Music Genre Classification

Michael Haggblade, Yang Hong, Kenny Kao
Department of Electrical Engineering, Stanford University

Project Overview

Music classification is an interesting problem that has many applications, from Drinkify (an program that generates cocktails to go with your music) to Pandora to dynamically generating images that complement your music. We compared the effectiveness of k-nearest neighbor (k-NN), k-means, SVM, and neural networks in classifying classical, jazz, pop, and metal music.

Results

SVM		Actual genre			
Predicted genre		Classical	Jazz	Metal	Pop
	Classical	29	4	1	1
	Jazz	1	20	1	0
	Metal	0	4	26	0
	Pop	0	2	2	29

NN		Actual genre			
Predicted genre		Classical	Jazz	Metal	Pop
	Classical	14	0	0	0
	Jazz	1	12	4	0
	Metal	0	0	13	0
	Pop	1	0	0	19

Kmeans		Actual genre			
Predicted genre		Classical	Jazz	Metal	Pop
	Classical	14	2	0	0
	Jazz	16	27	0	1
	Metal	0	1	27	1
	Pop	0	0	3	28

K-NN		Actual genre			
Predicted genre		Classical	Jazz	Metal	Pop
	Classical	26	9	0	2
	Jazz	4	20	4	1
	Metal	0	1	24	0
	Pop	0	0	2	27

We found that k-NN and k-means yielded similar accuracies of about 80%. A DAG SVM gave about 87% accuracy and neural networks gave 96% accuracy.

Sources

Chen, P., Liu, S. "An Improved DAG-SVM for Multi-class Classification"

Logan, B. "Mel Frequency Cepstral Coefficients for Music Modeling"

Mandel, M., Ellis, D. "Song-Level Features and SVMs for Music Classification"

Li, T., Chan, A., Chun, A. Automatic Musical Pattern Feature Extraction Using Convolutional Neural Network, IMECS 2010

Zhouyu Fu, Guojun Lu, Kai Ming Ting, and Dengsheng Zhang. A Survey of Audio-Based Music Classification and Annotation, *IEEE Transactions on Multimedia, Vol. 13, No. 2, April 2011.*

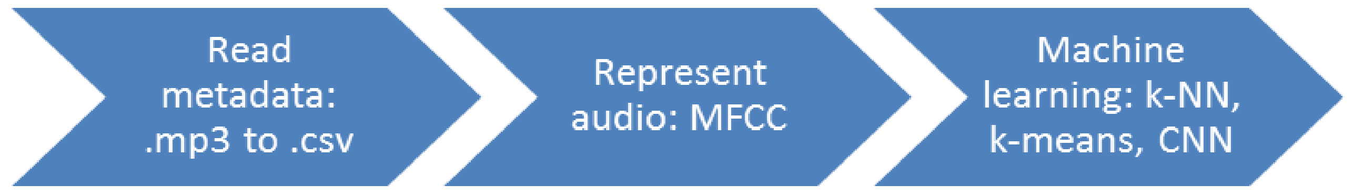
Feature Extraction

GTZAN Music Data Set

"This dataset was used for the well known paper in genre classification "Musical genre classification of audio signals" by G. Tzanetakis and P. Cook in IEEE Transactions on Audio and Speech Processing 2002.

The dataset consists of 1000 audio tracks each 30 seconds long. It contains 10 genres, each represented by 100 tracks. The tracks are all 22050Hz Mono 16-bit audio files in .wav format." [http://marsyas.info/download/data_sets]

Process Flow



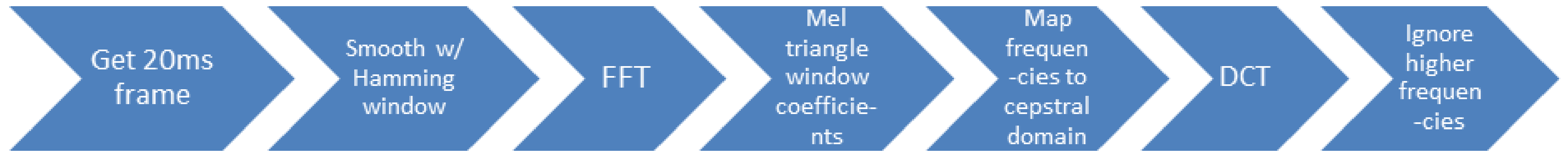
Read metadata: .mp3 to .csv

Represent audio: MFCC

Machine learning: k-NN, k-means, CNN

Mel Frequency Cepstral Coefficients (MFCC)

MFCCs are a way to efficiently represent time domain waveforms as just a few frequency domain coefficients. It bins the frequencies by mapping them to the mel scale, a logarithmic scale modeling human pitch change perception. We further reduce these collapsed frequencies by taking the mean and covariance as the final features for each song.



Get 20ms frame

Smooth w/ Hamming window

FFT

Mel triangle window coefficie-nts

Map frequen-cies to cepstral domain

DCT

Ignore higher frequen-cies

KL Divergence

To measure the "distance" between songs, we use KL divergence. Consider *p(x)* and *q(x)* to be the two multivariate Gaussian distributions with mean and covariance corresponding to those derived from the MFCC matrix for each song:

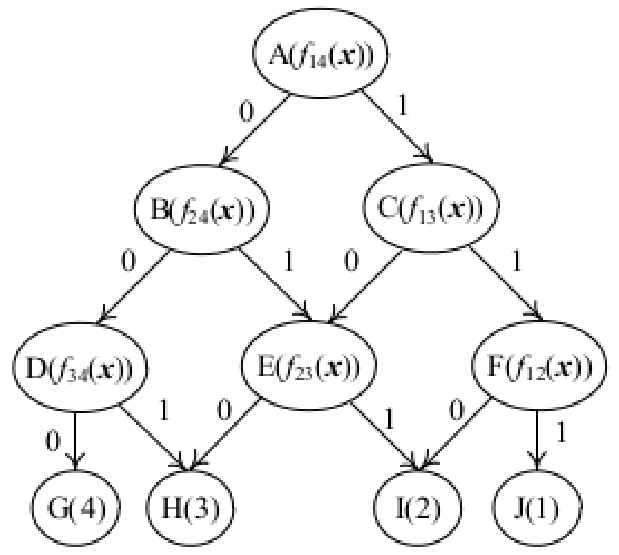
$$2KL(p||q) = \log\left(\frac{|\Sigma_q|}{|\Sigma_p|}\right) + Tr(\Sigma_q^{-1}\Sigma_p) + \left(\mu_p - \mu_q\right)^T \Sigma_q^{-1}\left(\mu_p - \mu_q\right) - d$$

However, since KL divergence is not symmetric but the distance should be symmetric, we have:

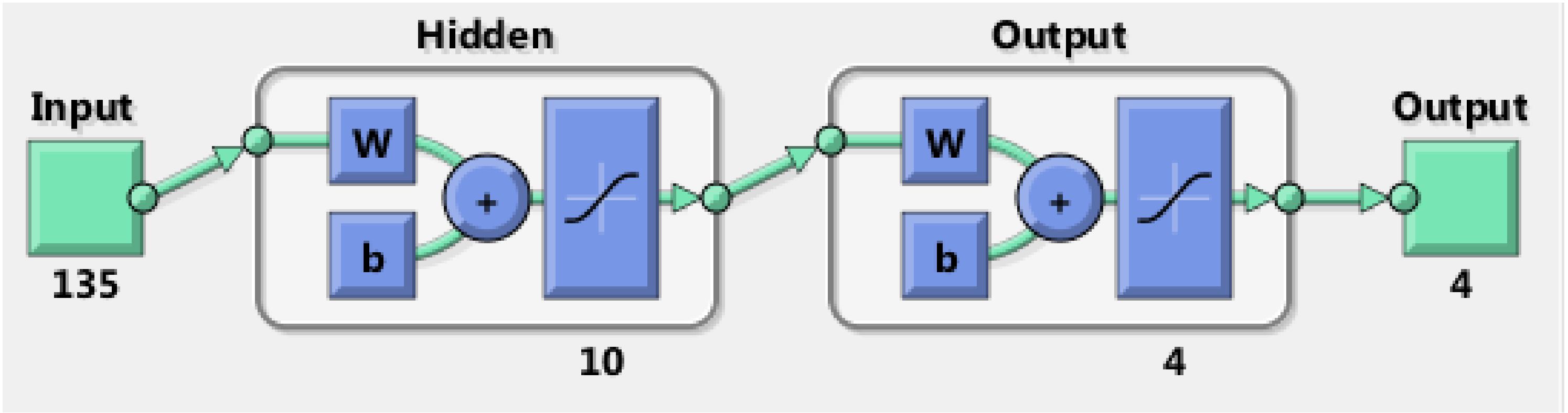
$$D_{KL}(p,q) = KL(p||q) + KL(q||p)$$

Machine Learning Algorithms

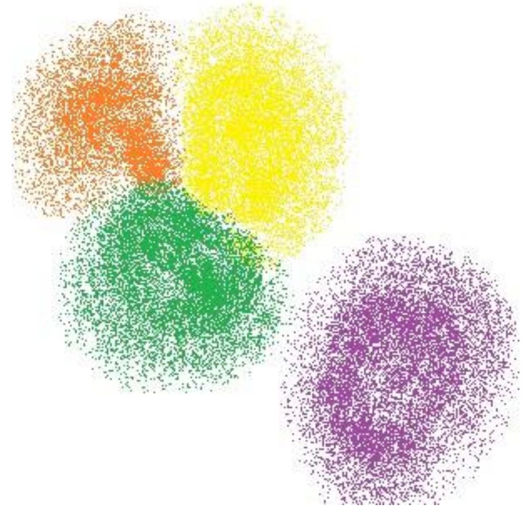
DAG SVM (multi-class)



Neural Networks



K-Means Clustering



K-Nearest Neighbors