Intelligent Systems: Reasoning and Recognition

James L. Crowley

Bayes Rule and Data Mining with EM

You are responsible for an International Masters program. Over the last 5 years, your program has accepted 200 students from three Universities. You wish to use the academic results of these students in two of your courses in order to provide guidelines for admissions for future students. For each student, you have the name of his university, his ranking in the last year of studies at his home university, and the grades that he has obtained in the two reference classes in your program, noted as numeric grade from 0 to 20.

a) Explain how to use a ratio of histograms to estimate the origin of a student from his grades in your reference classes. Present the formula and explain its terms. How large are the histograms? How many students from each University are necessary in order for a ratio of histograms to give a reasonable result? How can you estimate the probability of error?

b) Bad Luck! There are not enough students to use a ratio of Histograms. Explain how to use a normal (Gaussian) probability density functions to estimate the probability that a student is from one of the three Universities given his notes in your two reference classes. Explain how to estimate the parameters for the normal density functions. Can you determine a probability of error for the origin of the students? If yes, how is it determined?

c) You have really bad luck! Your secretary has erased the information about the origin of your students. All you have are their notes in your two reference classes. Explain how to use the EM algorithm to estimate the probability that each student belongs to each of the three universities. Explain how to initialize the parameters for the algorithm.