Exercises

1. Consider a local model for solving the POS tagging problem, using features from $[w_{i-2}, w_{i-1}, w_i, w_{i+1}, w_{i+2}]$ to predict $t_i \in L$ given an input sentence $s = W_1^n$, where L is the set of all possible POS labels.

   (a) Build a Naïve Bayes classifier by treating $[w_{i-2}, w_{i-1}, w_i, w_{i+1}, w_{i+2}]$ as a document, and t_i as its class label. What are the parameter types and parameter instances? Draw a figure to tell the generative story of the model.

   (b) Since the input is fix-sized, can you relax the iid assumption between words, modelling w_{i-2}, w_{i-1}, w_i, w_{i+1} and w_{i+2} differently, sentsitive to position? What are the parameter types and parameter instances now? Compare the size of the model with the model in question (a) with regard to the number of parameter instances.

(c) Build a discriminative linear model using the same parameter types as (b). What are the feature templates? What is the form of a feature vector? Can you use additional features?

(d) Implement the models of questions (a), (b) and (c) above, comparing the performances using a POS-tagging benchmark, which consists of training, development and test sets with manually assigned POS labels. For (c), use SVM, a log-linear model or perceptron.

2. Implement a first-order HMM and a second-order HMM for POS tagging. Compare the performances on the dataset of Exercise 1. A 0th-order HMM makes independence assumptions between output POS tags. Implement a 0th-order HMM, comparing it with the model of Exercise 1(a) and (b). What are the differences in model structures? What are the differences in empirical performances?

3. Given three sentences and their frequence counts in Table 7.2, suppose there are three possible hidden states $\{N, V, D\}$, finish the following exercises.

Sentence	Frequency
John loves the cat	10
John loves Mary	10
Mary loves the cat	20

Table 7.2: Example sentences.

(a) Estimate the parameters for a first-order HMM using EM.

(b) Suppose the hidden tagging sequences are partially observable, how to change the standard EM algorithms for parameter estimation. More specifically, when the word “the” is always associated with the tag “D”, what are the estimation results? Compare the estimation results with that of (a).

(c) As introduced before, we use $P(\mathbf{t}_1|\langle\text{B}\rangle)$ to describe the first tag being $\mathbf{t}_1$. This makes the estimation of the tag starting probabilitis is the same as that of the transition probabilities. Suppose now we do not use $\langle\text{B}\rangle$ in the beginning of the tagging sequence any more, and there is a parameter $\pi(\mathbf{t})$, which describes the probability of the first tag beging $\mathbf{t}$. Derive the estimations equations of EM algorithms for π.

4. Write out the pseudocode of EM for the second-order HMM.