Prof. Dr. Andreas Hotho,
M.Sc. Janna Omeliyanenko
Lecture Chair X for Data Science, Universität Würzburg

3. Exercise for "Sprachverarbeitung und Text Mining"

19.11.2021

1 Knowledge Questions

1. What is the difference between first-order Markov models and k-order Markov models?
2. What motivates the usage of order-k Markov models?
3. What is the problem concerning the data, when Markov models of too high order k are applied? What method did you learn in the lecture to address this problem? What is the intuition behind this method?
4. Is the application of the Viterbi-algorithm to Markov models of higher order more computationally expensive than to lower order models for the same POS-tag set with size T? Justify your answer.
5. Why must the condition $\lambda_3+\lambda_2+\lambda_1 = 1$ be satisfied for the following interpolation from the lecture: $P(t_i|t_{i-1}, t_{i-2}) = \lambda_3\hat{P}(t_i|t_{i-1}, t_{i-2}) + \lambda_2\hat{P}(t_i|t_{i-1}) + \lambda_1\hat{P}(t_i)$?
6. In the lecture we have transformed our optimization problem of finding the most probable sequence of POS-tags given an observed word sequence into the logarithmic space, forming the sum of logarithms of probabilities instead of the product over probabilities. Due to the monotonicity property of the logarithmic function, the maximum after transformation into the logarithmic space does not change. Given that a function $f : A \to \mathbb{R}$, where A is a subset of $\mathbb{R}$, is called strictly monotonically increasing if for all $(x, y) \in D$ with $x < y$ it holds that $f(x) < f(y)$, prove that the logarithm function is strictly monotonically increasing on the interval $(0, inf)$.
7. Why does the integration of a classifier for the estimation of the node score make for a more potent Markov model?

8. Given are the following table with the features, the counts observed in the dataset, and the counts predicted by two different trained classifiers CI and CII. Which of the two classifiers better reflects the data? Justify your answer.

Feature-Name = Value $\wedge$ CurrentLabel	Observed counts	CI predicted counts	CII predicted counts
Current Word = $w_1 \wedge$ ADJ	15	14.3	12.7
Current Word = $w_1 \wedge$ N	37	32.8	32.2
Current Word = $w_1 \wedge$ V	123	118.9	121
Current Word = $w_2 \wedge$ ADJ	43	38	38
Current Word = $w_2 \wedge$ N	1225	1222	1122
Current Word = $w_2 \wedge$ V	0	0.3	11
Previous Word = $w_1 \wedge$ ADJ	246	233.7	2463
Previous Word = $w_1 \wedge$ N	1	0.9	111
Previous Word = $w_1 \wedge$ N	587	587	7.2

2 POS Tagging - but backwards

Once again, consider the following highly simplified set of generalized word types (tagset) for part-of-speech tagging, where words are modeled as observations and associated word types as hidden variables:

DET	Determiner	the,a,...
N	Noun	year,home,costs,time,...
PRO	Pronoun	he,their,you
V	Verb	said,took,saw

Once again, using the table below, calculate the most likely sequence of tags for the following sentence:
„`I saw the saw`"

1. Apply the Viterbi algorithm. This time however, apply the algorithm backwards, so that the starting state is located at the end of the sequence and transition probabilities are applied from the current to the previous observation instead of the next.
2. Which changing behavior do you observe in comparison to the last exercise sheet?

Assume uniformly distributed <u>start</u> transition probabilities.

from\to	DET	N	PRO	V
DET	0.05	0.9	0.05	0
N	0.1	0.7	0.05	0.15
PRO	0.05	0.5	0.05	0.4
V	0.5	0.1	0.4	0

w\t	DET	N	PRO	V
I	0	0.0002	0.1	0
saw	0	0.0008	0	0.1
the	0.1	0	0	0