All homework problems are mandatory. You are allowed to work in groups to understand the problem and find a solution, as long as you list the group members, but you have to write the solution individually. Make sure you show your work not only the answer to the questions.

1. **Multi-attribute discrete splits.**

Classic classification trees allow splits on a single attribute. A natural extension is to consider splits involving multiple variables. In this problem we consider such splits for discrete attributes.

(a) Consider the problem of finding splits on two attributes simultaneously where the split is specified by an unrestricted partitioning of the cross product of the domains of the two attributes in two parts. Explain how the partitioning the minimizes the gini gain can be found efficiently for this situation.

(b) Consider the same problem as before but, in order to reduce the amount of space required to record the partitioning, we consider the partitioning of the domains of the two attributes independently in two parts, which would result in a four region partitioning of the crossproduct of the domains of the two attributes.

- Your job is to come up with convincing arguments that this problem is NP-complete. As a starting point you can take a look at the paper “Processing Complex Aggregate Queries over Data Streams” which you can find on my webpage at: http://www.cise.ufl.edu/ adobra/papers.html
- Can you come up with an heuristic that is better that first splitting on one attribute then on the other?
2. **Multi-attribute continuous splits.**

In this problem we consider the counterpart of the previous problem for continuous attributes. In this case, a fairly general partitioning can be specified by a hyperplane in the space of the two (or more) variables. This type of splits are called *oblique* splits.

(a) Find at least two papers that describe methods to efficiently find oblique splits and briefly describe them. Make sure you compare the methods you describe (scalability, effectiveness).

(b) Look carefully at the experimental result in the paper and identify in what practical situations oblique splits are preferable to normal splits. Back up your conclusions with concrete examples from papers.