SYDE 572 – Assignment 3

Calvin Tran – 20826392

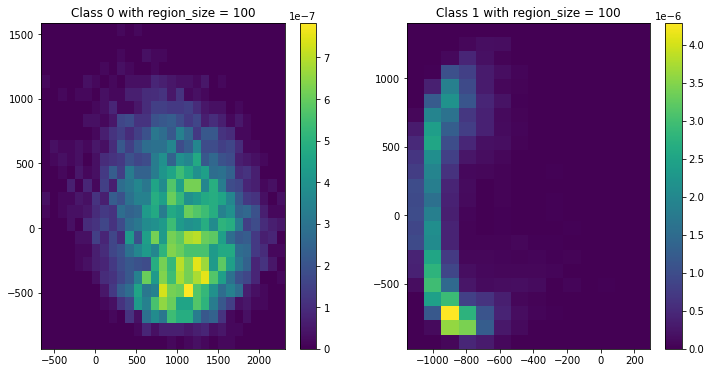
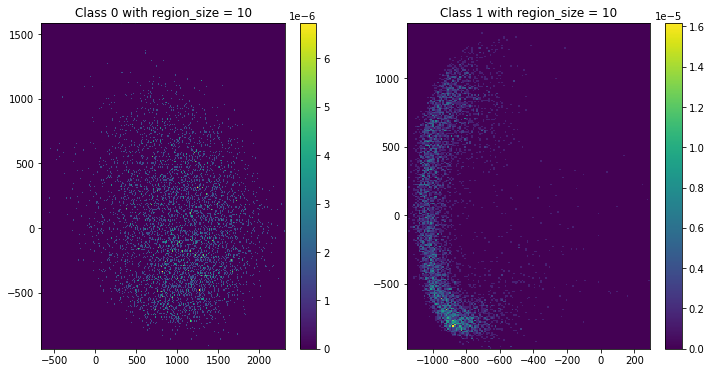
All code for each exercise can be found in Assignment\_3.ipynb.

# Exercise 1

The plots for each class and region size can be seen below.

A screenshot of a screen

Description automatically generated



A region size of 100 seems to be the best at capturing patterns in the data. Due to its high granularity, both region sizes of 1 and 10 have histograms that are too precise, making them extremely sensitive to noise and outliers. By having a larger region size and thus a larger sample size per bucket, the histogram becomes less sensitive to noise and outliers. Thus, the region size of 100 is the best at highlighting patterns in the data.

The following test error percentages were reported:

The percent error for a region size of 1 is 53.24%.

The percent error for a region size of 10 is 25.25%.

The percent error for a region size of 100 is 0.38%.

The percent error for a region size of 100 is significantly lower than the other two region sizes, making it the best option.

The graph below shows the graph using a kernel-based density estimation.

A collage of images of a comet

Description automatically generated

The percent error for this classifier was 0.28%, which was lower than all of the other histogram-based estimations.

The kernel-based estimation seems to best represent the data due to it having the lowest error.

While the histogram-based estimation with a region size of 100 also had a very low error rate, it also assumes that is constant throughout the region. With a larger region size, this becomes less true. Meanwhile, the kernel density estimation is able to create smoother visualizations, with points affecting an area around it, rather than just a single point. Although it may be more computationally intensive, the kernel-based estimation with a bandwidth of 20 performs better than the histogram-based estimation with a region size of 100.

Parametric methods rely on the assumption that data adheres to a particular distribution, such as Gaussian or exponential. This assumption only works well if the data actually adheres to the specified distribution. On the other hand, non-parametric methods, such as the ones above, refrain from making the same distribution constraints, allowing them to accommodate data exhibiting various patterns.

When the data's distribution is well-known and matches the assumptions, parametric methods can be more appropriate. However, when the data's distribution is uncertain or doesn't conform to parametric assumptions, non-parametric methods offer an attractive alternative. In this case, since we were given no information on the data’s distribution, the non-parametric approach for density estimation may be the better option. Nonetheless, it may be beneficial to perform experiments using both methods and comparing the results.

# Exercise 2