CS5033 - SL Project Proposal

Airi Shimamura Khoi Trinh

For our SL project, we want to build a few models to predict a user's preference for a song on Spotify.

Each song on Spotify have their own set of 11 features. These features are: danceability, energy, key, loudness, mode, speechiness, acousticness, instrumentalness, liveness, valence, and tempo.

A user can request their streaming history from Spotify via the $Privacy$ section of their account. For the purpose of this project, Khoi's one-year streaming history from February 2022 to February 2023 will be used as the 1 class, for songs we like; and about 4200 random songs which Khoi has never heard will be used as the 0 class, for songs we do not like.

The dataset will have 56,210 songs in class 1, and 4,170 songs in class 0. However, the songs in class 1 have duplicates, due to the nature of streaming a song multiple times. We expect this number will be much lower once the duplicates are removed. We anticipate around 2000 songs in class 1 at the end of the day.

Our intended contributions are: Airi will implement Naive Bayes and Random Forests; while Khoi will implement Logistics Regression and Decision Trees.

Naive Bayes is a group of probabilistic machine learning algorithms based on applying Bayes' theorem with the assumption of independence between features. This assumes that the presence or absence of one feature does not affect the presence or absence of any other feature within a class, and calculates the probabilities of each class label given the values of the features from the input. The class with the highest probability is then selected as the predicted class label for the given input.

Logistic Regression is a statistical machine learning algorithm predicting the probability of a target variable by fitting a logistic function to the input features. This optimizes the parameters of the logistic function by minimizing the cost function, which measures the difference between the predicted probabilities and the actual class labels.

Decision Tree is a machine learning algorithm to predict the class of an input based on its features. This algorithm partitions the input space into increasingly smaller regions based on the values of the input features by selecting the features and test conditions that best separate the training data into the different classes. This process is repeated recursively to create a tree-like structure until a final decision is made.

Random Forest is an ensemble learning method that combines multiple decision trees to improve the performance and reduce overfitting. Each decision tree is constructed by randomly selecting a set of features and samples from the training set. This randomization helps to make a model less sensitive to noise and outliers in the data, and outperforms other algorithms.

Our first hypothesis is that our supervised learning models will be able to classified these songs as likes or dislikes with at least 90% accuracy.

Our second hypothesis is that out of the four chosen algorithms; Random Forest will give us the best performance; followed by Decision Tree, then Logistic Regression, and finally Naive Bayes. We will evaluate each algorithm's performance by comparing the value of Precision, Recall, and produce ROC curves for each algorithm.

Specifically, we will do these calculation for Precision and Recall:

$$Precision = \frac{truepositives}{truepositives + falsepositives}$$

$$Recall = \frac{truepositives}{truepositives + falsenegative}$$