**Assignment 3: Data Lab, Naive Bayes Classifier**

* Due by 30th September, 2022 by 5pm IST.
* To be submitted to the following email address: office.of.gr@gmail.com
* The subject of the email should be: Assignment Number [3]: Data Lab, 2022
* Please clearly mention your name and roll number.
* Submit your work as a single pdf file. Additional material,

code, etc can/should also be submitted, but there should be atleast 1 pdf, which has the

entire assignment.

* Wherever there is code, in the assignments, the code should be

well documented and easy to understand / follow.

The objective of the assignments is three fold. One is to be able to develop expertise in writing and communicating about technical topics. This will be done by using the IEEE conference style format for all assignments. The other is to explain, in your own way, the mathematical ideas that are embedded within the technical topic of interest. For example, in this case it is naive bayes classifier. The third is to use the topic, in this case of naive bayes classifier, to understand a problem from the real world. So in a sense the objective is to write what one may call a mathematical essay on naive bayes classifier.

Title could be: Assignment 3: a mathematical essay on naive bayes classifier.

Abstract: Give a brief overview of your assignment.

Author: Name, Department, Institution, Email

**Section 1: Introduction**

In this section, the 1st paragraph should be on a broad overview of the topic. The 2nd paragraph should be an overview of the technical aspects (i.e. in this case it is a naive bayes classifier). The 3rd paragraph should be about the problem that you are aiming to solve/understand using naive bayes classifier. Finally, the 4th paragraph should give an overview of the paper.

**Section 2: Naive Bayes Classifier**

This section should outline the key principles underlying naive bayes classifier.

**Section 3: Data**

This data was extracted from the 1994 Census bureau database by Ronny Kohavi and Barry Becker (Data Mining and Visualization, Silicon Graphics). *The key task is to determine whether a person makes over $50K a year*, adult.csv contains the dataset required to solve the task.

| **Variable** | **Definition** | **Key** |
| --- | --- | --- |
| age | Age | Continuous |
| workclass | Work class | Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked |
| fnlwgt |  | Continuous |
| education | Level of education | Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool |
| education-num | No. of years of education | Continuous |
| marital-status | Marital status | Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse |
| Occupation | Occupation | Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces |
| relationship | Relationship | Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried |
| race | Race | White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black |
| sex | Gender | Female, Male |
| capital-gain | Capital gain | Continuous |
| capital-loss | Capital loss | Continuous |
| hours-per-week | Working hours / week | continuous |
| native-country | Native Country | United-States, Cambodia, England . . . |

**Section 4: The problem**

(a) Outline the problem, and plot/visualize the data.

(b) Make progress on the problem, by applying the techniques of naive bayes classifier to the problem at hand.

(c) Discuss any insights and observations.

**Section 5: Conclusions**

Write about 1 paragraph on the key insights that were obtained from your study and also outline any further avenues for investigation.

**References**

Please put in all the references that you have used during the assignment. The format should be the same as the IEEE conference format.