COMS W4701 HW4

Dawei He (dh3027)

April 10, 2022

Problem 1

(a)

By conditioning, we have

$$Pr(X_t|x_k, x_j) = \frac{Pr(X_t, x_k, x_j)}{Pr(x_k, x_j)}$$

because the denominator $Pr(x_k, x_j)$ is constant, so we can say $Pr(X_t|x_k, x_j)$ is proportional to $Pr(X_t, x_k, x_j)$, which can be represent as

$$Pr(X_t|x_k, x_j) \propto Pr(X_t, x_k, x_j)$$

(b)

We need to sum over all x except x_t, x_k, x_j, i.e.,

$$Pr(x_t, x_k, x_j) = \sum_{x_i,\ i=0,...,t\ \&\ i\neq j,k,t} Pr(x_0, x_1, ..., x_t)$$

.

(c)

Without losing generalization,

$$\begin{aligned}
&Pr(x_t, x_k, x_j) \\
&= \sum_{x_i,\ i=0,...,t\ \&\ i\neq j,k,t} Pr(x_0, x_1, ..., x_t) \\
&= \sum_{x_i,\ i=0,...,t\ \&\ i\neq j,k,t} Pr(x_0) \prod_{i=1}^{t} Pr(x_i|x_{i-1}) \\
&= \sum_{x_i, i=k+1,...t-1} \ \sum_{x_i, i=j+1,...k-1} \ \sum_{x_i, i=0,...j-1} Pr(x_0) \prod_{i=1}^{t} Pr(x_i|x_{i-1})
\end{aligned}$$

.

Consider the product $Pr(x_0) \prod_{i=1}^{j} Pr(x_i|x_{i-1})$, which is equal to

$$Pr(x_0)Pr(x_1|x_0)Pr(x_2|x_1) \cdots Pr(x_j|x_{j-1}) = Pr(x_0, x_1)Pr(x_2|x_1) \cdots Pr(x_j|x_{j-1}),$$

since $Pr(x_i|x_{i-1}) = Pr(x_i|x_{i-1}, ..., x_0)$ in Markov Chain, then $Pr(x_2|x_1) = Pr(x_2|x_1, x_0)$ etc. The product can be written as

$$Pr(x_0)\prod_{i=1}^{j} Pr(x_i|x_{i-1}) = Pr(x_0, x_1..., x_j).$$

And $\sum_{x_i, i=0,...j-1}$ means summing the first j elements over $Pr(x_0, x_1..., x_j)$, then get the marginal distribution of x_j, i.e,

$$\sum_{x_i, i=0,...j-1} Pr(x_0)\prod_{i=1}^{j} Pr(x_i|x_{i-1}) = Pr(x_j).$$

Now consider the product

$$\prod_{i=j+1}^{k} Pr(x_i|x_{i-1}) = Pr(x_{j+1}|x_j)Pr(x_{j+2}|x_{j+1})\cdots Pr(x_k|x_{k-1}),$$

since

$$\begin{aligned}
&Pr(x_{j+1}|x_j)Pr(x_{j+2}|x_{j+1}) \\
&= \frac{Pr(x_j)Pr(x_{j+1}|x_j)Pr(x_{j+2}|x_{j+1})}{Pr(x_j)} \\
&= \frac{Pr(x_j)Pr(x_{j+1}|x_j)Pr(x_{j+2}|x_{j+1}, x_j)}{Pr(x_j)} \\
&= Pr(x_{j+1}, x_{j+2}|x_j),
\end{aligned}$$

the product can be reduced as $\prod_{i=j+1}^{k} Pr(x_i|x_{i-1}) = Pr(x_{j+1}, x_{j+2}, ..., x_k|x_j)$, $\sum_{x_i, i=j+1,...k-1}$ means summing the middle $k-j-1$ elements over $Pr(x_{j+1}, x_{j+2}, ..., x_k|x_j)$, then get the marginal distribution given x_j, i.e,

$$\sum_{x_i, i=j+1,...k-1} \prod_{i=j+1}^{k} Pr(x_i|x_{i-1}) = \sum_{x_i, i=j+1,...k-1} Pr(x_{j+1}, x_{j+2}, ..., x_k|x_j) = Pr(x_k|x_j).$$

Then consider the product

$$\prod_{i=k+1}^{t} Pr(x_i|x_{i-1}) = Pr(x_{k+1}|x_k)Pr(x_{k+2}|x_{k+1})\cdots Pr(x_t|x_{t-1}).$$

same as before, the product can be reduced as marginal distribution given x_t, $Pr(x_t|x_k)$, when summing over x_i from $i = k+1$ to $i = t-1$.
Hence the formula the products of the three parts:

$$\begin{aligned}
&Pr(x_t, x_k, x_j) \\
&= \sum_{x_i, i=k+1,...t-1} \sum_{x_i, i=j+1,...k-1} \sum_{x_i, i=0,...j-1} Pr(x_0)\prod_{i=1}^{t} Pr(x_i|x_{i-1}) \\
&= Pr(x_j)Pr(x_k|x_j)Pr(x_t|x_k).
\end{aligned}$$

(d)

According to part(c),

$$Pr(x_t|x_k, x_j) = \frac{Pr(x_t, x_k, x_j)}{Pr(x_k, x_j)} = \frac{Pr(x_j)Pr(x_k|x_j)Pr(x_t|x_k)}{Pr(x_k, x_j)} = \frac{Pr(x_k, x_j)Pr(x_t|x_k)}{Pr(x_k, x_j)} = Pr(x_t|x_k),$$

and formula in part(c) holds true for all values of x_t, which means $Pr(X_t|x_k, x_j) = Pr(X_t|x_k)$ is true. Hence, X_t is conditionally independent of X_j and X_k.

Problem 2

(a)

The stochastic matrix T is

$$T = \begin{bmatrix} 0 & 1/3 & 0 & 1/2 & 0 & 0 \\ 1/2 & 0 & 1 & 0 & 1/3 & 0 \\ 0 & 1/3 & 0 & 0 & 0 & 0 \\ 1/2 & 0 & 0 & 0 & 1/3 & 0 \\ 0 & 1/3 & 0 & 1/2 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1/3 & 0 \end{bmatrix}$$

We have to compute the eigenvector with corresponding eigenvalue being 1 and normalize this eigenvector so all elements sum up to 1. Using numpy to compute above procedures and we get the stationary distribution π over the robot's predicted location:

$$\pi = \begin{bmatrix} P(A) & P(B) & P(C) & P(D) & P(E) & P(F) \end{bmatrix} = \begin{bmatrix} 1/6 & 1/4 & 1/12 & 1/6 & 1/4 & 1/12 \end{bmatrix}$$

(b)

The beginning state is $P(X_0) = (0, 0, 0, 0, 1, 0)$, we can get the probability of state X_1 by calculating

$$P(X_1) = T \cdot P(X_0) = (0, 1/3, 0, 1/3, 0, 1/3)$$

now we have to compute the probability of state X_1 given e_1.

$$P(X_1|e_1) = \frac{P(e_1|X_1)P(X1)}{P(e_1)} = \frac{1}{1/2 \times 0 + 1/4 \times 1/3 + 1/2 \times 0 + 1/4 \times 1/3 + 0 \times 0 + 1/4 \times 1/3} \begin{bmatrix} 1/2 \times 0 \\ 1/4 \times 1/3 \\ 1/2 \times 0 \\ 1/4 \times 1/3 \\ 0 \times 0 \\ 1/4 \times 1/3 \end{bmatrix} = \begin{bmatrix} 0 \\ 1/3 \\ 0 \\ 1/3 \\ 0 \\ 1/3 \end{bmatrix}$$

(c)

$$P(X_2|e_1) = \alpha_2' = T \begin{bmatrix} 0 \\ 1/3 \\ 0 \\ 1/3 \\ 0 \\ 1/3 \end{bmatrix} = \begin{bmatrix} 5/18 \\ 0 \\ 1/9 \\ 0 \\ 11/18 \\ 0 \end{bmatrix}$$

$$\alpha_2 = O_2 \begin{bmatrix} 5/18 \\ 0 \\ 1/9 \\ 0 \\ 11/18 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \times 5/18 \\ 0 \times 0 \\ 1/4 \times 1/9 \\ 1/4 \times 0 \\ 1/4 \times 11/18 \\ 1/2 \times 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 1/36 \\ 0 \\ 11/72 \\ 0 \end{bmatrix}$$

$$P(X_2|e_1, e_2) = \frac{P(e_2|X_2, e_1)P(X_2|e_1)}{P(e_2|e_1)} = \frac{1}{1/36 + 11/72} \begin{bmatrix} 0 \\ 0 \\ 1/36 \\ 0 \\ 11/72 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 2/13 \\ 0 \\ 11/13 \\ 0 \end{bmatrix}$$

(d)

According to (b),

$$P(X_1, e_1) = P(e_1|X_1)P(X1) = \begin{bmatrix} 0 \\ 1/12 \\ 0 \\ 1/12 \\ 0 \\ 1/12 \end{bmatrix}$$

we can get calculate $P(X_1, X_2, e_1, e_2)$

$$\begin{aligned} P(X_1, X_2, e_1, e_2) &= P(e_2|X_2)P(X_2|X_1)P(X_1, e_1) \\ &= P(e_2|X_2) \begin{bmatrix} 0 & 1/36 & 0 & 1/24 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1/36 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1/36 & 0 & 1/24 & 0 & 1/12 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1/144 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1/144 & 0 & 1/96 & 0 & 1/48 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \end{aligned}$$

finally we can get $P(X_1, X_2|e_1, e_2)$

$$P(X_1, X_2|e_1, e_2) = \frac{P(X_1, X_2, e_1, e_2)}{P(e_1, e_2)} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2/13 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2/13 & 0 & 3/13 & 0 & 6/13 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

The most likely state sequence is $E \to F \to E$

Note: $P(X_2|X_1)P(X_1, e_1)$ is row element wise multiplication, $P(e_2|X_2) \cdot [P(X_2|X_1)P(X_1, e_1)]$ is column element wise multiplication.

(e)

$P(e_2|X_1)$ is β_1, we want to use backward algorithm to compute it. First we initialize

$$\beta_2 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

Then we can compute β_2'

$$\beta_2' = O_2\beta_3 = \begin{bmatrix} 0 \times 1 \\ 0 \times 1 \\ 1/4 \times 1 \\ 1/4 \times 1 \\ 1/4 \times 1 \\ 1/2 \times 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 1/4 \\ 1/4 \\ 1/4 \\ 1/2 \end{bmatrix}$$

We can get β_1

$$P(e_2|X_1) = \beta_1 = T^T\beta_2' = \begin{bmatrix} 1/8 \\ 1/6 \\ 0 \\ 1/8 \\ 1/4 \\ 1/4 \end{bmatrix}$$

The quantity means if we only know we are at state X_1 and don't know the observation e_1 at state X_1, what is the probability that we can observe # at state X_2.

(f)

$$P(X1, e_2|e_1) = P(X_1|e_1)P(e_2|X_1) = \begin{bmatrix} 1/8 \\ 1/6 \\ 0 \\ 1/8 \\ 1/4 \\ 1/4 \end{bmatrix} \times \begin{bmatrix} 0 \\ 1/3 \\ 0 \\ 1/3 \\ 0 \\ 1/3 \end{bmatrix} = \begin{bmatrix} 0 \\ 1/18 \\ 0 \\ 1/24 \\ 0 \\ 1/12 \end{bmatrix}$$

$$P(X1|e_2, e_1) = \frac{1}{1/18 + 1/24 + 1/12} \begin{bmatrix} 0 \\ 1/18 \\ 0 \\ 1/24 \\ 0 \\ 1/12 \end{bmatrix} = \begin{bmatrix} 0 \\ 4/13 \\ 0 \\ 3/13 \\ 0 \\ 6/13 \end{bmatrix}$$

According to (d)

$$P(X_1, X_2|e_1, e_2) = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2/13 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2/13 & 0 & 3/13 & 0 & 6/13 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

So we can marginalize X_2

$$P(X1|e_2, e_1) = \sum_{X_2} P(X_1, X_2|e_1, e_2) = \begin{bmatrix} 0 & 4/14 & 0 & 3/13 & 0 & 6/13 \end{bmatrix}$$

We can see two results of two methods are the same.

Problem 3

3.3

(a)

Train accuracy: 0.9555019712104628

Test accuracy: 0.8375544528500484

Because some words in testing dataset do not occur in the training dataset, the observation probability $P(e|x)$ will be given 1 for all POS state. So their state will be less accurate than those words observed in the training set. So the overall test accuracy will be less than training accuracy

(b)

There could be some small portion of invalid sentences in the training dataset, which means these sentence could have wrong sequence of POS. HMM will maximize the likelihood of POS for most of the correct sentences in the training set, and output their correct sequence of states. But if the sentence has wrong POS sequence, HMM can not output such wrong states by back propagation.

3.5

(c)

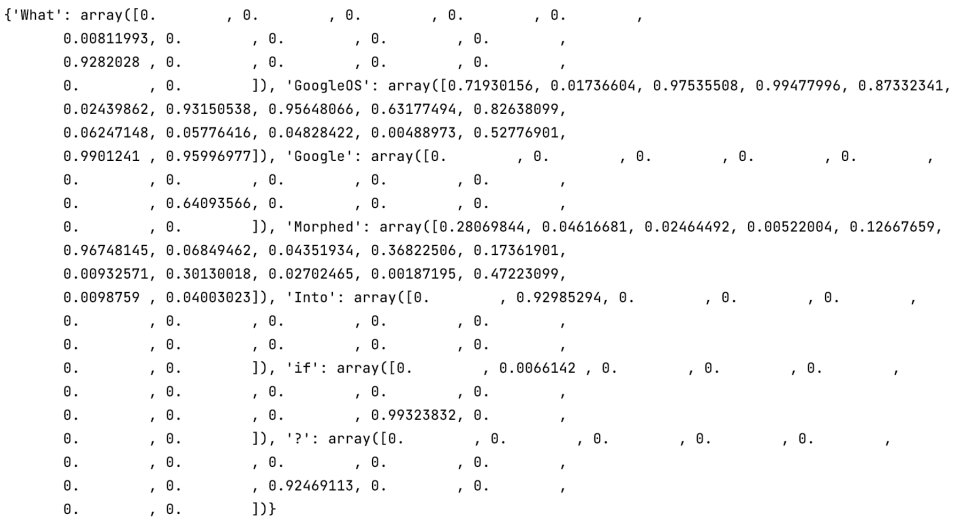

Figure 1: Observation probabilities for 3.5

$\arg\max_X P(e = What|X)$ =PRON, which is correct.
$\arg\max_X P(e = GoogleOS|X)$ =AUX, which is false.
$\arg\max_X P(e = Google|X)$ =PROPN, which is correct.
$\arg\max_X P(e = Morphed|X)$ =CCONJ, which is false.
$\arg\max_X P(e = Into|X)$ =ADP, which is correct.
$\arg\max_X P(e = if|X)$ =SCONJ, which is correct.
$\arg\max_X P(e =?|X)$ =PUNCT, which is correct.

POS of *GoogleOS* and *Morphed* are hard to identify and appear to be ambiguous. They both do not appear in the training dataset and *GoogleOS* is a compound word, so it is hard for model to correctly produce their POS.