NLP - EX 2

1) Prove that any generative probabilistic classification model that assume that
- the sequence of labels forms a Markov Chain
- the generation of x_i is conditionally independent given y_i of all other x_j and y_j

is an HMM model.

* Generative model: works with a joint probabilistic model of data. Assumes functional form for $\mathbb{P}(X|Y)$ and $\mathbb{P}(Y)$. Estimates parameters from data.

Prediction $$y^* = \arg\max_y \mathbb{P}(Y,X) \underset{\text{bayes rule}}{=} \arg\max_y \mathbb{P}(X|Y)\cdot\mathbb{P}(Y)$$

* Markov Chain: a finite state automaton with probabilistic state transitions.

Markov Assumption: next state only depends on the current state and independent of previous history given the current state.

A maximum likelihood estimator for a K^{th} order Markov language model is:

$$q_{ML}(y_m = t_j \mid y_{m-1} = t_{i1}, \dots, y_{m-k} = t_{ik}) = \frac{Count(t_j, t_{i1}, \dots, t_{ik})}{\sum_j Count(t_j, t_{i1}, \dots, t_{ik})}$$

transition probability (q_{ML}), label (y_m)

input sentence $X = x_1, x_2, \dots, x_m$ — עבור רצף של מילים x_i

tag sequence $y = y_1, y_2, \dots, y_{m+1}$ (y_{m+1} = STOP) — רצף של תגיות y_i

The most likely tag sequence for X is:

$$\arg\max_{y_1,\dots,y_{m+1}} \mathbb{P}(x_1,\dots,x_m, y_1,\dots,y_{m+1})$$

מודל גנרטיבי עם התפלגות משותפת $\mathbb{P}(x,y)$ כאשר x הם רצף של המילים ו-y התגיות המתאימות להן.

כמו כן, נתון כי $y_1,\dots,y_m$ יוצרות מודל מרקובי מסדר k.
עבור $\underline{k=1}$ $y_1,\dots,y_m$ יוצרות מודל מרקובי מסדר ראשון, כך ההתפלגות המשותפת של $y_1,\dots,y_m$ מתפרקת כמכפלה על פני פקטורים שתלויים רק בשכנים:

$$\mathbb{P}(y_1,\dots,y_m) = \prod_{i=1}^{m} \mathbb{P}(y_i \mid y_{i-1}) \quad \Leftarrow \text{Markov Chain}$$

$$= \prod_{i=1}^{m} \mathbb{P}(y_i \mid y_{i-1},\dots,y_0) \quad \text{transition probability}$$

במודל גנרטיבי אנו מסבירים כיצד לייצר משפט, ההסתברות נעשית על ידי כך שבהינתן אותו משפט, מהו הרצף הסביר ביותר של תגיות שייצר אותו.

$\mathbb{P}(X|Y)$ - לכל y_i יש סיכוי שונה עבור ההסתברות לראות את x_i בהינתן y_i

⊛ נוסף נתונה הנחה על פי הגלגל שסדרת x_i היוצרת את סדר המילים במשפט מהווים y_i

הנחה זו שמאוד לגיטימית היות שם מופע HMM הנשמעות היכן שסדר אותיות מאוד חשוב מסוים, אבל ישנם כאלו שסדר המילים במשפט לרוב יתן מופע נוסף על פיות הגלגל.

הנחה נוספת של מופע HMM היא שה-POS בוא משפיע רק על ה-POS הקודם, כלומר מופע מרקובי מסדר ראשון.

מטרה לקבוע מ $\mathbb{P}(x_1,\dots,x_m,y_1,\dots,y_m) = \left(\prod_{i=1}^{m+1} q(y_i|y_{i-1})\, e(x_i|y_i)\right) \cdot \mathbb{P}(y_1,\dots,y_m)$

$\mathbb{P}(x_1,\dots,x_m|y_1,\dots,y_m) = \left(\prod_{i=1}^{m+1} q(y_i|y_{i-1})\, e(x_i|y_i)\right)$

הסתברות הפלט

הסתברות המעברים

מופע מעבדי מסדר ראשון

רצפים שהופיעו ב training data

ועבור סדר א $\mathbb{P}(x_1,\dots,x_m,y_1,\dots,y_{m+1}) = \left(\prod_{i=1}^{m+1} q(y_i|y_{i-1},\dots,y_{i-k})\right)$

נשים ♡ כי

הגדרת זו תלות שם ⊛ הסתברות היציאה בין ליבון לגלגל

$\mathbb{P}(X|Y) = \frac{\mathbb{P}(X=x,\,Y=y)}{\mathbb{P}(Y=y)} \Rightarrow \prod_{i=1}^{m+1} e(x_i|y_i) = \mathbb{P}(y_1,\dots,y_m)\cdot \mathbb{P}(x_1,\dots,x_m|y_1,\dots,y_m)$

הגדרת זו תלות שניה הגדרת מרקובית

כאשר $y_{m+1} = y_{-1} = \dots = y_{-(k-2)} = *$ where * is START symbol

$y_{m+1} = \text{STOP}$

ומכלול מופע HMM שבידינו ניקח $x_1,\dots,x_m$ נרצה למצוא $y_1,\dots,y_m$ כך שיתקיים את ההסתברות המשותפת:

$$\arg\max_{y_1,\dots,y_m} \mathbb{P}(x_1,\dots,x_m,y_1,\dots,y_m)$$

זו ההסתברות הגבוהה שמתקבלת מעבר על כל הסדרות האפשריות של שרשרת מצבים $y_1,\dots,y_m$

2)

States = $\{H, L\}$

$\mathbb{P}(L|H) = 0.5$, $\mathbb{P}(H|L) = 0.4$

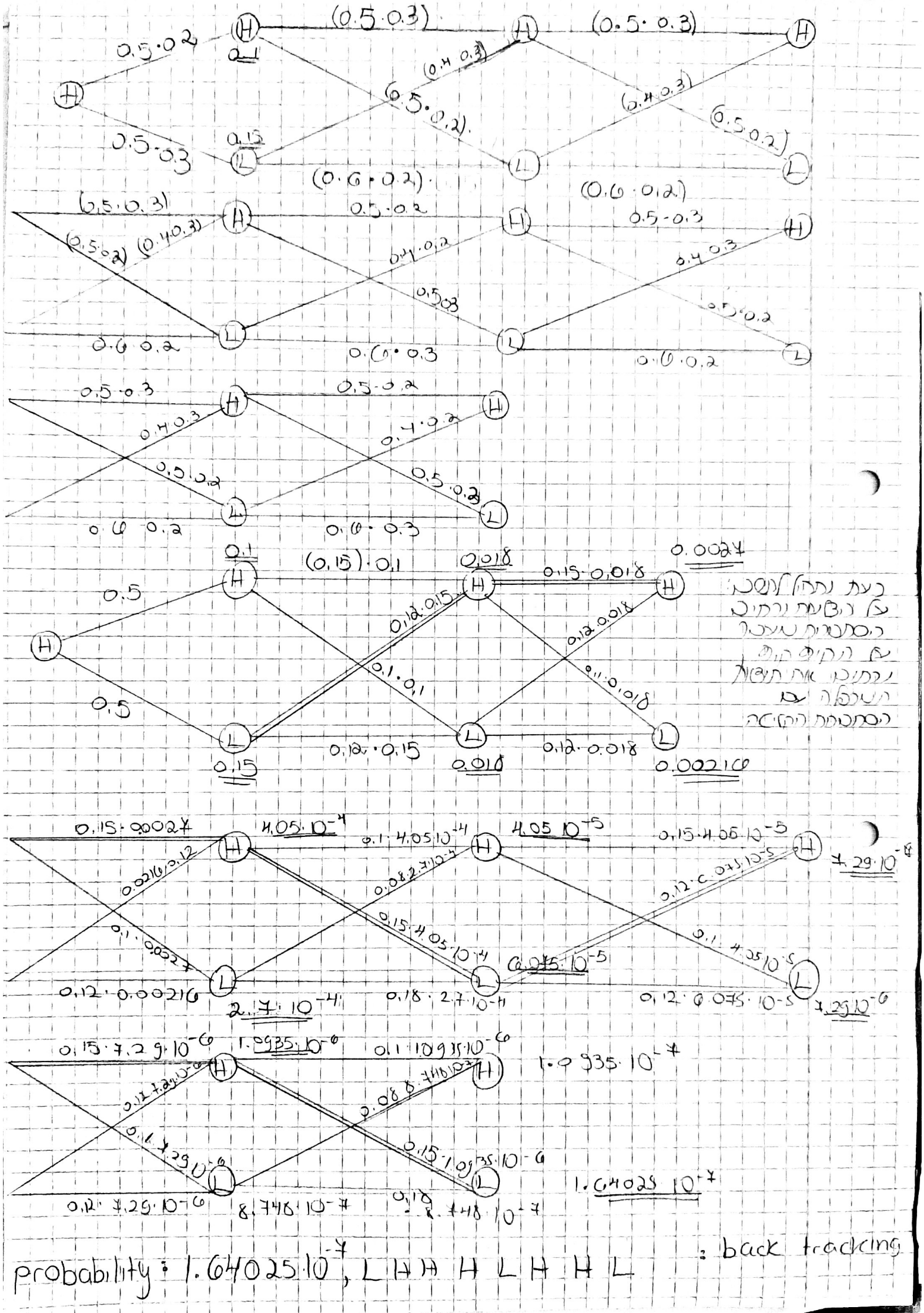
$\mathbb{P}(A|H) = 0.2$, $\mathbb{P}(C|H) = 0.3$, $\mathbb{P}(G|H) = 0.3$,

$\mathbb{P}(T|H) = 0.2$

$\mathbb{P}(A|L) = 0.3$, $\mathbb{P}(C|L) = 0.2$, $\mathbb{P}(G|L) = 0.2$, $\mathbb{P}(T|L) = 0.3$

S = A C C G H T G C A

previous state before S was H.


probability: 1.64025·10^-7, L H H H L H H L
: back tracking

3) Four-gram tagger:

$$\mathbb{P}(x_1,\dots,x_n,y_1,\dots,y_{n+1}) = \prod_{i=1}^{n} q(y_i \mid y_{i-3}, y_{i-2}, y_{i-1}) \prod_{i=1}^{n} e(x_i \mid y_i)$$

$y_0 = y_{-1} = y_{-2} = *$ where $*$ is the START symbol
$y_{n+1} = \text{STOP}$
$y_i \in K$ for $i=1,\dots,n$ where K is the set of all possible tags.

Viterbi algorithm that takes as input an integer n and finds

$$\max_{y_1,\dots,y_{n+1},\, x_1,\dots x_n} \mathbb{P}(x_1,\dots,x_n,y_1,\dots,y_{n+1})$$

Define a dynamic programming table

$\pi(k,t,u,v)$ = maximum probability of a tag sequence ending in tags t,u,v at position k

Input: An integer n (not a sentence $x_1 \cdots x_n$),
$q(w \mid t,u,v)$ - transition probabilities,
$e(x \mid s)$ - emission probabilities

Definition: Define K to be the set of all possible tags.
Define $K_{-2} = K_{-1} = K_0 = \{*\}$, $K_k = K$ for $k=1,\dots,n$
Define $\mathcal{V}$ to be the set of possible words

In generative model we generate a sentence, and predict the most probable sequence of tags. In this case the output is the probability.

Initialization: Set $\pi(0,*,*,*) = 1$

Algorithm: For $k=1,\dots,n$

For $t \in K_{k-2}$, $u \in K_{k-1}$, $v \in K_k$

$$\pi(k,t,u,v) = \max_{x_k \in \mathcal{V},\, w \in K_{k-3}} \{\pi(k-1,w,t,u) \cdot q(v \mid w,t,u) \cdot e(x_k \mid v)\}$$

Return $\max_{t \in K_{n-2}, u \in K_{n-1}, v \in K_n} \{\pi(n,t,u,v) \cdot q(\text{STOP} \mid t,u,v)\}$

We are not given a sentence $x_1,\dots,x_n$, so we should find x_k from the vocabulary

Running Time: $n \cdot |K|^3$ entries in π to be filled in (for $t,u,v \in K$)
$O(|K|\,|\mathcal{V}|)$ time to fill one entry (for $w \in K_{k-3}$, for $x_k \in \mathcal{V}$)
$\Rightarrow$ time to calculate $q(v \mid w,t,u) \cdot e(x \mid w)$ for all k,w,t,u,v is $O(n \cdot |K|^4 |\mathcal{V}|)$

	Most likely tag baseline	HMM	Add-one smoothing	Pseudo Words	Pseudo Words & Add-one smoothing
Known Words Error Rate	6.10%	23.00%	17.60%	21.90%	19.40%
Unknown Words Error Rate	73.20%	78.30%	72.70%	52.60%	55.50%
Total Error Rate:	11.90%	27.90%	22.40%	24.60%	22.50%