COMPUTER SCIENCE TRIPOS Part IA – 2024 – Paper 3

7 Machine Learning and Real-world Data (sht25)

A language school for English receives students from different countries and with different skill levels. Before entering, students perform an English test, which decides which class they are assigned to (B1, B2 or I). After studying for a week, students are sometimes reassigned to a different level better reflecting their actual language ability.

(*a*) Professor M is unhappy with this process and the test's ability to predict student level. She suggests that the school should derive students' final level directly by machine learning, based on the students' age, their first language (L1), and how long they studied English before. Several years' data from previous students is available. Describe how the task could be accomplished using a Naive Bayesian Classifier. Apply smoothing if this is appropriate. Give all relevant formulae for parameter estimation and classification. [4 marks]

(*b*) Calculate all relevant probabilities for features age, L1 and experience, for your classifier defined in a), using the following sample of student data. . [6 marks]

	Test performance										Student Stats			Level	
ID	Total	Question												Assignment	
	Score	1	2	3	4	5	6	7	8	9	Age	L1	Exper.	Initial	Final
A	4		●	●			●			●	[21-24]	C	[5-7]	B2	B2
B	5		●		●	●	●		●		[13-16]	F	[1-4]	B2	B2
C	1		●								[21-24]	F	[1-4]	B1	I
D	3				●	●				●	[17-20]	C	[5-7]	B1	B2
E	7	●	●	●		●		●	●	●	[13-16]	C	[5-7]	I	B2
F	5	●	●			●	●	●	●●		[21-24]	F	[≥ 8]	B2	I
G	6	●	●		●		●		●●		[17-20]	C	[5-7]	I	I
H	2		●						●		[17-20]	C	[1-4]	B1	B1
I	8	●	●	●	●	●			●	●	[13-16]	F	[≥ 8]	I	I
J	5	●	●			●	●		●		[21-24]	F	[1-4]	B2	B1

(*c*) A new student enters the school whose L1 is "C", who has studied English for 5 years, and who is 18 years old. Which level will this student be assigned to by your classifier from a), as trained in b), and why? [2 marks]

(*d*) Some features influence the prediction of the classifier more than others. How could you use the data available to you to determine the relative relevance of individual features? Describe at least two methods and give a numerical illustration for at least one of your methods, using the above data. [4 marks]

(*e*) The school are now re-thinking how they create classes. Rather than relying on level descriptions such as B1, they want to define classes by grouping incoming students of similar ability into classes of roughly the same size. Describe a method how this could this be achieved. You may use all information in the table above, excluding the information on levels. [4 marks]