INFO523 SVM & ANN

Noah Giebink & Sebastian Deimen

4/3/2020

Classify track explicitness using sociopolitical variables (SVM)

Data Preprocessing

```
spot <- read_csv('../data/spot_clean.csv') %>% select(country, track.explicit, happiness,
                                                      median_age, percent_urban,
                                                      percent_internet_users, density_sqkm,
                                                      freedom, gdp)
```

Table 1. There is only one observation of sociopolitical variables per country, but ~50 unique tracks. Therefore, both levels of track.explicit (T/F) are associated with many duplicate observations, making classification difficult. To illustrate, we show the first handful of observations for the Dominican Republic.

country	track.explicit	happiness	median_age	percent_urban	percent_internet_users	density_sqkm	freedom	gdp
Dominican Republic	FALSE	5.433216	26.1	80.3	63.9	230	3	3.52
Dominican Republic	TRUE	5.433216	26.1	80.3	63.9	230	3	3.52
Dominican Republic	FALSE	5.433216	26.1	80.3	63.9	230	3	3.52
Dominican Republic	TRUE	5.433216	26.1	80.3	63.9	230	3	3.52
Dominican Republic	FALSE	5.433216	26.1	80.3	63.9	230	3	3.52
Dominican Republic	FALSE	5.433216	26.1	80.3	63.9	230	3	3.52

```
exp <- group_by(spot, country, track.explicit) %>% tally() %>% filter(track.explicit==TRUE) %>%
  select(country, num_explicit = n)
songs_count <- group_by(spot, country) %>% summarise(count = n())
exp_count <- left_join(exp, songs_count, by = 'country')
exp_freq <- mutate(exp_count, freq_explicit = num_explicit/count) %>% select(country, freq_explicit)
exp_freq <- mutate(exp_freq, explicit_label = ifelse(freq_explicit > median(exp_freq$freq_explicit), 1,

spot_exp <- left_join(exp_freq, spot, by = 'country') %>% select(-track.explicit, -freq_explicit)
spot_exp <- unique(spot_exp)
```

Table 2. We solve this problem by labelling countries' taste as particularly explicit (+1) or not particularly explicit (-1) and retaining only unique observations (i.e. only one observation per country).

country	explicit_label	happiness	median_age	percent_urban	percent_internet_users	density_sqkm	freedom	gdp
Argentina	-1	5.792797	30.8	91.7	71.0	16.60	2.0	3.9800
Australia	-1	7.176993	37.4	85.9	88.2	3.31	1.0	2.4700
Brazil	-1	6.190922	31.3	86.3	60.9	25.60	2.0	2.2400
Canada	1	7.175497	40.5	81.3	91.2	4.14	1.0	1.7700
Chile	1	6.436221	33.7	87.5	83.6	24.80	1.0	4.2100
China	-1	5.131434	37.0	58.0	53.2	152.00	6.5	7.6000
Colombia	1	5.983512	30.1	80.4	58.1	45.30	3.0	4.3700
Dominican Republic	1	5.433216	26.1	80.3	63.9	230.00	3.0	3.5200
France	1	6.665904	41.2	80.2	79.3	120.00	1.5	0.3390
Germany	1	7.118364	45.9	77.3	89.6	237.00	1.0	-0.0344
Guatemala	1	6.626592	21.3	50.7	34.5	167.00	4.0	3.5800
India	-1	3.818069	26.7	33.6	29.5	465.00	2.5	4.3700
Indonesia	-1	5.340296	28.0	54.7	25.4	150.00	3.0	5.3300
Italy	1	6.516527	45.9	70.1	61.3	201.00	1.0	-2.1300
Mexico	-1	6.549579	27.5	79.9	59.5	68.90	3.0	1.0600
Peru	1	5.679661	27.5	77.7	45.5	26.00	2.5	4.6100
Philippines	-1	5.869173	24.1	46.7	55.5	368.00	3.0	7.1000
Russia	-1	5.513500	38.7	74.3	73.1	8.78	6.5	1.2300
South Africa	-1	4.883922	26.1	65.8	54.0	48.40	2.0	1.0100
Spain	-1	6.513371	43.2	80.1	80.6	93.10	1.0	-0.3140
United Kingdom	1	7.233445	40.2	83.1	94.8	278.00	1.0	1.7600
United States	1	6.882685	37.6	82.1	76.2	36.20	1.5	2.0300