Assignment 4

Objective of this assignment is to design multivariate classifiers from first principles. You may choose to either extend the univariate methods from assignment 1, for the multivariate scenarios or design new approaches. However, do not use any off the shelf classification algorithms.

Please note that this assignment does not require you to submit the code. However, you may choose to implement your methods to explore concepts.

The task of your classifier is gender identification, based on measured parameters. We have a toy data set of 1000 male and 1000 female (labeled) samples. We attempt a height based univariate classifier and realize that due to overlap in heights, there are limitation on improving accuracy meaningfully without exploring other features. So, we decide to identify and add new features to our learning algorithms to reduce the prediction error further.

Consider following scenarios of increasing complexity.

1. Uncorrelated input features (5 marks)
    a. We have two input features say, Height(in cm) and Heamoglobin levels measured for all 2000 samples. Let's assume that both these features are normally distributed within each gender. These features are pretty much uncorrelated within each gender.
    b. Can you design approaches to train a classification algorithms to predict gender?
2. Input features with non-zero correlation (3 marks)
    a. In this scenario, we have two input features say, Height(in cm) and weight(in kg) measured for all 2000 samples. Both are normally distributed within each gender. The correlation between these features is 0.6 within each gender.
    b. Which of the algorithms you designed for uncorrelated features would work as is ? If they don't, what changes can you make to your algorithms to accommodate correlations
3. How far can we go? (2 marks)
    a. We observe that accuracy improves with addition of one new feature in both of the above scenario. Can we reach a conclusion that accuracy can be improved further by adding multiple such features to the input? How many such features would you add in

your quest to improve accuracy? Would addition of new features require any changes to the experimental set up?

Guidelines:

1. Don't use any GenAI tools to find solutions.
2. Document the algorithms in your own language with all the steps clearly mentioned.
3. Highlight your conclusions (including limitations and risks if there any for each algorithm).
4. Submit the above in PDF document by Monday 17th Feb 2025 by 10:00 am.