Learnings from Assignment 1

This document captures the objectives of the assignment 1. It is divided in two parts.

The part one, titled as *observations* is what students are expected to come-across while experimenting with parameter values. Though they are not expected to know the concepts formally including their names, the phenomenon should be evident in the output in some form.

The part two is titled as *further investigations.* And students are not expected to conduct this investigation as first part of their assignment. However, the challenges in the assignment, throws open multiple questions. Investigating these challenges, presents an opportunity to introduce and develop foundational theory of machine learning. Students are encouraged to explore the concepts highlighted in ***green font***

Observations

1. Univariate Classification is feasible even with preliminary methods when the classes have small overlap.
2. The misclassification rate is proportional to size of the overlap across the three methods.
3. Size of the overlap in turn depends on distance between means of the two genders and their respective standard deviations
    a. As the distance between the means reduces, the overlap between the two distributions and hence the misclassification rate increases
    b. As the standard deviations of the two genders increases, the overlap between the two distributions and hence the misclassification rate increases
4. Accuray of both the threshold-based method and local/binning/quantization based method is impacted not just by the features of data such as means and standard deviations but also by the parameters we use in algorithms to find the optimal solution.

a. In threshold-based method, the value by which we increment the threshold impacts both the misclassification rate and optimal threshold value.
b. In quantization-based method, the interval length used for binning impacts the misclassification rate.
c. If you observed this in your experiment, Congratulations!!! You have stumbled upon an important concept in ML. Such algorithm specific parameters are called ***hyper-parameters*** (which are different from the ***model parameters***) and the process of finding their optimal value is called ***hyperparameter tuning***.

5. For threshold-based method –
    a. As we reduce the increment size, the misclassification rate reduces initially as the increment size drops. And new optimal thresholds appear. However, the progress saturates soon.
    b. Please observe this in the following table

increment_size	optimal threshold	misclassificaiton_rate(%]
5	161	11.15
2	159	10.25
1	159	10.25
0.5	159.5	10.15
0.1	158.8	10.15
0.05	158.55	10.1
0.01	158.55	10.1
0.005	158.545	10.05
0.001	158.543	10.05
0.0001	158.5425	10.05

    c. As the threshold increments become too small, though the time algorithm takes to find optimal solution increases drastically, the gain in accuracy start to be too small to justify the time/computational overhead.
6. For Quantization based method:
    a. This approach has capability to reduce misclassification to almost zero.
    b. The so-called accuracy keeps on improving as the bin size reduces. At the same time, the number of points within each bin also drop significantly as seen in below table

bin_size	Misclassificaion_rate
1	10.25
0.5	10.15
0.1	9.2
0.05	7.85
0.01	3.75
0.001	0.7
0.0001	0
1.00E-05	0
1.00E-06	0

c.

i. For a very small bin size, majority of the bins have no data point. So, there is no way to predict a gender if new data points fall into one of these now empty bins
ii. The minority of the bins which have data, have mostly single data point or very small number of data points. This significantly affects the reliability of predictions in these bins.
iii. Thus, while the accuracy improves, model turns to be completely unreliable.
iv. If you observed this irony, it is your first encounter with ***overfitting*** – where the model starts to hallucinate that it has learned everything. Yet all it has done is- just memorised the labels at some locations. Though it has achieved very high accuracy in the data on which it is trained, it is bound to perform terribly on yet unseen data points. Thus, it fails to ***generalise***.

Further investigation:

1. As is observed in Threshold based classifier and Quantization based classifier, it is futile to chase accuracy forever. Beyond a point, either the gains are not worth the computational effort or the increased accuracy is absolutely useless. This warrants investigation into following questions
    a. How do we decide at what point should we stop chasing accuracy while training ML models?
        i. Concepts developed to address these concerns are collectively called as ***regularisation methods***

b. Can it be done with just one dataset that we use for training or do we need additional test dataset to check if the claimed accuracy on training data set holds true on unseen data.
    i. Does this hint at using a ***test-train split*** of the data to ensure model generalization?
        1. Use the train part of the data to train model but measure and report accuracy by evaluating the model on separate test data.
c. If we try to tune the hyper-parameter by testing their accuracy on the test data, is the so-called test data really the 'test data' since it was used to tune/learn a hyper-parameter? Does this warrant the need for more than two-way test-train split?
    i. Does this present an opportunity to split the train data and use a part of this split as a ***validation set*** to tune the hyper-parameters? Can we reduce the risk further by using multiple such splits of the train data---- ***Cross-validation***?