# Assignment #6 Mixture Models and EM Algorithm

# 1 Mixture Models

1. If the author is and the probability of writing “the” is:

If the author is and the probability of writing “the” is:

Finally, the probability of the first letter being “the” is

1. Since each word is written independently, the probability of the second letter being “the” is the same
2. The probability of the first letter is “data” and written by author :

The probability of the first letter is “data” and written by author :

The probability of the first letter is “data”:

The probability that letter “data” was written by author is:

1. The letter “data” will occur least frequently, as it has both least frequencies (0.1 for and 0.1 for ) among these 5 words for both authors.
2. With the ,

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | the | computer | data | baseball | game |
|  | 0.3 | 0.3 | 0.2 | 0 | 0.2 |

The probability of the letter “computer” is

The probability of the letter “game” is

# 2 EM Algorithm

1. The probability of generating a word is
2. The log-likelihood of the whole document is:
3. We need || binary hidden variables in total.

Because for each occurrence of a word , a binary hidden variable

will be needed to indicate whether the word has been “generated” from the background model or the topic model

1. The redundancy of w.r.t. as

Define

The corresponding missing data ,

where ,

The complete likelihood of model give data:

For simplicity, I will use , , , , stands for background model

In the E-step, to compute expectation of over :

In the M-step, find to maximize :

Compute derivation:

Therefore, ,

#3 PLSA

1. Yes. We can randomly select some document as training set and use PLSA to generate topics learned during training. Then we can use this generated document-topic distribution information for classification.
2. Yes. We can these city-included document as labeled data and apply PLSA to generate topics models. Then using these topic model to classify corpus.
3. We can use PLSA-EM algorithm
4. Repeat Estep and Mstep until converge:

Estep – calculate

Mstep – calculate and with some prior knowledge of city-included documents

2) Classify corpus based on generated topic models .

1. I will apply the PLSA generated classier on different labeled datasets to check the classification accuracy.
2. If given city-included document is limited, short or misleading, it will have worse performance as PLSA cannot have enough information to generate a satisfying topic model.

In my assumption, these city-included labeled document should be rich and correct at topic information in order to allow PLSA to generate appropriate topics learned from training set.