8 Machine Learning and Real-world Data (sht25)

You want to determine which exact crops were grown on a particular field in mediaeval times in each year. You have records of the overall yield of the field (classified into good, average and poor), but the records don't say which crop was grown. You know empirically that certain crops tend to yield more than others:

Rye (R)	good: 50%, average: 40%, poor: 10%
Beans (B)	good: 20%, average: 30%, poor: 50%
Clover (C)	good: 10%, average: 60%, poor: 30%

An historical document provides a sample of consecutive years' crops, which shows that the villagers did not keep to a strict crop rotation:

R C C R B C B B R C C C B B R C B R C B B C C C R

You want to apply a Hidden Markov Model (HMM) to the task of predicting the crop sequences for years outside of your sample.

(*a*) Define the components of an appropriate First-Order Hidden Markov Model. Estimate the missing parameters from the information given. You may assume that each of the crops is equally likely to start a sequence. Apply smoothing. Ignore the end state (treat the sequence as if it ran forever). [6 marks]

(*b*) The Viterbi algorithm can be applied to infer a sequence of crops given an observation sequence.

(*i*) State the purpose of the variable $\delta_j(t)$ in the Viterbi algorithm and give its defining equation.

(*ii*) Consider the partial observation sequence `good, good, average, ...`, with the HMM trained as above. At t=2, the following δ have been calculated: $\delta_R(1) = \frac{1}{6}, \delta_B(1) = \frac{1}{15}, \delta_C(1) = \frac{1}{30}, \delta_R(2) = \frac{2}{11 \cdot 15}, \delta_B(2) = \frac{1}{8 \cdot 15}, \delta_C(2) = \frac{1}{8 \cdot 12}$. Simulate the Viterbi algorithm at t=3, i.e., the point when `average` is encountered, showing intermediate results. [6 marks]

(*c*) Which assumptions does an HMM make? To which degree are these assumptions justified in the situation described above? [4 marks]

(*d*) In order to mitigate the effects of the potentially violated assumptions mentioned in Part (*c*), somebody suggests an increase in the order of the HMM. Do you think this would have the desired effect, and why (or why not)? [2 marks]

(*e*) You were instructed to smooth the HMM in Part (*b*) above. There is also an argument for not applying smoothing. What would happen if the estimates above were not smoothed, and what is potentially desirable about this? [2 marks]