Analyse de données et modèles stochastiques

November 16, 2018

Duration: 2 hours

Lecture handouts and notes authorized. Calculators are authorized.

Exercises are independent. Question within exercises can mostly be considered as independent. The number of points in front of each exercise is given as an indication and is subject to changes within reasonable bounds.

When answering the questions, do not limit yourself to reporting just a number or a trivial answer. Rather provide the details of your calculations and of your reasoning. What I care about is evaluating what you understood from the lectures, not your calculation skills!

Exercise 1 – Estimation (6 points)

The two questions are independent.

2.1 Empirical mean estimators

We want to study the mean of a population from a sample of $n = 2m$ independent measures, $X_i = x_i$, $i \in [1, n]$, sorted in ascending order. We have at our disposal two estimators:

$$E_1 = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad E_2 = \frac{X_m + X_{m+1}}{2} \ .$$

Discuss the respective qualities of these estimators, in particular in terms of bias and variance, comparing the two. Which one would you choose for noisy data exhibiting a significant amount of outliers?

2.2 Moment estimation

We now assume that on the n samples, we measure a value e_1 for E_1. We also measure s_1 as the value for the empirical estimator of the standard deviation given by

$$S_1^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - e_1)^2 \ .$$

We assume that the distribution of the random variables X_i is a Gaussian of mean μ and standard deviation σ. Applying the moment estimation technique, give the estimate of μ and σ as a function of e_1 and s_1.

Exercise 2 – Sequence labeling with hidden Markov models (8 points)

You might use the simplified logarithm table in Appendix for numerical applications.

Labeling sequences of observations is a very common problem for which hidden Markov models are well-suited (e.g., finding the syntactic tag for a word in natural language processing, the letter corresponding to a portion of an image in optical character recognition). The tags to be found are considered as the hidden states which have to be found from the observations (words in case of syntactic tags, portion of images in OCR, etc.)

In this exercise, we consider such a model with four tags: A, B, C and D. We assume that the sequence of tags is a Markov process of order 1 with the following transition probabilities

	A	B	C	D
A	0.2	0.6	0.1	0.1
B	0.5	0.0	0.1	0.4
C	0.6	0.1	0.2	0.1
D	0.4	0.1	0.3	0.2

We will take as initial probabilities $P[\text{A}] = P[\text{B}] = P[\text{D}] = 1/3$.

1. What is the probability of the sequence of tags D C A B A?
2. What is the sequence of tags corresponding to the observation

   `a b c d b`

   assuming the following association probabilities between states and tags

	a	b	c	d
A	0.5	0	0	0.5
B	0.6	0.4	0	0
C	0.3	0.4	0.3	0
D	0	0	0	1

   Please explain your calculation to get to the result (a result with no explanation will be considered as invalid).

We now are interested in estimating the parameters of the model, i.e., the transition probabilities A_{ij} with $(i, j) \in \{A, B, C, D\}^2$ and state conditionnal probabilities B_{ik} with $i \in \{A, B, C, D\}$ and $j \in \{a, b, c, d\}$, from a set of training samples.

3. Assuming we know the complete data, i.e., a set of sequences with their observation as illustrated below, how would you estimate the parameters? Provide the exact equations, introducing the necessary notations.
4. Using the equations from the previous question, give the estimation of the quantities $P[x|B]$ for all $x \in \{a, b, c, d, e\}$, $P[D|D]$ and $P[A|B]$ according to the following samples

a	b	c	d
D	B	B	C

e	c	b	a	d	a
D	C	A	C	B	C

e	c	a	b	d	d	a	e
D	A	D	D	A	B	D	A

5. Explain in plain (and simple) words how the EM algorithm circumvents the fact that the hidden state sequence is unknown. In the example above, $P[a|A] = 0$ and $P[D|C] = 0$. Would the EM algorithm prevent these probabilities from being null assuming the observation sequences from the previous question (i.e., not knowing the hidden variables)?

Exercise 3 – Hypothesis testing (6 points)

We have designed a new sort algorithm that is faster than the traditionnal quicksort. To verify this claim, we test our algorithm on a number of platforms and architectures, running first the standard quicksort then our approach on the same data. We observe the following execution times (in seconds) on 9 distinct hardware:

qsort()	9.22	9.38	12.30	9.77	10.70	13.00	9.44	12.70	7.08
mysort()	8.10	10.74	11.456	8.50	10.98	11.5	10.85	11.19	6.13

Can we legitimately claim, with a risk of error of 5 %, that our algorithm is faster than quicksort?

We recall that for a random variable X following a Student distribution with 8 degress of freedom, we have $P[X \leq 1.86] = 0.95$.

A 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