DATA.STAT.840 Statistical Methods for Text Data Analysis

Exercises for Lectures 8-9: HMMs continued, and Probabilistic Context-free Grammars

Exercise 8.1: HMM modeling of sentences.

The file "hmm_sentences.txt" provided with this exercise pack contains 1000 sentences, one on each line, which have been generated from a relatively simple probabilistic context free grammar with a limited vocabulary. Some example sentences are:

- i will explain .
- they are cautious , however , i am strong and then the cautious robot will feed me , but you will feed a small fox and he is strong , however , is the fox strong ?
- they are wise , but then a insightful strong cat is cautious , however , are you cautious ?
- where will you explain ?
- the bird is insightful .

These sentences contain statements and questions which can be concatenated together, and each statement/question can contain different kinds of subjects and possibly also objects. In this exercise, the idea is to see how well such simplified sentences (which do not represent the full variety of language) can be modeled by a hidden Markov model.

Use the **HMMlearn** Python library, or another library of your choice in your chosen programming language, to learn a hidden Markov model for this set of example sentences. Use either 5 or 10 hidden states. Inspect the resulting emission probabilities of the states, and the transition matrices between the states. Do the states seem to correspond to meaningful properties of the simplified language?

Report your code, the resulting emission and transition probabilities, and your analysis.

Exercise 8.2: grammar for equations.

While equations are not natural language, they are featured in many scientific publications and are a useful example for grammars. Propose a context-free grammar that can create equations of the kind used in simple calculators: nonnegative integer numbers, a limited set of variables (x, y, z), operators (+, -, *, /, ^), parentheses, a limited set of functions (log, exp, sin, cos, tan). Example equations that your grammar should be able to generate:

- 1+2*3^10
- 24-(3*x-(z^y))/(3-x)
- sin(2*x-cos(y))^(2-10)

Then, show how an example equation is derived from your grammar: show the derivation tree of applying rules to arrive at the final equation.

Note: in this exercise you do not need to assign probabilities to the rules of your grammar.

Report the definition of your grammar, and the derivation of your example equation.

(exercises continue on the next page)

Exercise 8.3: Chomsky normal form.
The probabilistic context free grammar below is a simplified version of the one used to generate "hmm_sentences.txt" in exercise 8.1. Transform the grammar into Chomsky normal form, so that the possible sentences and their probabilities are the same as in the grammar below. Report the definition of your resulting Chomsky normal form grammar (rules and their probabilities).

S --> STMANY	1.0
STMANY --> S1 .	0.6
STMANY --> S1 , but STMANY	0.4
S1 --> SUBJ QVERB1 QVERB2 OBJ	1.0
SUBJ --> ARTICLE DESC NOUN	1.0
DESC --> ADJECTIVE	0.7
DESC --> ADJECTIVE DESC	0.3
OBJ --> ARTICLE DESC NOUN	1.0
QVERB1 --> can	0.2
QVERB1 --> will	0.5
QVERB1 --> may	0.3
ARTICLE --> a	0.6
ARTICLE --> the	0.4
QVERB2 --> explain	0.4
QVERB2 --> help	0.2
QVERB2 --> answer	0.4
ADJECTIVE --> wise	0.3
ADJECTIVE --> friendly	0.5
ADJECTIVE --> insightful	0.2
NOUN --> cat	0.7
NOUN --> dog	0.2
NOUN --> fox	0.1

Exercise 9.1: Inside-outside algorithm.
Use the inside-outside algorithm for the Chomsky normal form grammar you produced in exercise 8.3, to calculate the probability of the sentence "a wise fox can help the friendly insightful cat". Report your computation and the resulting probability. Hint: the resulting probability should be the same as it is in the original grammar given in exercise 8.3.

Exercise 9.2: Can you explain the Inside-outside algorithm better than an LLM?
Use a large language model (AI chatbot; see previous exercises for examples of what could be used) to generate an explanation of how the inside-outside algorithm works in PCFGs. Then modify the LLM's answer to add new facts or details the LLM did not state and to fix errors or other issues in the LLM's answer.

In your solution, report what LLM you used, report the original LLM answer, and report your modified answer with the added and modified parts highlighted.