Data analysis and stochastic models

November 10, 2023

Duration: 1h30

> Materials and course notes authorized. Calculators allowed.
>
> Exercises are independent. Most questions within exercises can be considered independent. The number of points per exercise is given as an indication and may be modified later, within reasonable limits.
>
> When answering questions, don't limit yourself to a simple result. Give details of your calculations and explain your reasoning.

Exercise 1 - Hidden Markov Chains (10 points)

> For this exercise, use the simplified logarithm table given in the appendix.

We're interested in an athlete's state of fitness over time. He/she may be resting (R), resuming activity (A), in shape (F), or in super shape (S). Measurements always start when the athlete is resuming activity (A), and the probabilities of moving from one state to another are as follows:

	to R	*to A*	*to F*	*to S*
R	0.3	0.7	0.0	0.0
A	0.0	0.3	0.7	0.0
F	0.0	0.0	0.8	0.2
S	0.9	0.0	0.0	0.1

.

Questions

1. Draw the associated automaton.
2. What is the probability of the state sequence $A\ F\ S\ S\ S$?

During training, the athlete's times can be a, b, or c, and depend on the state of form as follows:

$P(time|form) =$

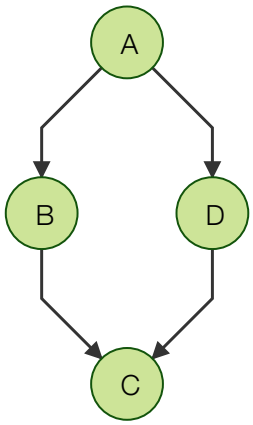
	to a	*to b*	*to c*
R	0.9	0.1	0.0
A	0.2	0.8	0.0
F	0.1	0.5	0.4
S	0.1	0.3	0.6

.

Questions

- 3. Using the Viterbi algorithm, give the most likely sequence of states for the sequence of observations $a\ b\ b\ c\ a$?
- 4. Some form states have typical durations (e.g. no more than a few weeks for S) which are not taken into account by current modeling. Explain how conditional random field (CRF) modeling would improve this fact.

Exercise 2 - Bayesian networks (6 points)



Consider the Bayesian network shown above. The variables are binary and the model is defined by the following parameters:

$P[A=0]=0.3$ $P[B=0 \mid A]=$

A	
0	0.5
1	0.7

$P[D=0 \mid A]=$

A	
0	0.2
1	0.6

$P[C=0 \mid B,D]=$

B	D	
0	0	0.3
0	1	0.1
1	0	0.8
1	1	0.5

.

Questions

1. Give the factorized form of the joint probability $P[A,B,C,D]$.
2. Draw the corresponding factor model.
3. Calculate $P[C=1|A=0,D=0]$.
4. Indicate two variables that are independent conditionally on another variable and prove this independence by factoring the joint probability $P[A,B,C,D]$.

Exercise 3 - Hypothesis testing (4 points)

You want to compare your algorithm (Our) with the state-of-the-art one (SOTA) in terms of computation time. You run them in 5 different contexts and obtain calculation times (in seconds) :

Setting	*S*1	*S*2	*S*3	*S*4	*S*5
Our	1	1	3	2	3
SOTA	3	1	2	3	6

Questions

1. Is the computation time of Our different from the computation time of SOTA with a confidence level of 5%?

Recall that for a random variable X following a Student's distribution with k degrees of freedom, the quantile of order 0.95 is :

- 2.353 for $k = 3$,
- 2.132 for $k = 4$,
- 2.015 for $k = 5$,
- 1.943 for $k = 6$,
- 1.895 for $k = 7$.

Appendix

x	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$ln(x)$	-2.3	-1.6	-1.2	-0.9	-0.7	-0.5	-0.4	-0.2	-0.1