UiT THE ARCTIC UNIVERSITY OF NORWAY

Assignment 2. Due: Sunday 08.10.2023 23:59

FYS-2021 Exercises
Department of Physics and Technology
Faculty of Science and Technology

Before you start

Learning to write a scientific report is an important skill that many of the courses at the Faculty of Science and Technology, including this one, aim to improve. Therefore any question that you answer should be contained within the report of this assignment. Answers outside of the written report, (e.g. in the comment of the code, or within a Jupyter Notebook), will not be considered as a part of your answer of the problem. You can structure your report by having a separate (sub)section with the answer for each question. The report and code should be your own individual work. Remember to cite all sources.

Make sure your report shows that you understand what you are doing. More specifically, it is important to elaborate your answers such that essential theory, equations, and intuition is included in your answers. However, your answers should still remain concise and stay focused on the core problem, e.g. there is no need to derive or prove an equation unless the problem asks you to.

Problems that ask for numeric values or plots should include these in the answer of the report. The code should be commented in such a way that any person with programming knowledge should be able to understand how the program works. Like your report, the code must be your own individual work.

You are permitted to use standard built-in functions and/or packages (e.g. numpy, pandas and matplotlib in Python) for reading the data and basic calculations. However: make sure that the packages you use do not over simplify your implementation! Of course, all implementations asked for in the problems should be your own work.

Hand-in format

For your final submission, ensure the report is consolidated into a single `.pdf` file, with any associated code appended at the end. Alongside this report, submit a compressed `.zip` file that encompasses both the `.pdf` report and the code files, adhering to the directory structure below. The naming convention for these files should be `assignment2_firstname_lastname.pdf` and `assignment2_firstname_lastname.zip` respectively.

```
assignment2_firstname_lastname.zip
|
|-- report
|     |
|     |-- assignment2_firstname_lastname.pdf
|
|-- src
      |
      |-- Your code goes in this directory
```

Submission

Please upload BOTH your `.pdf` and `.zip` files to Canvas under 'Assignments' > 'Mandatory assignment 2' by the announced submission due.

Plagiarism

Plagiarism is a serious academic offense and will not be tolerated. Always ensure that you provide proper attribution for any work, ideas, or concepts that are not originally yours. Should the Canvas plagiarism detection system flag your submission, your report will not be considered for assessment.

Resources

All datasets required to answer the exercises can be found in the Canvas room for the course.

Problem 1

(1a) Data is appended as `data_problem1.csv`. Load the data and report general information of the data. Additionally plot (as histograms) the data and discuss the separability.

(1b) Lets assume that data from class $\mathcal{C}_0$ follows a Gamma distribution:

$$p(x|\mathcal{C}_0) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-\frac{x}{\beta}}$$

and that data from class $\mathcal{C}_1$ follows a Gaussian distribution:

$$p(x|\mathcal{C}_1) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

The Gamma distribution has the parameters α and β, in this case α is known: $\alpha = 2$, but β is unknown. The Gaussian distribution has the parameters μ and σ where both are unknown.

Remember that the Gamma *function* for $n \in \{0, 1, 2, ...\}$ can be computed as:

$$\Gamma(n) = (n-1)!$$

Show that that the maximum likelihood estimations of the parameters are:

$$\hat{\beta} = \frac{1}{n_0\alpha}\sum_{j=1}^{n_0} x_0^j, \quad \hat{\mu} = \frac{1}{n_1}\sum_{j=1}^{n_1} x_1^j, \quad \hat{\sigma}^2 = \frac{1}{n_1}\sum_{j=1}^{n_1}(x_1^j - \hat{\mu})^2$$

where $x_0^1, ..., x_0^{n_0}$ are the training samples from $\mathcal{C}_0$ and $x_1^1, ..., x_1^{n_1}$ are the training samples from $\mathcal{C}_1$.

(1c) Split the data into training and test data. Use the maximum likelihood estimations to estimate the parameters based on the training data. Use the point-estimations of the parameters to implement a Bayes' classifier. Report the test accuracy.

(1d) Explain why the Bayes' classifier minimizes the probability of miss-classification when the probability distribution of the data is known.

Show the missclassified data in a plot along with the rest of the data and explain why it was miss-classified. Does it follow the conclusion in *a*)?

Problem 2

(2a) An important concept for the decision tree is entropy defined by the equation:

$$H(p) = -\sum_{i=1}^{n} p_i \cdot \log_2(p_i) \tag{1}$$

Explain what entropy is, plot the entropy H for a 2-class problem, as a function of p_1 where $p = (p_1, p_2)$, $p_2 = 1 - p_1$, in the interval $p_1 \in [0, 1]$. What are p_1 and p_2 for our classification case? Elaborate on how is this used in a decision tree classifier.

(2b) We will be working with a movie dataset containing 5000 movies from IMDB. The dataset contains several features, however for this classification task we will be working with the following features: *Year*, *Rating*, *Runtime(Mins)* and *Total_Gross*. Keep also the title of the movie, you will need it for the last question.

The task will be to predict if a movie is good or bad, where the good movies has a rating above 6,5. Perform the following:

1. Extract the relevant features mentioned above from `imdb_5000_preprocessed.csv`.
2. Create labels from the *Rating* column (*Rating* >= 6.5: 1, *Rating* < 6.5: 0)
3. Split the dataset into a training and test set. Use 20% of the samples for the test set. Make sure to retain the distributions of each class the same in the training and test set!

(2c) Implement your own decision tree with a maximum depth of 2. Train it using the training set, report the accuracy and confusion matrix on both the training and test set. Discuss the results. Remark: it is not mandatory to use a recursion if you are not comfortable with it, you may use loops.

(2d) (BONUS) Implement the decision tree using recursion and make a tree of depth 4. Compare (shortly) the results with the ones of the previous question.

(2e) What is the best feature for the classification task and does it do a good job? List a few films that were not classified correctly. Why do you think they were not correctly classified?