

問題簡述:

大略如作業 1.1，但是這次沒有給 **encode** 和 **bigram** 的機率。因此作業目的為利用任意演算法從資料中 **train** 出第一字、**n-gram** 和 **encode** 的機率，再利用這個機率預測 **encode** 後的文章。

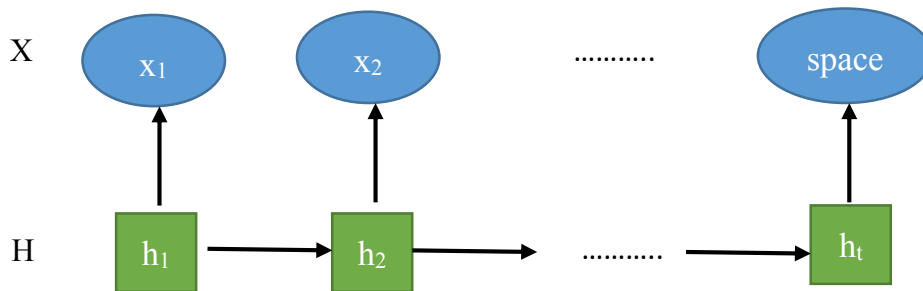
問題假設

根據題意，我的 **graphic model** 有以下基本假設：

1. 每個字元只跟前一個字元和 **encode** 後的字元相關
2. 每個單字結尾都有空白(除了最後一個單字)
3. 空白到任何字元的機率都固定，因為既然有算每個單詞的第一字的機率，那麼這個機率就不重要了。

解題方法:

graphic model 和上一題一樣(一個 **training example** 皆為一個單字)：



作法則是標準的 EM 演算法：

1. E step:

- (1) 利用 forward-backward 演算法求出 $\alpha_t(h_t)$ 和 $\beta_t(h_t)$:

$$\alpha_t(h_t) = P(x_{1:t}, h_t) = P(x_t | h_t) \sum_{h_{t-1}} \alpha_{t-1}(h_{t-1}) P(h_t | h_{t-1})$$

$$\beta_t(h_t) = P(x_{t+1:n} | h_t) = \sum_{h_{t+1}} \beta_{t+1}(h_{t+1}) P(h_t | h_{t+1}) P(x_{t+1} | h_{t+1})$$

- (2) 利用 $\alpha_t(h_n)$ 求出 $P(X)$ ($P(X) = P(x_{1:n})$)

$$P(X) = \sum_{h_n} \alpha_t(h_n)$$

- (3) 求出相關參數：

$$P(h_t = i, X) = \alpha_t(h_t = i) \beta_t(h_t = i)$$

$$P(h_1 = i, X) \text{ 為 } t = 1 \text{ 時的特例}$$

$$P(h_t = i, x_t = x, X) = P(h_t = i, X) \text{ (if } x_t = x), \text{ else equal } 0$$

$$P(h_t = i, h_{t+1} = j, X) = \alpha_t(h_t = i) P(h_{t+1} = j | h_t = i) P(x_{t+1} | h_{t+1} = j) \beta_{t+1}(h_{t+1} = j)$$

並且

$$P(h_1 = i | X) = \frac{P(h_1 = i, X)}{P(X)}$$

$$P(h_t = i, x_t = x | X) = \frac{P(h_t = i, x_t = x, X)}{P(X)}$$

$$P(h_t = i, h_{t+1} = j | X) = \frac{P(h_t = i, h_{t+1} = j, X)}{P(X)}$$

(4) 計算出累計機率：

$$C(h_1 = i) = \sum_X P(h_1 = i | X) \quad (\text{將所有單字的機率累計})$$

$$C(h = i, x) = \sum_X \sum_{t=1}^n P(h_t = i, x_t | X)$$

$$C(h_t = i, h_{t+1} = j) = \sum_X \sum_{t=1}^{n-1} P(h_t = i, h_{t+1} = j | X)$$

2. M step:

$$P(h_1 = i) = \frac{C(h_1 = i)}{C(h_1)}$$

$$P(x | h = i) = \frac{C(h = i, x)}{C(h = i)}$$

$$P(h_{t+1} = j | h_t = i) = \frac{C(h_t = i, h_{t+1} = j)}{C(h_t = i)}$$

解題技巧:

1. 為了避免 underflow，我用了上課的方法對所有的機率都取 log。並且為了避免 overflow 和乘到機率 0 的問題(在 log domain 中為加上 -inf)，當機率遇到 -inf 就直接跳過不做這一步，因為這一組加上 -inf 的機率會變成 -inf 最後也不影響答案，因此直接跳過即可。
2. 加快演算法的方法：
利用 forward-backward 演算法，如此可避免在指數複雜度下做 M-step 的運算。用此方法的複雜度為 $O(n \cdot h^2)$ (n 為文章長度， h 為 hidden unit 的個數)。
3. 因為 encode table 大部分值是 0，因此我用 vector list 把 encode table 和 decode table 存起來，因為每個單字的 decode 結果都知道，所以用 decode table 就可反推可能的 hidden unit，如此就不用遍歷所有的可能 unit。

參考資料:

1. 上課講義 L5_Efficient EM Training.pptx p. 4~P. 19
2. Karl Stratos, Unsupervised Learning 101: the EM for the HMM
3. Michael Collins, The EM algorithm for HMMs
4. Christopher M. Bishop. Pattern Recognition and Machine Learning (Information Science and Statistics). 2006. P. 618 ~ P. 621
5. 和 R05922027 江東峻 同學討論