Statistical Analysis of Spotify Streaming History

Seth Ketron

June 2020

Abstract

The purpose of this analysis is to determine relationsips in features of music. The dataset comes from the author's spotify streaming history for one year from May 2019 - May 2020. The dataset consists of 1,451 tracks and 13 different features. The first model fit is a simple logistic regression model. Multiple logistic regression was then performed using all predictor variables and then performed using the lasso with modality as the response variable. The validation set approach was used to determine estimates for the test error rate. The conclusion of the analysis reports that multiple logistic regression using the lasso provides the lowest estimated test error rate and that certain key signatures, danceability, speechiness, and acousticness are the most important variables in predicting the modality of a track.

1 Materials and Methods

Streaming services have become a premiere way to consume music. Spotify is a popular music streaming service that offers users the ability to request their streaming data and utilize the Spotify API to recover various features of each track. These features can be used with statistical learning methods to provide insight into a user's listening habits.

The dataset being explored is the author's streaming history from May 2019 - May 2020. The dataset consists of 1,451 tracks and 13 features. Some of the features include the key of the track and the modality of the track. The key denotes the key signature the track is composed in, and the modality indicates whether the track uses minor or major scales in its composition. Some other subjective features include ratings of listening features such as danceability which indicates a track's suitability for dance. Valence is a measure of the degree to which a song invokes positive feelings. A higher valence rating indicates a song that invokes more positive feelings. Speechiness is a measure of the presence of spoken words and acousticness is a confidence measure of whether a track is acoustic. In utilizing statistical learning methods, we determine a response variable that we will use to fit a model. Given our dataset $\mathbf{X}$ with columns indicating each feature and rows indicating each song, we have

$$\mathbf{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,12} \\ x_{2,1} & x_{2,2} & \dots & x_{2,12} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1451,1} & x_{1451,2} & \dots & x_{1451,12} \end{bmatrix}$$

where each $x_{i,j}$ with $i = 1, \dots, 1451$ denoting the song and $j = 1, \dots, 12$ indicating the feature for that respective song. Modality is the decided upon response which we denote by $\mathbf{Y}$ where

$$\mathbf{Y} = \begin{bmatrix} y_{1,1} \\ y_{2,1} \\ \vdots \\ y_{1451,1} \end{bmatrix}$$

where each $y_{i,j}$ with $i = 1, \dots, 1451$ denoting the song and $j = 1$ indicating the modality for that respective song.

Logistic regression is a statistical learning method that is useful when the response variable is qualitative. A qualitative variable is a variable that is non-numeric, and given that the modality of a song is either minor or major, the response can take on values of two distinct classes. The purpose of logistic regression is to use $\mathbf{X}$ to determine the probability that $\mathbf{Y}$ takes on a value of 0 or 1 where 0 denotes a minor modality and 1 indicates a major modality. Since probabilities take on values between 0 and 1, we determine a threshold for a classifier that will assign a song to either class. If $p(x_i)$ is greater than 0.5, then the classifier will assign y_i to the major class.

To determine $p(\mathbf{X})$, logistic regression uses the logistic function which is given by

$$p(\mathbf{X}) = \frac{e^{\beta_0 + \beta_1 \mathbf{X}_1 + \dots + \beta_{12} \mathbf{X}_{12}}}{1 + e^{\beta_0 + \beta_1 \mathbf{X}_1 + \dots + \beta_{12} \mathbf{X}_{12}}}. \tag{1}$$

$\beta_0, \dots, \beta_{12}$ are called regression coefficients and will be discussed later. The model is fit using a method called maximum likelihood. With some algebraic manipulation, we see that

$$\frac{p(\mathbf{X})}{1 - p(\mathbf{X})} = e^{\beta_0 + \beta_1 \mathbf{X}_1 + \dots + \beta_{12} \mathbf{X}_{12}}. \tag{2}$$

The left hand side of (2) is called the odds. The odds can take on values between 0 and ∞, where values close to zero represent very low probabilities and very high values represent very high probabilities. By taking the log on both sides of equation (2) we have

$$log\left(\frac{p(\mathbf{X})}{1 - p(\mathbf{X})}\right) = \beta_0 + \beta_1 \mathbf{X}_1 + \dots + \beta_{12} \mathbf{X}_{12}. \tag{3}$$

The left hand side of (3) is called the logit and we note that it is linear in $\mathbf{X}$ on the right hand side.

We determine estimates for the regression coefficients $\beta_0, \ldots, \beta_{12}$, which we will denote as $\hat{\beta}_0, \ldots, \hat{\beta}_{12}$, by finding $\hat{\beta}_0$, $\ldots$, $\hat{\beta}_{12}$ such that the predicted probability of x_i, namely $\hat{p}(x_i)$, is as close to the observed y_i as possible. This is where the maximum likelihood method comes into play. The method utilizes a function called the likelihood function, given by

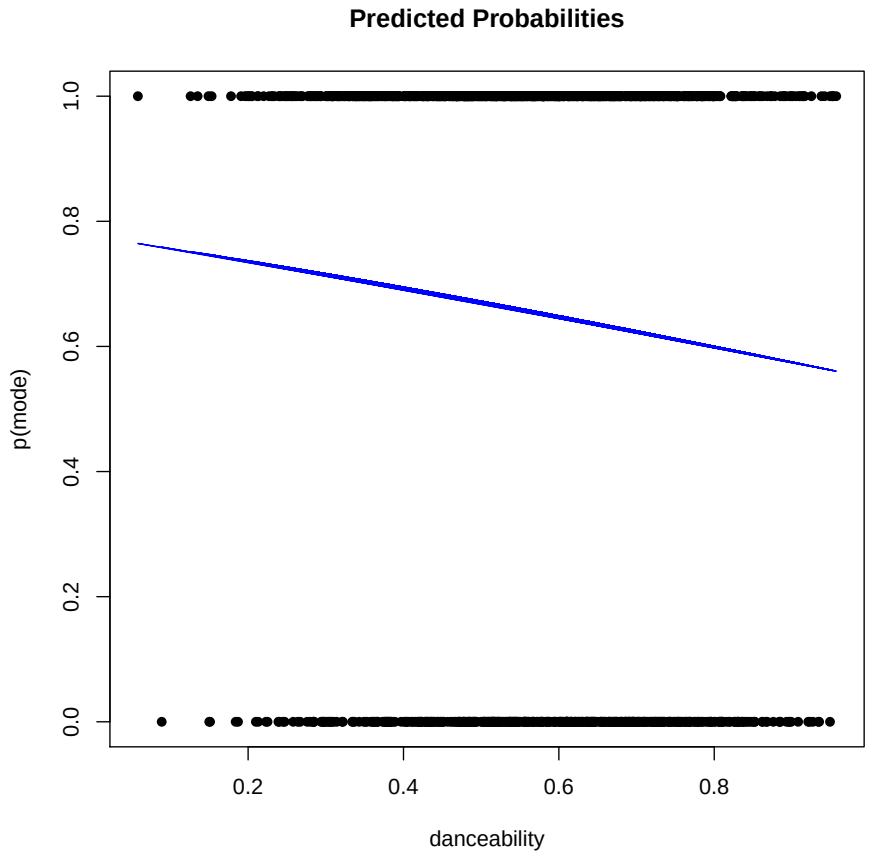
$$\ell(\beta_0, \ldots, \beta_{12}) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} (1 - p(x_{i'})). \tag{4}$$

The estimates $\hat{\beta}_0, \ldots, \hat{\beta}_{12}$ are selected to maximize (4).

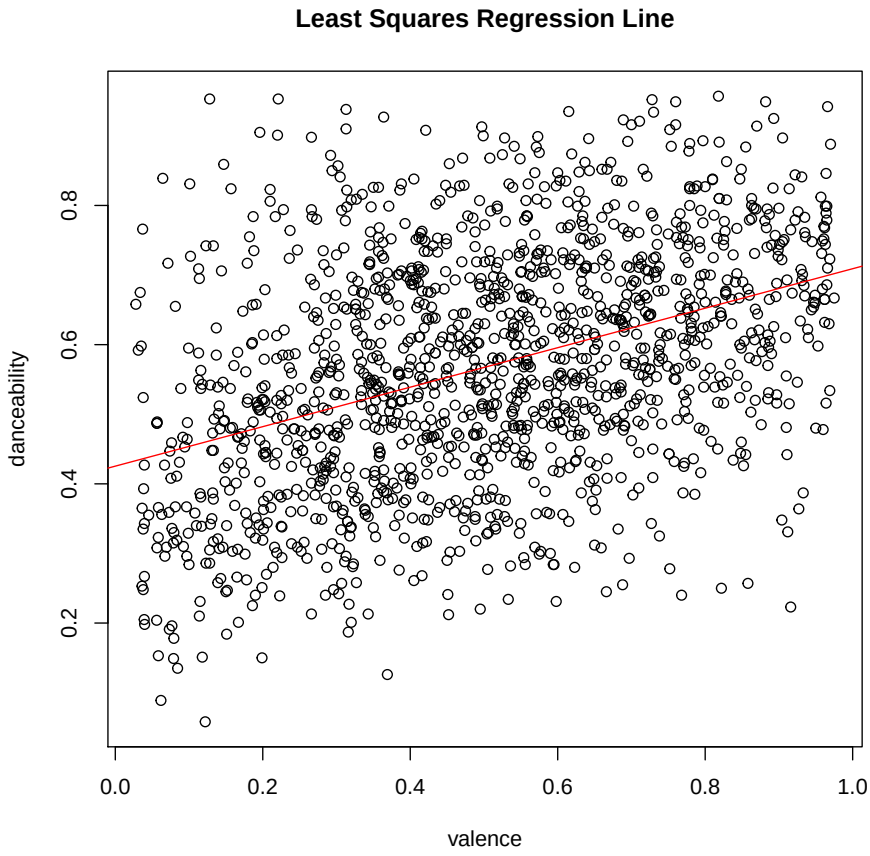
The lasso is a method that will help determine the most important features in predicting modality. The lasso provides an alternative approach to estimating the regression coefficients by shrinking some of them to be exactly zero. This is done by minimizing a slightly different function instead of maximizing the likelihood function. The method gives an alternate set of regression coefficient estimates, denoted by $\hat{\beta}^L_\lambda$. The idea behind the lasso is that it offers better model interpretability due to the fact some of the coefficients are shrunk to exactly zero, and coefficients remaining are associated with the most important features in prediction.

2 Results and Discussion

To begin the discussion on results, we fit a logistic regression model using only one of the predictors. We fit this model with modality as the response variable and danceability as the sole predictor variable. Below is a plot of the predicted probabilities.



We can see that as the danceability rating of a song increases, the probability that the song is in major scale decreases. This is an interesting result as songs composed using major scales tend to be associated with more positive emotions and that has a linearly increasing relationship with danceability as we can see from the plot below.



From the results given below of the logistic model summary we can see that danceability has a significant p-value of 0.00177, which suggests there is a statistically significant relationship in making modality predictions based on a song's danceability rating. From previously, since this logistic model only makes use of one predictor variable, we will have regression coefficient estimates for β_0 and β_1, namely $\hat{\beta}_0$ and $\hat{\beta}_1$. From the summary, these estimates are $\hat{\beta}_0 = 1.2385$ and $\hat{\beta}_1 = -1.0402$.

```
##
## Call:
## glm(formula = spotify$mode ~ spotify$danceability,
    family = binomial,
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q
```

```
## -1.6865  -1.4022   0.8647   0.9351
##     Max
##  1.0761
##
## Coefficients:
##                       Estimate
## (Intercept)             1.2385
## spotify$danceability  -1.0402
##                       Std. Error z value
## (Intercept)               0.1998   6.198
## spotify$danceability      0.3327  -3.127
##                       Pr(>|z|)
## (Intercept)           5.73e-10 ***
## spotify$danceability   0.00177 **
## ---
## Signif. codes:
##   0 '***' 0.001 '**' 0.01 '*' 0.05
##   '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be
   1)
##
##     Null deviance: 1869.0  on 1450  degrees of freedom
## Residual deviance: 1859.1  on 1449  degrees of freedom
## AIC: 1863.1
##
## Number of Fisher Scoring iterations: 4
```

To determine how well our model performs we will use it to make modality predictions on the validation set. Below we can see our classifer does not make any minor modality predictions, but it correctly classifies the modality of 65% of songs with a test error rate of 35%.

```
##
## simplog.pred   0   1
##            1 500 951

## [1] 0.6554101
```

Following this, a multiple logistic regression model was fit making use of all predictor variables to see how well it performs in making predictions. The idea was to determine a model that will make both minor and major modality predictions. The summary is given below, and we can see that the model

determines certain key signatures and danceability to be the most statistically significant predictors.

```
##
## Call:
## glm(formula = spotify$mode ~ spotify$key +
    spotify$time_signature +
##     spotify$danceability + spotify$energy +
    spotify$loudness +
##     spotify$speechiness + spotify$acousticness +
    spotify$instrumentalness +
##     spotify$liveness + spotify$valence + spotify$tempo
     + spotify$duration_ms,
##     family = binomial, data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q
## -2.2898  -1.1091   0.6414   0.9163
##     Max
##  1.5901
##
## Coefficients:
##                             Estimate
## (Intercept)                2.685e+00
## spotify$key1              -8.132e-01
## spotify$key2              -4.026e-02
## spotify$key3              -1.644e+00
## spotify$key4              -1.544e+00
## spotify$key5              -9.230e-01
## spotify$key6              -1.525e+00
## spotify$key7              -2.903e-01
## spotify$key8              -4.901e-01
## spotify$key9              -1.125e+00
## spotify$key10             -1.823e+00
## spotify$key11             -1.956e+00
## spotify$time_signature4    3.870e-02
## spotify$time_signature5   -8.357e-01
## spotify$danceability      -1.230e+00
## spotify$energy            -9.202e-01
## spotify$loudness           3.372e-02
## spotify$speechiness       -7.456e-01
## spotify$acousticness       5.420e-01
## spotify$instrumentalness   1.747e-01
## spotify$liveness           5.874e-01
## spotify$valence            4.138e-01
```

```
## spotify$tempo              2.208e-03
## spotify$duration_ms       -9.136e-07
##                          Std. Error
## (Intercept)               7.253e-01
## spotify$key1              2.816e-01
## spotify$key2              3.026e-01
## spotify$key3              4.027e-01
## spotify$key4              2.808e-01
## spotify$key5              2.927e-01
## spotify$key6              2.908e-01
## spotify$key7              2.850e-01
## spotify$key8              3.295e-01
## spotify$key9              2.616e-01
## spotify$key10             3.063e-01
## spotify$key11             2.842e-01
## spotify$time_signature4   2.439e-01
## spotify$time_signature5   5.948e-01
## spotify$danceability      4.521e-01
## spotify$energy            5.276e-01
## spotify$loudness          2.258e-02
## spotify$speechiness       5.807e-01
## spotify$acousticness      2.965e-01
## spotify$instrumentalness  2.822e-01
## spotify$liveness          3.847e-01
## spotify$valence           3.091e-01
## spotify$tempo             2.048e-03
## spotify$duration_ms       6.707e-07
##                          z value
## (Intercept)                3.701
## spotify$key1              -2.888
## spotify$key2              -0.133
## spotify$key3              -4.083
## spotify$key4              -5.500
## spotify$key5              -3.154
## spotify$key6              -5.245
## spotify$key7              -1.018
## spotify$key8              -1.487
## spotify$key9              -4.302
## spotify$key10             -5.952
## spotify$key11             -6.883
## spotify$time_signature4    0.159
## spotify$time_signature5   -1.405
## spotify$danceability      -2.722
## spotify$energy            -1.744
## spotify$loudness           1.493
```

```
## spotify$speechiness       -1.284
## spotify$acousticness       1.828
## spotify$instrumentalness   0.619
## spotify$liveness           1.527
## spotify$valence            1.339
## spotify$tempo              1.078
## spotify$duration_ms       -1.362
##                          Pr(>|z|)
## (Intercept)              0.000214 ***
## spotify$key1             0.003873 **
## spotify$key2             0.894153
## spotify$key3             4.45e-05 ***
## spotify$key4             3.80e-08 ***
## spotify$key5             0.001612 **
## spotify$key6             1.57e-07 ***
## spotify$key7             0.308472
## spotify$key8             0.136974
## spotify$key9             1.69e-05 ***
## spotify$key10            2.65e-09 ***
## spotify$key11            5.86e-12 ***
## spotify$time_signature4  0.873929
## spotify$time_signature5  0.160015
## spotify$danceability     0.006497 **
## spotify$energy           0.081144 .
## spotify$loudness         0.135454
## spotify$speechiness      0.199155
## spotify$acousticness     0.067514 .
## spotify$instrumentalness 0.535942
## spotify$liveness         0.126753
## spotify$valence          0.180613
## spotify$tempo            0.280961
## spotify$duration_ms      0.173170
## ---
## Signif. codes:
##   0 '***' 0.001 '**' 0.01 '*' 0.05
##   '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be
   1)
##
##     Null deviance: 1869.0  on 1450  degrees of freedom
## Residual deviance: 1708.3  on 1427  degrees of freedom
## AIC: 1756.3
##
## Number of Fisher Scoring iterations: 4
```

The multiple logistic regression model makes minor predictions but it does not perform any better than the simple logistic regression model reporting 65% of songs correctly classified and a test error rate of 35%.

```
##
## log.pred   0   1
##        0 301 304
##        1 199 647

## [1] 0.6533425
```

Following the multiple logistic model, we will make use of the lasso to see if it will help our model create better predictions by shrinking some of the coefficients to exactly zero and reporting only the most significant variables in prediction. Below we report the coefficients the lasso method determines. As we can see, certain key signatures are significant. The coefficients given are associated with the key signatures D, D#, E, F#, G, A#, and B. Danceability, speechiness, and acousticness are also important variables in prediction. The multiple logistic regression model utilizing the lasso reports 67% of songs are correctly classified, having a test error rate of 33%. This is a slight improvement over the two previous models.

```
## 24 x 1 sparse Matrix of class "dgCMatrix"
##                                      1
## (Intercept)              1.09595056
## train$key1               .
## train$key2               0.32553088
## train$key3              -0.12320759
## train$key4              -0.32574878
## train$key5               .
## train$key6              -0.20390087
## train$key7               0.29509690
## train$key8               .
## train$key9               .
## train$key10             -0.54805070
## train$key11             -0.73339571
## train$time_signature4    .
## train$time_signature5    .
## train$danceability      -0.63145684
## train$energy             .
## train$loudness           .
## train$speechiness       -0.11976697
```

```
## train$acousticness      0.07957112
## train$instrumentalness  .
## train$liveness          .
## train$valence           .
## train$tempo             .
## train$duration_ms       .

## [1] 0.6701031
```

We note that some of the coefficients using the lasso are negative and some are positive. This can be interpreted as for negative coefficients, they indicate that the associated feature is related with a higher probablility that the song will be predicted as being in minor mode, whereas positive coefficients indicate that the associated feature is related with a higher probability that the song will be predicted as being in major mode.

The lasso brings the advantage of interpretability to a model. By shrinking some of the coefficients to exactly zero, we end up with a model with less features to interpret and only the features that are most important in making predictions. The validation set approach benefits greatly from a larger dataset, so the lasso may perform even better on this model if a larger test set was available.