Categorical Prediction of Song Popularity Using Topological Data Analysis

Michael Lin, Callie Mao, Grace Mao, Kelvin Niu, Kevin Tie
Professor Sayan Mukherjee

7 December 2015
Statistics 561/Computer Science 571: Probabilistic Machine Learning

1 Introduction

What makes a song popular? With music being an important part of so many peoples lives and since so many new songs are released to the public constantly, it is interesting to consider whether or not there is an accurate way to predict song popularity based on the characteristics of its raw audio file. What are the qualities that most chart-topper singles possess that other lower-ranking songs do not? Perhaps the most downloaded songs share a particular type of rhythm or style that makes them stand out from the vast pool of songs more forgettable to the general audience. Exploring this possibility could yield some productive results.

The goal of this project is to determine whether differences in the audio files of popular modern songs are significant enough to allow for the possibility of classification and categorization. The scope of the project is limited to songs that have appeared on the Billboard 100 list in the past 13 months; in particular, the categories considered are the songs ranked 1, 21, 41, 61 and 81 for each month. While it is clear that the popularity of a song is highly affected by many extraneous factors, e.g. artist(s), marketing, genre, and competition, this project will focus on audio files only in order to maintain similar comparison standards. All data were obtained as .mp3 files from public Internet sources and parsed using discrete Fourier transform.

Should this project produce significant results, the objective is to be able to provide a quantitative way to distinguish among differently ranked songs. The project could then be extended to concentrate on analyzing the exact distinguishing features among modern songs, or to serve as a predictor for how popular any song will likely get even before its release.

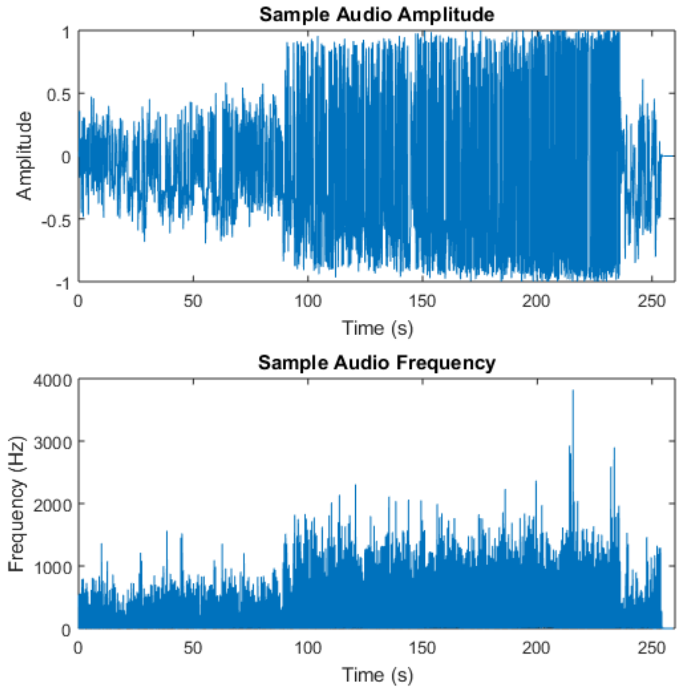
2 Data Collection

Top songs ranked 1, 21, 41, 61 and 81 from Billboard 100 over the last 13 months were chosen as the data set of songs for analysis. The songs were specifically chosen to be 20 ranks apart to maximize possible differences for classification.

Each audio file, obtained from the web, was in MP3 format and was imported into MATLAB as an n x 2 matrix, where n is the number of samples. The entries in the matrix are signed values between -1 and 1, and each column of the matrix represents one of two channels. The matrix represents the amplitude of the sound wave over time, normalized to between -1 and 1. For simplicity, a vector of only the first column (left channel) was used for analysis.

To compute the frequencies at each time point, the amplitude matrix was transformed using the Spectogram function in MATLAB, which calculated the discrete Fourier transform on each subset of the data. The magnitudes of these transformed values were then obtained and returned as a vector of frequencies.

For computational tractability, both the vectors for amplitude and frequency were compressed by only retaining the maximum magnitude value for each 0.1 second of audio. As a result, each vector should have 10 times as many data points as number of seconds in the song.



..

Figure 1: Amplitude and frequency diagrams for Photograph Ed Sheeran.

3 Parameter Feature Selection

Certain features of the parsed audio files were chosen for the analysis of each song. The features were chosen to roughly quantify certain characteristics of the songs while retaining as much information about them as possible. The following six song features were used for analysis:

- Sum of frequencies - summary of how high frequencies are
- Sum of absolute values of intensities - summary of how high intensities are (how loud the song is)
- Sum of absolute values of differences between consecutive frequencies - summary of frequency fluctuation
- Sum of absolute values of absolute differences between consecutive intensities - summary of intensity fluctuation
- Number of frequencies greater than one standard deviation of the total mean - summary of high-frequency outliers
- Percentage of amplitudes with absolute values between 0 and 0.5 - summary of small amplitudes

To capture the time progression of the songs, each song was broken down into three-second frames that overlapped by one second to retain a sense of the connectivity among points. For each frame, quantitative values for each of the six above parameters were calculated to give a 6-dimension point. Thus, each song was represented by a point cloud with (n/10 2) points, where n is the number of seconds in the song.

Finally, all calculated parameters were normalized to the global maximum by dividing each parameter value by the largest value of that parameter over all frames of all songs. This normalization shrinks all point clouds into a unit cube while simultaneously putting equal weight on each of the parameters.

4 Data Analysis

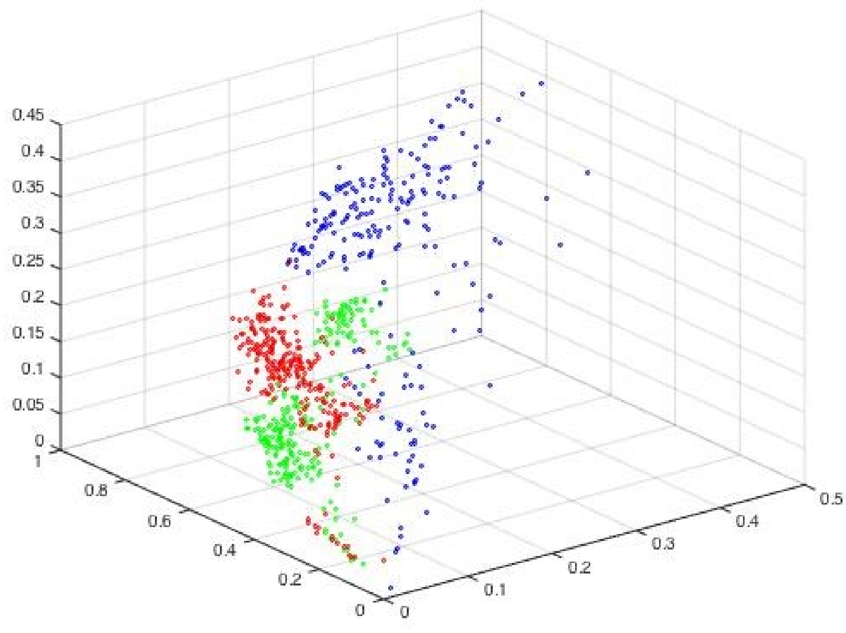
Each song represented as a point cloud in 6-D space normalized to a unit object was further processed in the following ways to produce several single-value characteristics per song:

- Sum the values of each parameter over all frames of one song to produce six distinct values per song
- Analyze the data via Topological Data Analysis (TDA) tools and find
  - The total number of 1-cycles per song
  - The sum of all persistencies of the 1-cycles

In geometry, a k-simplex is a k-dimensional polytope that is the convex hull of its k+1 vertices.[1] In algebraic topology, a k-chain is a linear combination of k-simplices.[2] Furthermore, the boundary of a k-chain is a linear combination of the boundaries of the simplices in the chain, and a k-chain is a k-cycle if its boundary is zero. A 1-cycle is thus a 1-dimensional boundary of any 2-dimensional object.

In the TDA, the radius of each point in the point cloud was expanded over time as the birth and death of 1-cycles were tracked. A 1-cycle is born when the radii of nearby points overlap and merge to form a 1-dimensional closed loop; it dies when the radii expand to fill the hole in the loop. The persistence of a 1-cycle is the difference in time between its birth and death. Analyzing 1-cycles and their persistences provide insight into the arrangement of the data points in space and how they may cluster.

Shown below are the point cloud representations of three songs ranked 1, 41 and 81 from February 2015 considering only the first three parameters for visualization purposes. Somewhat selective clustering of the rank 1 song (red) and rank 81 song (blue) can be seen.



..

Figure 2: Point cloud representation of songs from February 2015 on the first three parameters. Red = Uptown Funk Mark Ronson/Bruno Mars (rank 1), green = Trap Queen Fetty Wap (rank 41), blue = Say You Do Dierks Bentley (rank 81).

Upon obtaining the eight statistical values to represent each song as described above, we analyzed the significance between the different ranks for each of the statistics using permutation tests.

5 Permutation Test

Hypothesis tests using permutation (perm) test as test statistics were performed on the results from topological data analysis. A perm test computes a p-value to determine if the distinction between two sets is significant.[3] The procedure is as follows: given two sets of numbers A and B, the difference between the mean of the sets is first determined. If this difference is significant, then A and B can be called distinct. In order to determine significance, the labeling (A or B) of the values on the sets is randomized n times. The p-value is calculated as the number of times the new difference of the means exceeds than the original difference divided by the number of times the labeling was randomized. Let p-values less than 0.05 be significant and p-values above 0.05 be insignificant. Intuitively, if the clustering of two sets of points is significantly different, then random relabeling of the points would tend to give smaller differences between the means than the original difference between the means, leading to a smaller p-value. For each of the eight statistical values, we ran pairwise perm tests on all pairs of ranks with n = 100,000 and calculated p-values (e.g. rank 1 group vs. rank 41 group) and found the following results.

For the statistics calculated by summing the parameters for each song, we found the following significant values at the 0.05-level:

- Frequency fluctuations for (21, 61) and (21, 81) ranking group pairs
- Frequency outliers for (61, 81) ranking group pair
- Amplitude magnitude for (1, 61), (1, 81), and (41, 81) ranking group pairs

For number of one-cycles, we observed the following rank pairs to be significant (the two can be distinguished from each other): 1/41, 1/81, 41/61, 61/81. It seems that rank 1 and rank 61 songs can be distinguished from rank 41 and 81 songs. This feature reflects the spatial distribution of the points in the point cloud, which is an overall reflection of all the other parameters used for calculation. The rank pairs that can be distinguished from each other have very different point distributions with respect to the formation of one-cycles.

For the sum of all one-dimensional persistences, we observed the following pairs to be moderately significant ($0.05 < p < 0.20$): 1/41, 21/41. As opposed to simply calculating the number of one-cycles, this statistic considers the persistences of those one-cycles, or how long they survive. This gives a sense of how widely separated various clusters within a song are. The result from this statistic is as not as significant as that for just the number of 1-cycles, which may be because summing the persistences actually loses information about them.

6 Results and Conclusion

Considering only TDA heuristics, a significant difference was found between the following pairs of ranks: 1 and 41, 21 and 41, while a somewhat significant difference was found between the following pairs of ranks: 1/81, 41/61, 61/81.

Considering all heuristics used in the pair-wise permutation tests, the following shows order of decreasing significance among pairs:

$$1/61, 21/41, 41/61, 21/61, 1/81, 21/81, 41/81, 1/41, 1/21, 61/81$$

Although the data is by no means conclusive or decisive, roughly speaking there seems to be a moderate difference when comparing high ranks (1, 21) to mid ranks (41) and mid ranks to low ranks (61, 81), but not as much difference when comparing high ranks to low ranks.

Several factors may contribute to the lack of definitive results. First of all, the parsed audio files may contain exorbitant noise that could reduce the distinctiveness among the songs. The parameters we chose for characterization are rather rudimentary and may fail to consider certain characteristics of the songs, such as the singers gender, the presence of rhythmic patterns and the distinction between harmony and melody. Furthermore, calculating individual characteristics of the point clouds may have reduced the information leading to the final categorization.

7 Future Direction

Our results represent only a very rudimentary analysis of song classification by rankings. There are many possible improvements and directions to take this project. First of all, having a larger dataset of pieces for each rank would likely give a better standard for comparison. There are also other features that may be worth analyzing, such as melodic patterns, dynamic variations, tempo changes and rhythmic nuances. The parameters chosen to characterize the pieces could be tested and refined until an optimal set of parameters is found. We could also give different features different levels of importance by multiplying by different weights rather than normalizing all features to an interval between zero and one. Additionally, instead of running only pairwise permutations, running TDA on all groups simultaneously would allow for better holistic results. If we took this project one step further, we could use a K-nearest-neighbor classification for an unknown song to see what rank the model would predict the song to have. For example, once we calculate the statistical value or vector to represent an unknown song, we take the K nearest valued neighbors and classify the song under the rank with the most nearest neighbors.

8 References

[1] http://mathworld.wolfram.com/Simplex.html

[2] http://mathworld.wolfram.com/topics/AlgebraicTopology.html

[3] Legendre, P. & L. Legendre. 1998. Numerical ecology, 2nd English edition. Elsevier Science BV, Amsterdam

Appendix A – Songs

	Rank 1	Rank 21	Rank 41	Rank 61	Rank 81
2015 Nov	Adele - Hello	X Ambassadors - Renegades	Fetty Wap - Again	DLOW - Bet You Can't Do It Like Me Challenge	Demi Lovato - Cool for the Summer
2015 Oct	The Weeknd - The Hills	Fetty Wap - Come My Way	Travis Scott - Antidote	Carrie Underwood - Smoke Break	Gonna - Blake Shelton
2015 Sep	The Weeknd - Can't Feel My Face	Wiz Khalifa - See You Again	Maroon 5 - Sugar	She's Kinda Hot - 5 Seconds of Summer	Nothin' Like You Dan and Shay
2015 Aug	OMI - Cheerleader	Omarion ft. Chris Brown & Jhene Aiko - Post to Be	She's Kinda Hot - 5 Seconds of Summer	Lose My Mind - Brett Eldredge	Chris Young - I'm coming over
2015 Jul	OMI - Cheerleader	Omarion ft. Chris Brown & Jhene Aiko - Post to Be	House Party - Sam Hunt	Marvin Gaye - Charlie Puth	How Many Times - DJ Khaled
2015 Jun	Wiz Khalifa - See You Again	Thinking Out Loud - Ed Sheeran	Photograph - Ed Sheeran	Crash and Burn - Thomas Rhett	Kiss You in the Morning - Michael Ray
2015 May	Wiz Khalifa - See You Again	Chains - Nick Jonas	Nasty - Bandit Gang Marco	Heartbeat Song - Kelly Clarkson	Classic Man - Jidenna
2015 Apr	Wiz Khalifa - See You Again	Bitch Better Have My Money - Rihanna	Only - Nicki Minaj	Get Low - Dillon Francis	Homegrown Honey - Darius Rucker
2015 Mar	Uptown Funk - Mark Ronson/Bruno Mars	Shut Up and Dance - Walk the Moon	Omarion ft. Chris Brown & Jhene Aiko - Post to Be	Throw Sum Mo - Rae Sremmurd/Nicki Minaj	Luke Bryan - I See You
2015 Feb	Uptown Funk - Mark Ronson/Bruno Mars	I Don't Fk With You - Big Sean	Trap Queen - Fetty Wap	Often - The Weeknd	Say You Do - Dierks Bentley
2015 Jan	Uptown Funk - Mark Ronson/Bruno Mars	I Love Makonnen - Tuesday	Something in the Water - Carrie Underwood	Nicki Minaj - Feeling Myself	J-Cole - Apparently
2014 Dec	Taylor Swift - Blank Space	Calvin Harris - Blame	Fergie - L.A. Love	Luke Bryan - I See You	Matt McAndrew - Make It Rain
2014 Nov	Taylor Swift - Blank Space	Calvin Harris - Blame	Fergie - L.A. Love	Alesso - Heroes	Luke Bryan Roller Coaster

Appendix B – Topological Data Analysis Results

Rank 1

	Num Verts	Num Edges	Num 1 Cycles	Total Pers
15-11	290	27119	76	0.5607
15-10	232	17458	57	0.5324
15-9	216	20550	74	0.4702
15-8	177	12537	51	0.4334
15-7	177	12537	51	0.4334
15-6	224	14497	51	0.4648
15-5	224	14497	51	0.4648
15-4	224	14497	51	0.4648
15-3	268	33991	90	0.6024
15-2	268	33991	90	0.6024
15-1	268	33991	90	0.6024
14-12	261	28583	86	0.6991
14-11	261	28583	86	0.6991
Average	237.6923077	22525.46154	69.53846154	0.540761538

Rank 21

	Num Verts	Num Edges	Num 1 Cycles	Total Pers
15-11	192	17875	59	0.3247
15-10	284	36847	104	0.7167
15-9	224	14497	51	0.4648
15-8	226	23273	93	0.6947
15-7	226	23273	93	0.6947
15-6	265	34937	121	0.7102
15-5	211	21783	67	0.797
15-4	217	21123	81	0.7824
15-3	196	18738	53	0.2422
15-2	282	35709	101	0.6458
15-1	268	35778	92	0.1876
14-12	205	18235	46	0.4047
14-11	205	18235	46	0.4047
Average	230.8461538	24638.69231	77.46153846	0.543861538

Rank 41

	Num Verts	Num Edges	Num 1 Cycles	Total Pers
15-11	308	41982	131	1.0337
15-10	280	26806	78	0.581
15-9	236	25829	96	0.7543
15-8	217	19888	57	0.8948
15-7	186	17205	69	0.419

15-6	257	23190	77	0.52
15-5	158	12365	63	0.3693
15-4	308	44707	135	1.3286
15-3	226	23273	93	0.6947
15-2	243	29114	95	0.919
15-1	225	23013	83	0.8107
14-12	190	17736	59	0.7151
14-11	190	17736	59	0.7151
Average	232.6153846	24834.15385	84.23076923	0.750407692

Rank 61

	Num Verts	**Num Edges**	**Num 1 Cycles**	**Total Pers**
15-11	144	9464	43	0.6172
15-10	200	18039	66	0.5764
15-9	217	19888	57	0.8948
15-8	154	11466	48	0.4855
15-7	191	18145	68	0.2224
15-6	186	17183	55	0.5127
15-5	200	19006	76	0.4859
15-4	210	21852	65	0.6929
15-3	256	20857	85	0.6555
15-2	247	26977	107	0.9573
15-1	235	25411	92	0.9695
14-12	188	16980	65	0.4929
14-11	207	19718	56	0.4087
Average	202.6923077	18845.07692	67.92307692	0.613207692

Rank 81

	Num Verts	**Num Edges**	**Num 1 Cycles**	**Total Pers**
15-11	208	20757	55	0.2948
15-10	183	15477	58	0.5594
15-9	184	13325	58	0.4554
15-8	195	16899	80	0.5256
15-7	261	33092	123	0.8698
15-6	178	15753	60	0.3953
15-5	234	27255	115	1.0717
15-4	201	19827	77	0.4784
15-3	188	16980	65	0.4929
15-2	214	16758	71	0.635
15-1	291	39199	131	1.1549
14-12	214	19852	69	0.9286
14-11	257	31696	119	0.5576
Average	216	22066.92308	83.15384615	0.647646154

APPENDIX C – P-VALUES FROM DATA ANALYSIS

Sum of Persistences

	1 21	1 41	1 61	1 81	21 41	21 61	21 81	41 61	41 81	61 81
Mean Difference	0.0031	0.2096	0.0724	0.1069	0.2065	0.0693	0.1038	0.1372	0.1028	0.0344
P-values	0.9617	0.0082	0.2892	0.1939	0.0355	0.4213	0.2878	0.1606	0.3315	0.7244

Number of 1-Cycles

	1 21	1 41	1 61	1 81	21 41	21 61	21 81	41 61	41 81	61 81
Mean Difference	7.9231	14.6923	1.6154	13.6154	6.7692	9.5385	5.6923	16.3077	1.0769	15.2308
P-values	0.3526	0.1002	0.8093	0.1503	0.4978	0.2712	0.5840	0.0721	0.9132	0.1125

Sum of Frequencies

	1 21	1 41	1 61	1 81	21 41	21 61	21 81	41 61	41 81	61 81
Mean Difference	6.7213	4.6172	10.0790	8.4077	11.3385	16.8003	15.1290	5.4617	3.7904	1.6713
P-values	0.4188	0.5751	0.1723	0.2062	0.2113	0.0414	0.0422	0.5129	0.6277	0.7965

Sum of Absolute Intensities

	1 21	1 41	1 61	1 81	21 41	21 61	21 81	41 61	41 81	61 81
Mean Difference	8.3639	8.0610	13.1784	0.6238	16.4249	21.5423	8.9878	5.1173	7.4372	12.5545
P-values	0.5357	0.5500	0.3295	0.9585	0.2986	0.1758	0.5363	0.7420	0.6062	0.3879

Frequency Fluctuation

	1 21	1 41	1 61	1 81	21 41	21 61	21 81	41 61	41 81	61 81
Mean Difference	6.4096	6.0522	12.6494	11.2101	12.4618	19.0590	17.6197	6.5972	5.1579	1.4394
P-values	0.4807	0.5135	0.1253	0.1326	0.2116	0.0323	0.0273	0.4736	0.5566	0.8376

Intensity Fluctuation

	1 21	1 41	1 61	1 81	21 41	21 61	21 81	41 61	41 81	61 81
Mean Difference	18.2065	0.9831	17.0647	13.4698	17.2234	1.1418	4.7367	16.0816	12.49	3.594
P-values	0.1614	0.9472	0.1918	0.3246	0.1283	0.8815	0.5954	0.1581	0.305	0.693

High-Frequency Outliers

	1 21	1 41	1 61	1 81	21 41	21 61	21 81	41 61	41 81	61 81
Mean Difference	0.1716	4.5385	13.1302	1.1420	4.7101	13.3018	0.9704	8.5917	5.6805	14.2722
P-values	0.9841	0.5987	0.1526	0.8916	0.5025	0.0822	0.8862	0.2005	0.3364	0.0246

Percentage of Small Amplitudes

	1 21	1 41	1 61	1 81	21 41	21 61	21 81	41 61	41 81	61 81
Mean Difference	18.2357	1.9752	30.1464	33.5856	20.210	11.910	15.349	32.121	35.560	3.4392
P-values	0.4319	0.8677	0.0347	0.0180	0.4053	0.6582	0.5518	0.0655	0.0393	0.8596