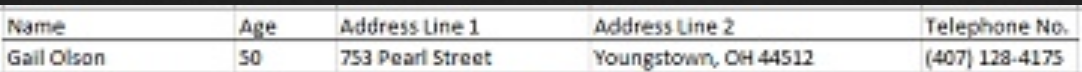
Reading Questions for Chapter 4 of Mitzenmacher-Upfal

When data comes to us in a spreadsheet, the rows are individuals and the columns are different variables/features we are measuring on each individual, e.g. in the following row



we have 5 features: Name, Age, …, Phone Number.

1. Give an example of a real world scenario where you might want to divide a dataset into two disjoint groups (as in 4.4) where each feature is balanced between the two groups.
2. Compute for the matrix A and the vector b below:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | 1 | 0 | 1 | 0 | 0 | 0 | 1 |
| 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 |
| 0 | 1 | 1 | 0 | 0 | 1 | 1 | 1 |

|  |
| --- |
| 1 |
| -1 |
| -1 |
| 1 |
| 1 |
| 1 |
| -1 |
| 1 |

1. Find a vector b that you think *minimizes* for the matrix above.
2. Can you think of any non-brute force way to search for b? If not, then perhaps the problem is NP complete. If you’ve learned about complexity theory, comment on this.
3. What’s the asymptotic lower bound on ? How far is the book’s algorithm off from optimal? Hint: the answer is in the reading, right before Theorem 4.11.
4. The proof of Theorem 4.11 breaks into two cases: k small vs. k large. Intuitively, keeping in mind that the goal is to find b to minimize , why is the k small case easy?
5. Consider the proof of Theorem 4.11, where the book says “clearly.” Recalling that the entries of b are all 1 or -1, what vector b maximizes |ci|? What is that maximum value of |ci| as a function of k? Is that line in the proof clear now?
6. For the k large case, we write the bad event as a union of n events. Write out the bad event as a sentence:
7. Are the n individual events mutually disjoint? Does this matter for the union bound?
8. How did the book get rid of the m and k in the last inequality of the proof?