Supervision questions: set3.

Markov assumption

State the two Markov assumptions, and explain why they are important in the definition of Hidden Markov Models.

HMM Artificial data

The data you were given with task 7 (parallel sequence of observations and states created by the "dice" HMM) was artificially created using an HMM (remember that we called HMMs and Naive Bayes **generative models**). In this exercise, you will explore how this was done.

1. What is the information you need in order to be able to design an algorithm for generating artifical data using an HMM?
2. Describe an algorithm for creating artificial data.
3. Transition probabilities into the final state are expressed as an extra parameter for an HMM. In some models these final transition probabilities are irrelevant. Under what circumstances would the prediction result be affected by transitions into the final state? Can you think of some examples of real world situations where this might happen?

Smoothing in HMMs

We did not smooth the Dice HMM in task 7 nor did you smooth the protein HMM in task 9.

1. In which situations can smoothing be counterproductive, and why?
2. In the case of the protein model, which of the two types of probability are better candidates for smoothing and why?

Viterbi and Forward algorithm

Study the Forward algorithm in the Jurafsky and Martin textbook. This is the algorithm for estimating the likelihood of an observation. It is another instance of the dynamic programming paradigm.

1. Give and explain the recursive formula for this dynamic programming algorithm in terms of a_{ij} and $b_i(o_t)$.
2. Explain why there is a summation over the paths.

Parts of Speech tagging with HMM.

Hidden Markov Models (HMM) can be used for **Part of Speech Tagging**. This is the task of assigning parts of speech, such as **verb, noun, pronoun, determiner** to words in a text sample.

A particular HMM is defined as follows: $S_e = \{s_1 = \text{verb};\ s_2 = \text{noun},\ s_3 = \text{personal pronoun},\ s_4 = \text{auxiliary verb}\}$; s_0, s_F designated start state and end state

$$A = \begin{bmatrix} a_{01} = 0.01 & a_{02} = 0.10 & a_{03} = 0.60 & a_{04} = 0.29 & \\ a_{11} = 0.02 & a_{12} = 0.63 & a_{13} = 0.07 & a_{14} = 0.13 & a_{1f} = 0.15 \\ a_{21} = 0.49 & a_{22} = 0.20 & a_{23} = 0.10 & a_{24} = 0.01 & a_{2f} = 0.20 \\ a_{31} = 0.40 & a_{32} = 0.05 & a_{33} = 0.05 & a_{34} = 0.40 & a_{3f} = 0.10 \\ a_{41} = 0.73 & a_{42} = 0.01 & a_{43} = 0.15 & a_{44} = 0.01 & a_{4f} = 0.10 \end{bmatrix}$$

$$B = \begin{bmatrix} b_1(fish) = 0.89 & b_2(fish) = 0.75 & b_3(fish) = 0 & b_4(fish) = 0 \\ b_1(can) = 0.10 & b_2(can) = 0.24 & b_3(can) = 0 & b_4(can) = 1 \\ b_1(we) = 0.01 & b_2(we) = 0.01 & b_3(we) = 1 & b_4(we) = 0 \end{bmatrix}$$

The observation sequence O is the following: $O =$ We can fish

1. Consider the two state sequences $X_a = s_0, s_3, s_4, s_1, s_f$ and $X_b = s_0, s_3, s_1, s_2, s_f$. Which interpretations of the above observation sequence do they represent?
2. Give the probabilities $P(X_a)$, $P(X_b)$, $P(X_a, O)$ and $P(X_b, O)$, and state which of these probabilities are used in the HMM.
3. Demonstrate the use of the Viterbi algorithm for deriving the most probable sequence of parts of speech given O above. Explain your notation and intermediate results.
4. Does the model arrive at the correct disambiguation? If so, how does it achieve this? If not, what could you change so that it does?
5. If a labelled sample of text is available, then the emission probability matrix B can be estimated from a labelled sample of text. Describe one way how this can be done.
6. The statistical laws of language imply that there is a potential problem when training emission probabilities for words. This problem manifests itself in the probability of the word *can* in the state sequence X_b from (a) above. What is the problem, and how could it be fixed?

Viterbi with higher order HMMs.

Viterbi is a clever algorithm that allows you to process the input in time that is linear to the observation sequence. With a first order HMM, we keep N (number of states) maximum probabilities per observation at each step.

1. How many states do we need to keep for an N order HMM?
2. What are the implications for the asymptotic complexity of Viterbi?