COMP90049: Assignment 2 Report

Anonymous

1 Introduction

1.1 Hypotheses

This report focuses on two hypotheses:

1. Training a model using both metadata features and audio features would perform better than training a model with just metadata or audio features.

2. Bootstrap aggregation would aid in the performance of the classifier.

1.2 Motivation

For the first hypothesis, after reading that combining the audio and lyrical features helped with the effectiveness of a classifier on genre classification [6], I wanted to see if this applied to metadata and audio. For the second hypothesis, as bagging does not appear have a large negative effect and increases accuracy on stable classification methods as seen in Breiman (1996), I wanted to see if this applied to this dataset when it came to genre classification.

2 Related Literature

2.1 The Million Song Dataset (MSD)

The source describes an attempt by researchers to provide a large dataset that comprises of metadata and audio features for a million different songs for algorithm development at scale. [1] The dataset contains audio features and metadata such as tags, similarity relationships and acoustic features and could be used in metadata analysis, artist recognition, recommendations and other uses. The main limitation of MSD is that the dataset lacks diversity due to not having much ethnic or classical music.

2.2 Capturing the Temporal Domain in echonest features for improved classification effectiveness

This paper evaluates the performance of the descriptors used in the MSD. The evaluation is based on 4 traditional MIR genre classification tests: GTZAN [9], ISMIR Genre and Rhythm [3] and the Latin Music Database [5]. The classifiers used are KNN, Support vector machines, J48, random forest and naive bayes. The findings of the paper were that there were good results with simple but short feature sets and even better results with complex feature sets. [8]

2.3 Music Genre Classification with the Million Song Dataset

This paper focuses on datamining music information on the MSD. Unlike most genre classification research, the authors focused on both musical features and lyrical features. A framework that uses model-blending combining text and audio sequence features was proposed. The findings were that timbre features from MDS helped with prediction, bagging of words with lyrical features also led to good predictions, especially for metal and hip-hop music. Finally, combining audio and lyrical features helped aid prediction. [6]

2.4 The Elements of Statistical Learning: Data Mining, Inference, and Prediction.

The source goes into depth about the various properties about decision trees and gaussian naive-bayes. The source describes classification trees as decision trees that predict outcomes where the predicted outcome is a discrete value. Within the trees, the "leaves" are the class labels whereas the branches represent different features where the dataset is split based

on rules. This process is then repeated untill splitting does not aid in the predictions. The GNB classifier is extended on the NB model where it is assumed that continuous data that is associated with each class has a gaussian distribution. [7]

2.5 Machine Learning

The source goes into detail about methodologies that aid in learning through statistical methods. One method is the use of a multi-layer perceptron (MLP). An MLP is a non-linear classifier that utilizes one input layer, at least one hidden layer and one output layer of perceptrons. The hidden layers and the output layers utilize a non-linear activation function. The classifier's perceptrons then learns by altering the weights as each set of data is processed. An MLP is a non-linear classifier as the decision boundary created is a non-linear function of inputs and can work with interactions between features and automatically learn features. [4]

2.6 Bagging predictors

The source describes how bagging is an ensemble algorithm meant to improve stability and accuracy of machine learning algorithms. It can be used with many types of classifiers with the help of model averaging. This helps to avoid overfitting of the model and reduce variance. [2]

3 Approach

3.1 Features

Metadata features and audio features were processed from the dataset and split into training and evaluation features and labels except text based features such as the title. This was done in the preprocessing stages of the model creation. Lyrical features were not used due to the complexity of processing the data in a manner that would be usable.

3.2 Classifiers

Decision Trees were chosen as it can handle categorical and numerical data, resistant to outliers and is a simple model to implement and compute. [7] The Gaussian Naive-Bayes classifier was chosen as it is a simple model that is scalable. [7] Although the assumptions are naïve and the independence assumption is void, the model tends to work well as the classifier is seen to be optimal due to the cancelling out of dependence distribution even when there are strong dependencies. [10] A multi-layer perceptron model was chosen as it is a more complex model than the above two and as the dataset does not appear to be linearly separable, the features are not high-dimensional and some of the features are not interpret-able. [4] Bagging was chosen as it is seen as an effective and simple method to improve the accuracy of a model by combining the predictions of multiple classifiers. [2]

3.3 Methodology

3.3.1 First Hypothesis

Firstly, a baseline model was created using sklearn's most-frequent strategy with a zero-rule classifier. This produces an accuracy of 0.1222 (4 d.p) with "classic pop and rock" being the predicted genre for predictions. The baseline accuracy is used to determine if any models created are more accurate than a zero-rule.

Secondly, three different classifiers, decision tree, gaussian naïve-bayes and a multi-layer perceptron classifier were selected to test my hypothesis due to the variety of representation of types of classifiers that these three provide. Each classifier was trained with metadata features, audio data features and finally a combination of both metadata and audio data features. The classifiers were then used to predict the class labels of the training features to get the estimated apparent accuracy of the classifiers. Learning curves were plotted to analyse the models.

For the decision tree classifiers, a max depth parameter was found by iterating over different depths and finding the depth that provide high accuracy without having too large of a depth (arbitrarily less than 10 to avoid overfitting). For metadata feature classifier, a max depth of 4 was used. For the audio feature classifier, a max depth of 8 was used. For the metadata and audio combination features classifier, a max

depth of 2 was used. Attribute selection was done by using information gain and entropy as a parameter.

Finally, the classifiers were then used on the validation dataset to determine the estimated true accuracy of the classifiers. The output from the predictions was then compared to the estimated apparent accuracy and the accuracy of the various classifiers were compared to test the first hypothesis.

3.3.2 Second Hypothesis

Decision tree, MLP and Gaussian NB classifiers were used as the base estimators in a bagging implementation in sklearn and were trained using the metadata and audio combination dataset. The results were then compared to the results obtained from the predictions made using the classifiers without bagging.

4 Evaluation & Analysis

	Accuracy Rate (6 d.p)
Features	**Decision Tree**
Metadata (Training)	0.376921
Metadata (Validation)	0.284444
Audio (Training)	0.596769
Audio (Validation)	0.4
Metadata & Audio (Training)	0.369106
Metadata & Audio (Validation)	0.333333
Metadata & Audio - Bagging (Validation)	0.313333

Table 1: Accuracy Rates from the Decision Tree on the datasets

4.1 Accuracy Rate Analysis

All classifiers performed better than the baseline. The estimated apparent and true accuracy rates were calculated for each classifier for each dataset. The estimated apparent accuracies of the DT and MLP classifiers were higher than their true accuracies. However, the estimated

	Accuracy Rate (6 d.p)
Features	**Gaussian NB**
Metadata (Training)	0.358687
Metadata (Validation)	0.268888
Audio (Training)	0.418859
Audio (Validation)	0.491111
Metadata & Audio (Training)	0.436833
Metadata & Audio (Validation)	0.506667
Metadata & Audio - Bagging (Validation)	0.508889

Table 2: Accuracy Rates from the Gaussian Naive Bayes Classifier on the Datasets

	Accuracy Rate (6 d.p)
Features	**Multi-Layer Perceptron**
Metadata (Training)	0.347225
Metadata (Validation)	0.306667
Audio (Training)	0.448685
Audio (Validation)	0.313333
Metadata & Audio (Training)	0.486194
Metadata & Audio (Validation)	0.377778
Metadata & Audio - Bagging (Validation)	0.411111

Table 3: Accuracy Rates from the Multi-Layer Perceptron on the Datasets

accuracy of the GNB classifier was lower than the true accuracy when audio and metadata & audio features were used.

4.2 Learning Curve Analysis

Learning curves were plotted with the training data to ascertain the model bias and variance of each classifier when used with the three datasets.

4.2.1 Metadata

The DT classifier (figure 1 in the appendix) has a low-medium variance as the gap between the curves are converging while it has a high bias

as the accuracy, although higher than the baseline, is relatively low, and the accuracy of the training set is decreasing while the validation set accuracy has already plateaued. This would imply that this model is underfitting and does not generalise well.

The gaussian naive-bayes (GNB) (figure 2) has a low variance as the accuracy of the training and cross-validation set have almost converged. It has low-medium bias as the accuracy is not high. This would imply that the model is underfitting and does not generalise well. However, it appears to have the best fit among the curves.

The multi-layer perceptron (MLP) (figure 3) has a low variance as the gap between the two accuracies are small and it has a medium bias as the accuracy is low. This would imply that the model is underfitting and does not generalise well.

4.2.2 Audio

The DT classifier (figure 4) has a high variance as the distance between the training and cross-validation accuracy are large. It has a medium bias as although the curves have not converged, the accuracy for both are relatively high. This would imply that the model is overfitting and does not generalise well.

The GNB classifier (figure 5) has a low variance due to the small gap and a low-medium bias as the accuracy is relatively high and is increasing with more data. This would imply that the model is underfitting and does not generalise well.

The MLP classifier (figure 6) has a low variance due to the small gap and a medium bias as the accuracy is relatively high and has almost converged. This would imply that the model is underfitting and does not generalise well.

4.2.3 Metadata Audio

The DT classifier (figure 7) has a low variance due to the small gap between the curves. It has a medium-high bias due to the low accuracy of both training and cross-validation scores. This would imply that the model is underfitting and does not generalise well.

The GNB classifier (figure 8) has a low variance due to the small gap between the two curves. It has a low-medium bias as the accuracy rates are relatively high. This would imply that the model is underfitting and does not generalise well.

The MLP classifier (figure 9) has a low-medium variance due to the medium gap between the two curves. It has a medium-high bias as although the accuracy rates are relatively high, the curves are decreasing. The medium bias and low variance would imply that the model is underfitting and does not generalise well.

4.3 Comparison of Using Audio & Metadata Features and Both Individually

As seen in tables 1, 2 and 3, the accuracy rate obtained when the classifiers using the validation data set for prediction and using metadata & audio features, are lower for the decision tree classifier than the rates from using the audio features (0.333333 <0.4) but was higher than when compared to when using metadata features (0.284444 <0.333333).

This is contrary to the gaussian naive-bayes model where accuracy rates obtained when using a combination of metadata and Audio features were higher than when used individually (0.268888 & 0.491111 <0.506667). And using audio features in the classifier outperformed a classifier that used metadata features. This could be due to the ability of the NB model to scale to more features.

When a combination of features was used, the MLP classifier outperformed the same classifier when either audio or metadata features were used independently (0.306667 & 0.313333 <0.377778).

4.4 Effectiveness of Bagging on the Performance of the Classifier

As seen in tables 1, 2 and 3, when comparing the accuracy for each classifier when using the metadata and audio features to train the model and when using bagging with these classifiers, it can be seen that in the decision tree classifier, bagging resulted in a decrease in accuracy (0.313333 <0.333333). However, when using the gaussian naive-bayes classifier and the multi-layer perceptron classifier, bagging resulted in a slight increase in accuracy. (0.5066667 <0.50888888) and (0.37777777 <0.4111111)

5 Conclusion

The use of a combination of metadata and audio features to train a classifier might not necessarily increase the accuracy of a model when compared to using each individually. This is due to the properties of each different model as seen in the results above where the use of the combination of features resulted in a higher accuracy for the gaussian naive-bayes classifier and the multi-layer perceptron classifier but not the decision tree classifier.

The use of bootstrap aggregation might not necessarily result in a higher performance classifier. This can be seen in the results obtained where bagging only slightly increased the accuracy of the gaussian naive-bayes and multi-layer perceptron classifiers. And on contrary, resulted in a slight decrease in accuracy for the decision tree. This could be due to the nature of how bagging works through model averaging.

References

[1] Thierry Bertin-Mahieux, Daniel PW Ellis, Brian Whitman, and Paul Lamere. The million song dataset, 2011.

[2] Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

[3] Pedro Cano, Emilia Gomez, Fabien Gouyon, Perfecto Herrera, Markus Koppenberger, Beesuan Ong, Xavier Serra, Sebastian Streich, and Nicolas Wack. Ismir 2004 audio description contest. technical report. 2006.

[4] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* 2009.

[5] C.N. Silla Jr, A.L. Koerich, P. Catholic, and C.A.A. Kaestner. The latin music database. page 451, 2008.

[6] Dawen Liang, Haijie Gu, and Brendan O'Connor. Music genre classification with the million song dataset. *Machine Learning Department, CMU*, 2011.

[7] Tom Mitchell. *Machine Learning*. 1997.

[8] Alexander Schindler and Andreas Rauber. Capturing the temporal domain in echonest features for improved classification effectiveness. In *International Workshop on Adaptive Multimedia Retrieval*, pages 214–227. Springer, 2012.

[9] G. Tzanetakis and P. Cook. *Musical genre classification of audio signals*. 2002.

[10] Harry Zhang. Exploring conditions for the optimality of naive bayes. *International Journal of Pattern Recognition and Artificial Intelligence*, 19(02):183–198, 2005.

A Appendix

A.1 Learning Curves

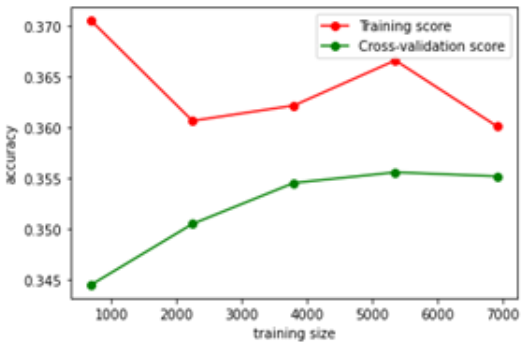

Figure 1: Metadata: Decision Tree

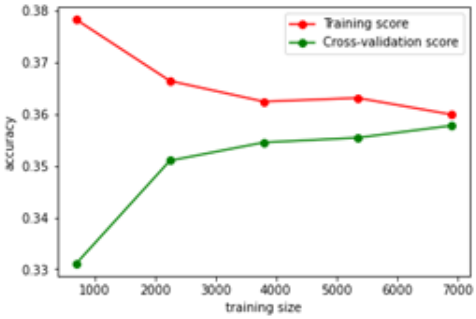

Figure 2: Metadata: Gaussian NB

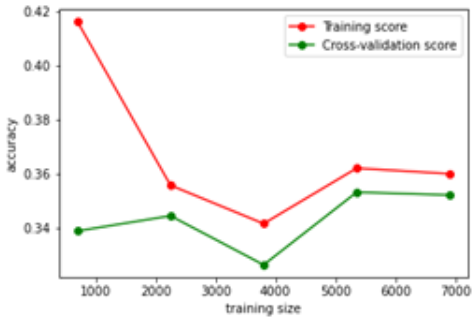

Figure 3: Metadata: Multi-Layer Perceptron

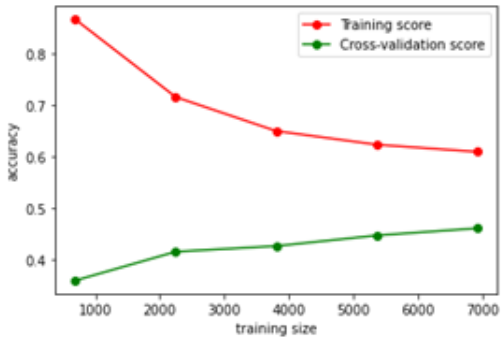

Figure 4: Audio: Decision Tree

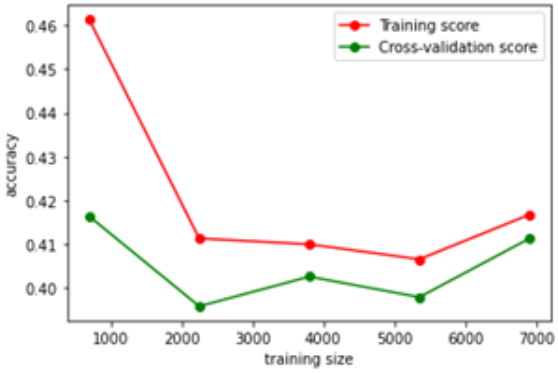

Figure 5: Audio: Gaussian NB

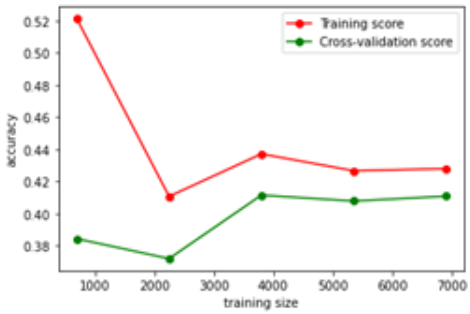

Figure 6: Audio: Multi-Layer Perceptron

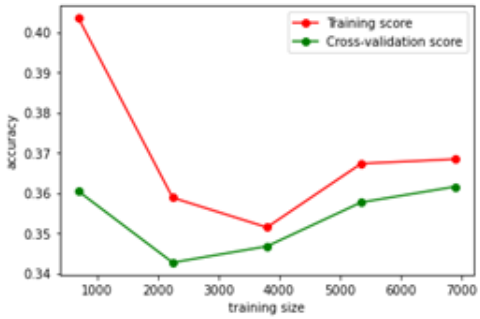

Figure 7: Metadata and Audio: Decision Tree

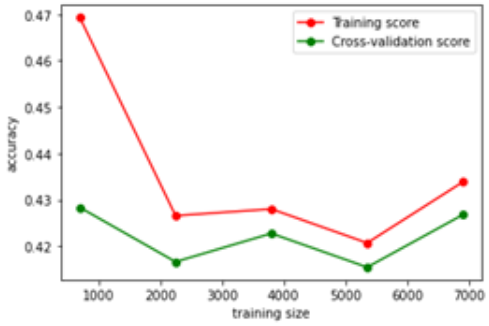

Figure 8: Metadata and Audio: Gaussian NB

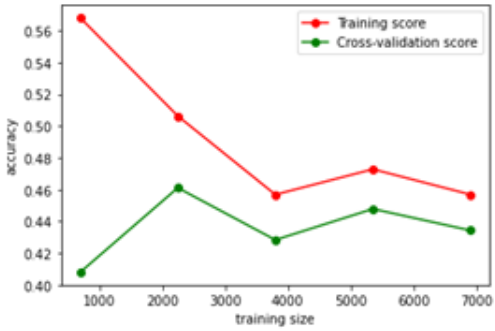

Figure 9: Metadata and Audio: Multi-Layer Perceptron