6.4 Summary

In this chapter we have introduced:

- The concept of hidden variables
- Hard and soft variations of the Expectation Maximization (EM) algorithm
- The correlation between EM and MLE for training probabilistic models
- EM for unsupervised text classification
- IBM model 1 for statistical machine translation
- Probabilisitc latent semantic allocation.

Exercises

1. Give a collection of six documents as shown in Table 6.2, use the unsupervised Naïve Bayes model to cluster the documents into: (1) 2 classes (2) 3 classes. Initialize the model parameter $P(\mathbf{h}|\Theta)$ with $\frac{1}{K}$ for every class $\mathbf{h}$, where K is the total number of classes. For every class $\mathbf{h}$, inititalize the model parameter $P(\mathbf{w}|\mathbf{h},\Theta)$ with $\frac{1}{|V|}$, where $|V|$ is the vocabulary size. Estimate the model parameters $P(\mathbf{h}|\Theta)$ and $P(\mathbf{w}|\mathbf{h},\Theta)$ according to Eq 6.21 and Eq 6.22, respectively. Compare the the 2-class clustering results with the 3-class ones.

Document	Document
Apple released iPod .	Tom bought one iPod .
Apple released iPhone .	Tom bought one iPhone .
Apple released iPad .	Tom bought one iPad .

Table 6.2: A collection of documents for clustering.

2. Consider a semi-supervised settings for the Naïve Bayes model, where a set of labeled documents $D = \{(d_i, c_i)\}|_{i=1}^{N}$ and a set of unlabeled documents $U = \{d_i\}|_{N+1}^{N+M}$ are avaliable. The training objective is to maximize

$$L(\Theta) = \sum_{i=1}^{N} \log P(d_i, c_i|\Theta) + \sum_{j=N+1}^{N+M} \log P(d_j|\Theta)$$

   (a) Describe how to train the model parameters Θ using an algorithm similar to Algorithm 2.

Id	Source	Target
1	他(he) 住(live) 在(in) 上海(Shanghai)	He is living in Shanghai
2	他(he) 喜欢(like) 上海(Shanghai)	He likes Shanghai
3	他(he) 喜欢(like) 住(live) 在(in) 上海(Shanghai)	He likes living in Shanghai

Table 6.3: A parallel corpus.

(b) What is the role of the unlabeled data in this training objective? If we add a hyper-parameter λ to indicate that how much attention we should pay to the unlabeled data, the training objective becomes,

$$L(\Theta) = \sum_{i=1}^{N} \log P(d_i, c_i|\Theta) + \lambda \sum_{j=N+1}^{N+M} \log P(d_j|\Theta)$$

Compare the second term $\lambda \sum_{j=N+1}^{N+M} \log P(d_j|\Theta)$ with the L2-regularizer introduced in Chapter 3.

3. Given a parallel corpus as shown in Table 6.3, (1) execute IBM model 1 for one iteration and show the model parameters; (2) suppose that we already have a location dictionary, indicating that "Shanghai" and "上海" should always be connected, which means $P(\text{上海}|Shanghai) = 1$, run IBM model 1 from scratch again and show the model parameters.

4. Prove the last step of Eq 6.25.

(Hint: Calculate $\sum_{a_{|X|}=0}^{|Y|} \prod_{i=1}^{|X|} P(x_i|y_{a_i}) = \left(\prod_{i=1}^{|X|-1} P(x_i|y_{a_i=0})\right) \sum_{j=0}^{|Y|} P(x_{|X|}|y_j)$ first, and then $\sum_{a_{|X|-1}=0}^{|Y|} \sum_{a_{|X|}=0}^{|Y|} \prod_{i=1}^{|X|} P(x_i|y_{a_i}) = \prod_{i=1}^{|X|-2} P(x_i|y_{a_i=0}) \sum_{j=0}^{|Y|} P(x_{|X|-1}|y_j) \sum_{j=0}^{|Y|} P(x_{|X|}|y_j)$ before deriving $\sum_{a_1=0}^{|Y|} \sum_{a_2=0}^{|Y|} \cdots \sum_{a_{|X|}=0}^{|Y|} \prod_{i=1}^{|X|} P(x_i|y_{a_i}) \prod_{i=1}^{|X|} \sum_{j=0}^{|Y|} P(x_i|y_j)$.)

Id	Document	Id	Document
1	World Cup, Russia, host	2	World Cup, boost, Russia, economy
3	Russia, bid, World Cup	4	Russia, economy, growing, oil
5	Russia, economy, recover, continue	6	Russia, oil, dependence

Table 6.4: A collection of documents for latent topic analysis.

5. Prove the last step of Eq 6.27.

6. Given a document collection as shown in Table 6.4, suppose that there are two latent topic, one is about "World Cup" and the other is about "Russia's economy", (1) use PLSA to estimate the document-topic and

topic-word probabilities; (2) compare the similarilities of document pairs $< d_1, d_3 >$, $< d_4, d_5 >$ and $< d_2, d_5 >$ using the document-topic distribution.

7. Self-training in Chapter 4 and hard EM are somehow similar. They both predict labels for unlabeled instances and make use of automatically generated labels for iterative training. Show similarities and differences between self-training and hard EM.