes model
  - Support vector machines
  - Decision Trees
- Assumption: labeled articles available for training
  - Hand labeling is expensive
  - Lots of unlabeled documents available

Text Categorization

- Supervised learning: categorization
  - Data: all training documents labeled with category
  - Alg: Relative frequency counts (maximum likelihood)
- Unsupervised learning: clustering
  - Data: no training document labeled with category
  - Alg: Expectation maximization (approx. max. likelihood)
- Semi-supervised learning: mix of categorization and clustering
  - Data: some labeled documents with some unlabeled documents
  - Alg: Expectation maximization (approx. max. likelihood)

Supervised Maximum Likelihood

- Notation:
  - h: hypothesis
  - $\boldsymbol{E}$: observable variables
  - $\boldsymbol{e}$: evidence (data)
- Supervised maximum likelihood

$$h_{ML} = \text{argmax}_h \Pr(\boldsymbol{e}|h)$$

- Supervised text categorization

$$h_{ML} = \text{argmax}_h \Pr(t, w_1, w_2, w_3, \ldots|h)$$

Unsupervised Maximum Likelihood

- Notation:
  - $\boldsymbol{Z}$: hidden (unobservable) variables
- Unsupervised Maximum likelihood

$$h_{ML} = \text{argmax}_h \Pr(\boldsymbol{e}|h)$$
$$= \text{argmax}_{\boldsymbol{h}} \sum_Z \Pr(\boldsymbol{Z}, \boldsymbol{e}|h)$$

- Unsupervised Text Categorization

$$h_{ML} = \text{argmax}_h \Pr(w_1, w_2, w_3, \ldots|h)$$
$$= \text{argmax}_h \sum_T \Pr(T, w_1, w_2, w_3, \ldots|h)$$

"Direct" maximum likelihood

- Notation: $V_1, V_2, \ldots$ (all hidden and obs. variables)
  - E.g.: $V_1 = T,\ V_2 = W_1, V_3 = W_2, \ldots$
- Directly maximizing likelihood in the unsupervised case is difficult
- $h_{ML} = argmax_h \sum_Z \Pr(Z, e|h)$
  $= argmax_h \sum_Z \prod_l \Pr(V_l|parents(V_l))$
  $= argmax_h \log \sum_Z \prod_l \Pr(V_l|parents(V_l))$

Problem: can't push log past sum to linearize product

Expectation-Maximization (EM)

- EM algorithm
  - Intuition: if we knew the missing values, computing h_{ML} would be trivial
- Guess h_{ML}
- Iterate
  - Expectation: based on h_{ML}, compute expectation of the missing values
  - Maximization: based on expected missing values, compute new estimate of h_{ML}

Expectation-Maximization (EM)

- More formally:
  - Approximate maximum likelihood
  - Iteratively compute:

$$h_{j+1} = argmax_h \underbrace{\sum_Z \Pr(Z|h_j, \boldsymbol{e}) \log \Pr(\boldsymbol{e}, \boldsymbol{Z}|h)}_{\text{Expectation}}$$

Maximization

Expectation-Maximization (EM)

- Derivation

$$\begin{aligned}\log \Pr(e|h) &= \log\left[\frac{\Pr(e,Z|h)}{\Pr(Z|e,h)}\right] \\ &= \log \Pr(e,Z|h) - \log \Pr(Z|e,h) \\ &= \sum_Z \Pr(Z|e,h) \log \Pr(e,Z|h) \\ &\quad - \sum_Z \Pr(Z|e,h) \log \Pr(Z|e,h) \\ &\geq \sum_z \Pr(Z|e,h) \log \Pr(e,Z|h)\end{aligned}$$

- EM finds a local maximum of

$$\sum_Z \Pr(Z|e,h) \log \Pr(e,Z|h)$$

which is a lower bound of log P(**e**|h)

Expectation-Maximization (EM)

- Log inside sum can linearize product

$$\begin{aligned}h_{j+1} &= argmax_h \sum_{\boldsymbol{Z}} \Pr(\boldsymbol{Z}|h_j, \boldsymbol{e}) \log \Pr(\boldsymbol{e}, \boldsymbol{Z}|h) \\ &= argmax_h \sum_{\boldsymbol{Z}} \Pr(\boldsymbol{Z}|h_j, \boldsymbol{e}) \log \prod_l \Pr(V_l|parents(V_l), h) \\ &= argmax_h \sum_{\boldsymbol{Z}} \Pr(\boldsymbol{Z}|h_j, \boldsymbol{e}) \sum_l \log \Pr(V_l|parents(V_l), h)\end{aligned}$$

- Monotonic improvement of likelihood

$$\Pr(\boldsymbol{e}|h_{j+1}) \geq \Pr(\boldsymbol{e}|h_j)$$

Algorithm

- Iterative Optimization

$h_{j+1} = argmax_h \sum_{\boldsymbol{Z}} \Pr(\boldsymbol{Z}|h_j, \boldsymbol{e}) \sum_l \log \Pr(V_l|parents(V_l), h)$

Initialize h_0

Alternate between

1) Estimate Z based on h_j by computing

$$\Pr(\boldsymbol{Z}|h_j, \boldsymbol{e})$$

2) Find h_{j+1} according to

$h_{j+1} = argmax_h \sum_{\boldsymbol{Z}} \Pr(\boldsymbol{Z}|h_j, \boldsymbol{e}) \sum_l \log \Pr(V_l|parents(V_l), h)$

Text Categorization Example

- Let $\theta = \Pr(T = sports)$

  $\theta_{si} = \Pr(W_i = 1|T = sports)$

  $\theta_{fi} = \Pr(W_i = 1|T = finance)$

- Corpus:

  $\#s$: # of documents about sports

  $\#f$: # of documents about finance

  $\#s1i$: # of documents about sports with w_i

  $\#s0i$: # of documents about sports without w_i

  $\#f1i$: # of documents about finance with w_i

  $\#f0i$: # of documents about finance without w_i

EM for Text Categorization

Algorithm summary

Initialize $\theta, \theta_{si}, \theta_{fi}$ (with any values)

Repeat

For each document d compute

$$prob(d,s) \leftarrow \Pr\left(T^d = sports \middle| \theta, \theta_{si}, \theta_{fi}, w_i^d \ \forall i\right)$$

$$prob(d,f) \leftarrow \Pr\left(T^d = finance \middle| \theta, \theta_{si}, \theta_{fi}, w_i^d \ \forall i\right)$$

$E[\#s] \leftarrow \sum_d prob(d,s), \quad E[\#f] \leftarrow \sum_d prob(d,f)$

$E[\#s1i] \leftarrow \sum_d \delta(\mathrm{w_i^d} = 1)\, prob(d,s),\ E[\#s0i] \leftarrow \sum_d \delta(\mathrm{w_i^d} = 0)\, prob(d,f) \ \ \forall i$

$E[\#f1i] \leftarrow \sum_d \delta(\mathrm{w_i^d} = 1)\, prob(d,s),\ E[\#f0i] \leftarrow \sum_d \delta(\mathrm{w_i^d} = 0)\, prob(d,s) \ \ \forall i$

$\theta \leftarrow \frac{E[\#s]}{E[\#s]+E[\#f]}, \quad \theta_{si} \leftarrow \frac{E[\#s1i]}{E[\#s1i]+E[\#s0i]}, \quad \theta_{fi} \leftarrow \frac{E[\#f1i]}{E[\#f1i]+E[\#f0i]} \ \forall i$

Until convergence

Data

- Data: 1000 documents
- Frequency: # of documents

W_1	W_2	W_3	Frequency
1	1	1	273
1	1	0	93
1	0	1	104
1	0	0	90
0	1	1	79
0	1	0	100
0	0	1	94
0	0	0	167

Initialization

- Set the parameters to some initial values
  - Make educated guess
  - Since EM converges to local optimum, initial values influence the final values that it will converge to
- Example:
  $\theta = 0.5,\ \ \theta_{si} = 0.8\ \ \forall i,\ \ \theta_{fi} = 0.3\ \ \forall i$

Estimate T

- 8 types of documents
- Estimate T for each type of document

W_1	W_2	W_3	$prob(d, s)$	$prob(d, f)$
1	1	1	$k(0.6)(0.8)(0.8)(0.8) = 0.966$	$k(0.4)(0.3)(0.3)(0.3) = 0.034$
1	1	0	$k(0.6)(0.8)(0.8)(0.2) = 0.753$	$k(0.4)(0.3)(0.3)(0.7) = 0.247$
1	0	1	$k(0.6)(0.8)(0.2)(0.8) = 0.753$	$k(0.4)(0.3)(0.7)(0.3) = 0.247$
1	0	0	$k(0.6)(0.8)(0.2)(0.2) = 0.246$	$k(0.4)(0.3)(0.7)(0.7) = 0.754$
0	1	1	$k(0.6)(0.2)(0.8)(0.8) = 0.753$	$k(0.4)(0.7)(0.3)(0.3) = 0.247$
0	1	0	$k(0.6)(0.2)(0.8)(0.2) = 0.246$	$k(0.4)(0.7)(0.3)(0.7) = 0.754$
0	0	1	$k(0.6)(0.2)(0.2)(0.8) = 0.246$	$k(0.4)(0.7)(0.7)(0.3) = 0.754$
0	0	0	$k(0.6)(0.2)(0.2)(0.2) = 0.034$	$k(0.4)(0.7)(0.7)(0.7) = 0.966$

Expected Frequencies

$$E[\#s] = 0.966(273) + 0.753(93 + 104 + 79) + 0.246(90 + 100 + 94) + 0.034(167) = 547$$
$$E[\#f] = 1000 - 547 = 453$$

$$E[\#s11] = 0.966(273) + 0.753(93 + 104) + 0.246(90) = 434$$
$$E[\#s12] = 0.966(273) + 0.753(93 + 79) + 0.246(100) = 418$$
$$E[\#s13] = 0.966(273) + 0.753(104 + 79) + 0.246(94) = 425$$

$$E[\#f11] = 0.034(273) + 0.247(93 + 104) + 0.754(90) = 126$$
$$E[\#f12] = 0.034(273) + 0.247(93 + 79) + 0.754(100) = 127$$
$$E[\#f13] = 0.034(273) + 0.247(104 + 79) + 0.754(94) = 125$$

Revised Parameters After One Iteration

$$\theta = \frac{547}{1000} = 0.547$$

$$\theta_{s1} = \frac{434}{547} = 0.793, \qquad \theta_{f1} = \frac{126}{453} = 0.278$$

$$\theta_{s2} = \frac{418}{547} = 0.764, \qquad \theta_{f2} = \frac{127}{453} = 0.280$$

$$\theta_{s3} = \frac{425}{547} = 0.777, \qquad \theta_{f3} = \frac{125}{453} = 0.276$$

Typical Log-Likelihood Curve