COMP300027 Machine Learning - Project 1 (Music Genre Classification with Naïve Bayes) Report

Sammi Li 1271851

TASK 1: Pop vs. classical music classification

The model trained by the “pop_vs_classical_train.csv” dataset, evaluated on the “pop_vs_classical_test.csv" has an accuracy of 0.97674, precision of 0.95238, and the recall of 1.0.

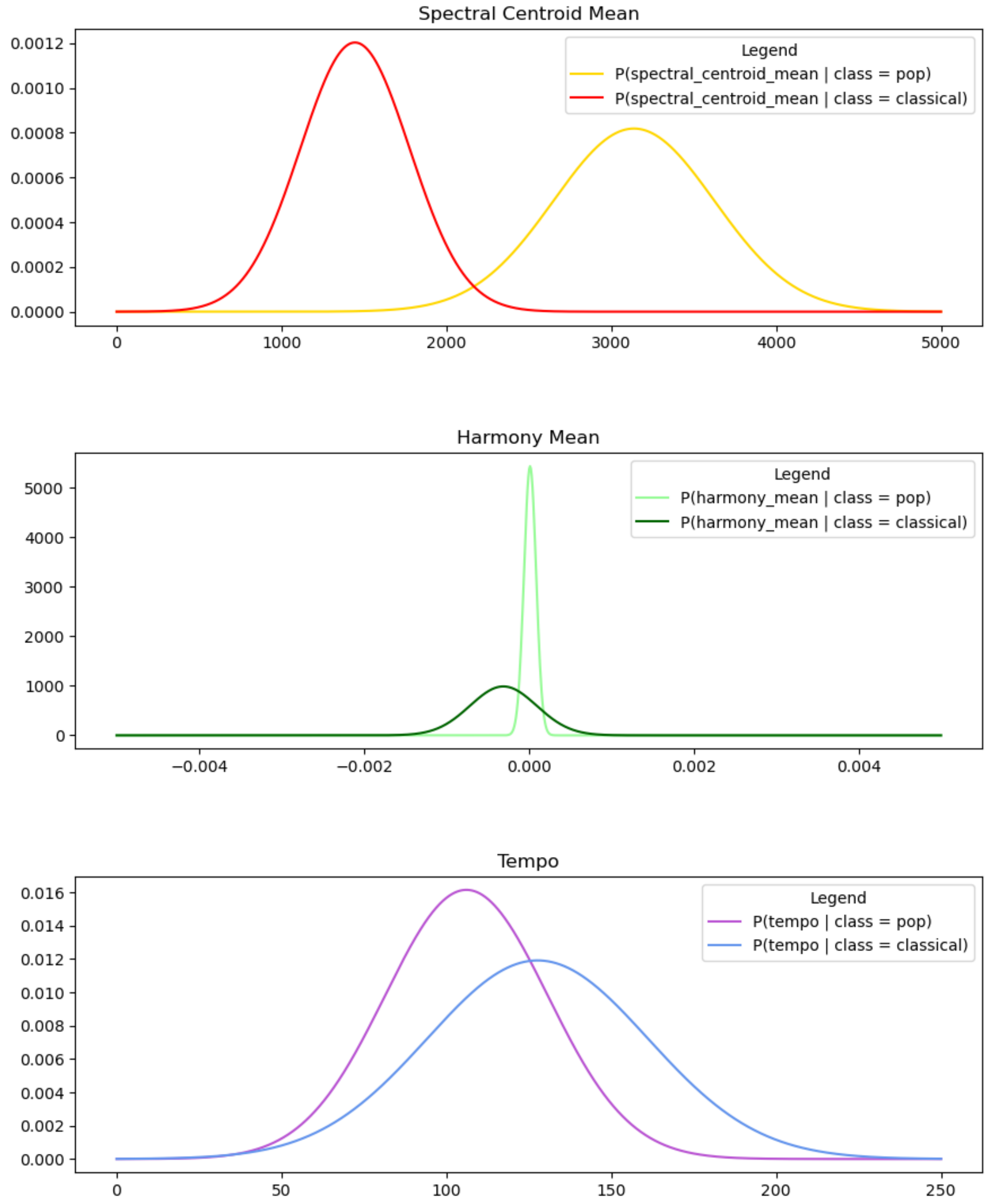

Figure 1: Probability density functions for features, spectral centroid mean, harmony mean and tempo.

If I had to classify pop vs. classical music using just one of these three features, I would use the spectral centroid mean. This is because as shown in the plot, there is little overlap between the spectral centroid mean of the pop genre and the classical genre. The pop genre has spectral centroid mean values of approximately 1900 to 4600, whilst the classical genre has values of approximately 500 to 2400. Thus it would be easier to differentiate between classical and pop genres with these values. In comparison to the probability density function of harmony mean, which suggests that the harmony mean of both pop and classical genres highly overlap and are dense in a small range, indicating that the harmony means of pop music fit in the range of the harmony means of classical music. Additionally, the distribution of tempo also has a lot of overlap. Therefore, it will be difficult to differentiate between classical and pop genres using tempo and harmony mean.

TASK 2: 10-way music genre classification

In this segment, I modified my naive Bayes model to handle missing attributes and tested it on various test sets with different amounts of missing data. As shown in figure 1, initially, with no missing attribute, the model achieves an accuracy of 0.495. Then, a random attribute was removed from the first 50 instances, first 100 instances, and all 200 instances, which resulted in accuracy values of 0.455, 0.355, and 0.95 respectively. Columns were also removed from the test set, varying from two missing columns to seven. The accuracy for all these tests is 0.95. Therefore, from these values computed, it suggests that as the number of missing instances increases, the accuracy of the Naive Bayes model decreases. Although, the plateau into an accuracy of 0.95 after there is one missing value for each instance, may suggest that with the limited attributes, the model is able to predict at least 9.5% of the instances accurately.

Amount of missing attributes	Accuracy
Zero missing attributes	49.5%
1 missing attribute for the first 50 instances	45.5%
1 missing attribute for the first 100 instances	35.5%
1 missing attribute for the all 200 instances	9.5%
1 missing attribute for each instance + 2 missing columns	9.5%
1 missing attribute for each instance + 7 missing columns	9.5%
1 missing attribute for each instance + 7 missing columns + extra missing chunks	9.5%

Figure 2: Table of accuracies according to number of missing attributes