Text-based Audio Classification Assignment Report

1) **Dataset** : It comprises of 800 sound samples from the categories(200 per category): {“Percussion”, “Wind instrument, woodwind instrument”, “Bowed string instrument”, “Domestic sounds, home sounds”}
   Description: The dataset is well balanced as all the categories have equal number of samples. I will be using audio features along with the tags to classify these sounds. The sounds from each category are perceived remarkably different than those from the other categories from our dataset. Since we are using hand-crafted audio features in this assignment, reasons for choosing a feature set will be motivated by the criticality of those features in differentiating the sounds perceptually. For example, pitch salience will be highest in the most melodic category that we have - Bowed string instrument. Hence, perceptual information helps us to choose the audio features which will best classify the sounds.

2) **Methodology:** The keywords for each of the sounds are extracted from metadata. The number of tags to be considered in the feature set is configurable. We consider top N tags which appear most often.
   Training Data: 75% of dataset(randomized), **Testing Data**: 25% (randomized)
   The SVM classifier and decision trees are trained separately and tested over test data. Next, we add the audio features to this feature set that are most relevant for perceptually differentiating these sound categories. The audio features that I consider promising are:
   - **hfc mean** : High Frequency Content feature is useful in identifying onsets in sounds which are percussive.
   - **Average Loudness** : The domestic sounds which are representing a candid auditory scene unlike music being performed, can be expected to have a lower loudness than compared to the musical instruments which are played with conspicuous motives. Also, the role of percussion and melodic instruments are reflected in the average loudness.
   - **Mean Dissonance** : This feature characterizes the inharmonicity in the sounds. It will be very helpful in separating out the melodic instruments which have very low dissonance than compared to others.
   - **Pitch Salience :** This feature is somewhat like the perceptual complement of mean dissonance. This will also help separate out melodic instruments which have very high pitch salience than compared to the rest.
   - **Spectral Flux :** This feature is indicative of timbre of the sound. It tells us how fast a spectrum changes. Hence, this is an important feature to classify even among melodic instruments

.

3) **Results and Observations:** (Due to the limit of 2 pages, I have moved the plots here)

a) Tag Only Based:

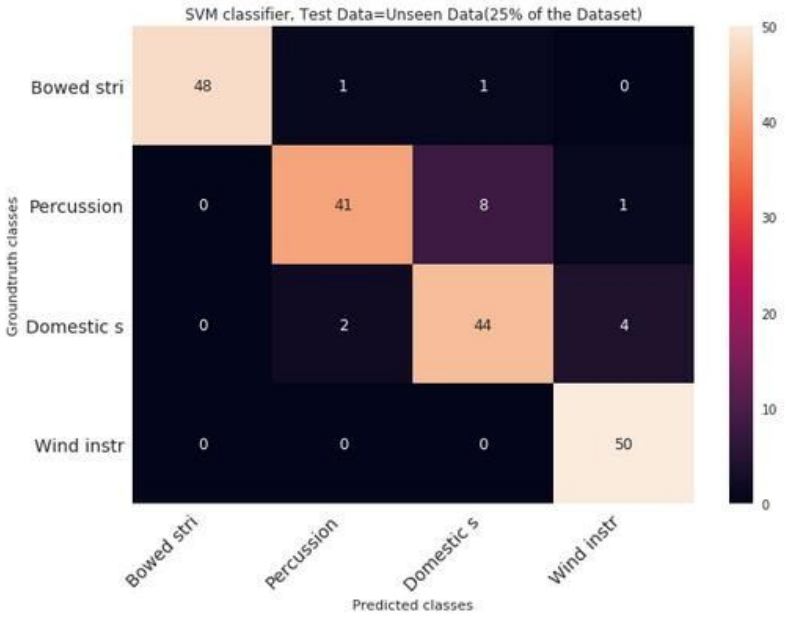
Number of Tags	TREE		SVM	
	On Training Set	On Test Set	On Training Set	On Test Set
25	72.33 %	68 %	86.33 %	84 %
50	67.5 %	65 %	86.67 %	89.5 %

SVM outperforms TREE classifier. The increase in number of tags increases the accuracy of both SVM and TREE classifiers. Training set accuracy is not above 90%, hence it does not seem to be the case of overfitting.

b) Tags + Audio Features Based:

Number of Tags	TREE		SVM	
	On Training Set	On Test Set	On Training Set	On Test Set
25	81.33 %	76 %	88.67 %	85.5 %
50	75.83 %	72.5 %	89.17 %	91.5 %

Including audio features increases the accuracy both for SVM and TREE classifiers. SVM outperforms TREE classifier. The increase in number of tags increases the accuracy of both SVM and TREE classifiers. Training set accuracy is not very high above 90%, hence it does not seem to be the case of overfitting.



Performance of SVM classifier with 50 tags+audio features on test data

The above confusion matrix reveals that Percussion and Domestic category sounds are often misclassified between each other. This is understandable as the sounds from these categories can be very similar than compared to any other pair of categories that we have. This trend is noticed in other plots as well(here).

4) **Conclusions:**
   - We would like to test on the training set to check the extent of models getting overfitted. Since none of the training set accuracies were very high above 90%, the models do not seem to over fit.
   - More the features, better the results. However, the models will overfit after we reach a particular number of features.
   - Cross validation is required to get truthful results.
   - Using multimodal information is beneficial in music information retrieval tasks in general.