This list contains exercises of the type you will find in an exam for the course *Natuurlijke Taalmodellen en Interfaces*.

Contents

Points

Question:	1	2	3	4	5	Total
Points:	2	1	4	3	3	13

1 Markov models

1. Consider the probability of a sentence as given by the following factorisation

$$\begin{aligned} P_S(x_1^n) &= P_N(n) P_{S|N}(x_1^n|n) \\ &= P_N(n) \prod_{i=1}^{n} P_{X|H}(x_i|x_{<i}) \end{aligned}$$

where S is a random sentence, N a random length, X a random word, and H a random history.

(a) (½ point) Select appropriate descriptions for x_1^n

- ◯ an outcome of S
- ◯ a sequence of n random words
- ◯ n outcomes of S

(b) (½ point) Select appropriate descriptions for n

- ◯ a random length
- ◯ a random noun
- ◯ the length of the outcome of S

(c) (½ point) Select appropriate descriptions for x_i

- ◯ a random word
- ◯ the ith element of the outcome of S
- ◯ the ith random sequence

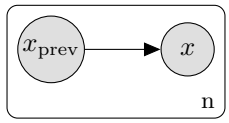
(d) (½ point) Select appropriate descriptions for $x_{<i}$

- ◯ a word if $i = 2$
- ◯ a random sequence
- ◯ the ith random history

Total for Question 1: 2

2. (1 point) Let x_1^n be the outcome of a random sentence S, and let $P_{S|N}(x_1^n|n)$ denote its probability value (given length n) under a **unigram** language model. Write down the expression that corresponds to this probability value.

3. Answer questions about the graphical model below, where X is a random variable over exactly v English words.



(a) (½ point) Which language model (LM) is this?
A. unigram LM B. bigram LM C. hidden Markov LM

(b) (½ point) How many conditional probability distributions (cpds) are there in the model (ignore the *length* distribution)?
A. one B. two C. n D. v

(c) (½ point) Is $P_{X|X_{\text{prev}}=x_{\text{prev}}}$ a tabular cpd or an inferred distribution?
A. tabular B. inferred

(d) (½ point) Is $P_{S|N=n}$ a tabular cpd or an inferred distribution?
A. tabular B. inferred

(e) (1 point) Write down the expression of the probability value $P_S(x_1^n)$ (you may assume appropriate padding exists).

(f) (½ point) Assume that the probability value $P_{X|X_{\text{prev}}}(x|x_{\text{prev}})$ can be assessed in constant time. Express the complexity of computing $P_{S|n}(x_1^n|n)$ as a function of sentence length (use *big-O-notation*).

(g) (1/2 point) Suppose we have exactly v words in the vocabulary, and we use a Categorical distribution for each cpd in the model. What is the representation cost of this model (use *big-O-notation*)?

Total for Question 3: 4

4. Consider the following unigram language model, where EoS is a special symbol deterministically added to the end of every sentence, and answer the questions below. In this exercise you are

X	$\text{Cat}(x\|\boldsymbol{\theta})$
a	θ_{a}
b	θ_{b}
c	θ_{c}
d	θ_{d}
EoS	θ_{EoS}

expected to pad sentences with a BoS token, which **is not** modelled, and an EoS token, which **is** modelled.

(a) (½ point) What is the probability of the sentence <u>a b c a d</u> given its length?

(b) (½ point) What is the probability of the sentence <u>a b b d c a a f</u>?

(c) (1 point) What is the role of smoothing?

(d) (1 point) Answer true (T) or false (F).

i. ____ The sentence <u>a a b c</u> has the same probability as the sentence <u>a b a c</u>.

ii. ____ The unigram language model is sensitive to word order.

iii. ____ A smoothed unigram language model has infinite support.

iv. ____ Without smoothing, and without taking padding into account, the support of the unigram language model above is the set of strings in $\{\text{a}, \text{b}, \text{c}, \text{d}\}^*$.

Total for Question 4: 3

5. Consider the generative story below

$$N \sim P_N$$
$$X_i | X_{i-1} = x_{i-1} \sim \text{Cat}(\theta_1^{(x_{i-1})}, \ldots, \theta_v^{(x_{i-1})}) \qquad \text{for } i = 1, \ldots, n$$

and assume that every distribution in the model gives support to $\{\text{a}, \text{b}, \text{c}, \text{d}, \text{UNK}\}$, where UNK is a special token to which we map all unseen words, in addition to a EOS padding symbol that occurs always at the end of strings.

(a) (1 point) Draw the graphical model using plate notation.

(b) Pad the sentence <u>a b c a b</u> and answer the questions below.

i. (1 point) List its bigrams and their counts.

ii. (1 point) What is the probability of the sentence given its length? Express probability values in terms of the parameters shown in the generative story.

Total for Question 5: 3

Assessment

Question	Points	Score
1	2	
2	1	
3	4	
4	3	
5	3	
Total:	13	