Logistic regression and Bayes-classifier study of classification of songs to genres based on timbre, pitch and rhythm of the music signal

John Doe
Aalto University, P.O. Box 11100, FI-00076 Aalto, Finland
(Dated: December 2, 2017)

A genre of a song can be estimated based on its music signal's characteristics . In this work we use two classifiers, Bayes Classifier and Logistic classifier, to classify songs into one of ten possible genres. The two classifiers are trained against the training data and their performance is compared against each other. In this work we show that both classifiers perform much better compared to random guess. However their capability to classify all songs is clearly limited the accuracy for both classifiers being around 60%.

I. INTRODUCTION

An automatic music transcription, *i.e.*, notating a piece of music to a speficic genre, *e.g.*, Blues, dates back to 1970s when first attempts towards automatic music transcription were made[1]. Since then interest in automatic transcription of music has grown rapidly and various approaches, statistical methods, modelling human auditory system, have been applied to music transcription problem. However even today an expert human musician often beats a state-of-the-art automatic transcription system in accuracy.

Characteristics of music signal that are useful in classification of a song are *timbre*, *rhythm*, *pitch*, *loudness* and *duration*[1] from which the three first one, described below are used in this work.

- The timbre of the music can be most easily described as the factor which separates two sources of music from each other. For example if the same song is played by violin or a guitar the timbre is called the character which separates the violin from the guitar.
- The pitch is related to frequency scale of a song a can be defined as the frequency of the sine-wave fitted to target sound by human listener.
- The rhythm of the music can be described as arrangement of sounds as time flows.

In classification problem the object is classified into a certain class based on it's characteristics called features. A linear classifier does the classification by making a linear combination of the features and converting the resulting value into a class or a probability that the object belongs to given class. In logistic regression the feature vector of the object is transformed into a probality by taking a linear combination of features and mapping the result into interval $[0, 1]$ using a sigmoid function. The Bayes-classifier in contrast assumes that the feature vector is drawn from a multidimensional-Gaussian distribution. The posterior probability of the object belonging to a certain class is then obtained as a product of the prior of the class and the probability to sample the given feature vector from the multidimensional Gaussian distribution.

The paper is organized as follows. The used data-set and the computational methods are described in detail in Sec. II. In Sec. III the results for the both logistic regression- and Bayes-classifier are given. Sec. IV is a summary of the results and the differences between the two classifiers are discussed.

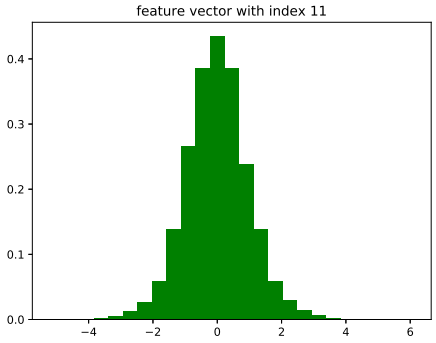
II. USED DATA-SET AND COMPUTATIONAL METHODS

A. Used data-set

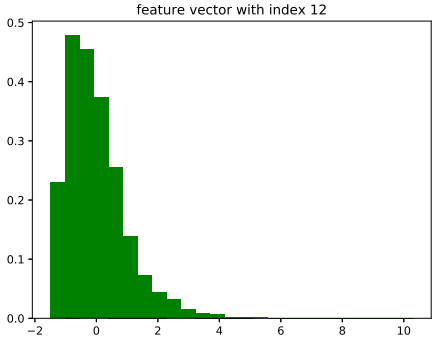
The data-set consisted of 4363 songs and was divided into training and test data sets including every third song to test set and rest of the songs to training set. Each song contained 264 features and the songs were labeled to 10 different categories. The gatecories were: 1 Pop Rock, 2 Electronic, 3 Rap, 4 jazz, 5 Latin, 6 RnB, 7 International, 8 Country, 9 Reggae and 10 Blues. The musical characteristics of the songs were packed to a feature vector of length 256. The first 48 elements in the feature vector can be associated to timbre, the next 48 elements to pitch and the final 168 features to rhythm. The distribution of the features resembled in most cases a Gaussian distribution or a skew symmetric distribution. This is illustrated figures 1a and 1b.

B. Computational methods

In this work two different methods were used to classify the songs to different genres. First method is logistic-regression method in which the logistic-loss is minimized iteratively using the gradient descent method. The other method used is the Bayes-classifier which classifies the song to certain category that gives the maximum posterior probability with respect to label i. Both methods are described below in detail. In addition we studied the effect of feature extraction and for that purpose we used principal component analysis method to exclude features with little impact.



(a)



(b)

FIG. 1: Visualization of typical distributions of features, Gaussian distribution (a) and skew-symmetric distribution (b).

1. *Logistic-regression*

In logistic regression for binary classifier problem the starting point is the minimization of the loss function

$$\mathcal{E}((\mathbf{x},y);\mathbf{w}) = \min_{\mathbf{w}\in\mathbf{R}^2} (1/N)\sum_{i=1}^{N} L((\mathbf{x}^{(i)},y^{(i)});\mathbf{w}),$$

where the logistic loss $L((\mathbf{x},y);\mathbf{w})$ is defined as $L((\mathbf{x},y);\mathbf{w}) = \ln\left(1+\exp\left(-y(\mathbf{w}^T\mathbf{x})\right)\right)$ and $\mathbf{x}$ is the feature vector of a music signal, $\mathbf{w}$ are the coefficients of the linear expansion and y is the label 1 or -1 whether the song belongs to certain category or not. The minimization problem can be further converted to

$$\begin{aligned}\mathcal{E}((\mathbf{x},y);\mathbf{w}) &= \frac{1}{N}\max_{\mathbf{w}\in\mathbf{R}^2}\sum_{i=1}^{N}\ln\left(\frac{1}{1+\exp(-y^i(\mathbf{w}^T\mathbf{x}))}\right)\\ &= \frac{1}{N}\max_{\mathbf{w}\in\mathbf{R}^2}\Big(\sum_{y^i=1}\ln\left(\frac{1}{1+\exp(-\mathbf{w}^T\mathbf{x}^{(y^i)})}\right)\\ &\quad+\sum_{y^i=-1}\ln\left(1-\frac{1}{1+\exp(-\mathbf{w}^T\mathbf{x}^{(y^i)})}\right)\Big) \quad (1)\\ &= \max_{\mathbf{w}\in\mathbf{R}^2}\Big(\sum_{y^i=1}\ln(p_{y^i=1})\\ &\quad+\sum_{y^i=-1}\ln(1-p_{y^i=1})\Big), \quad (2)\end{aligned}$$

where $p_{y^i=1}$ is the probability that the song i is labeled belonging to certain category. There is no closed form solution for equation (2) and for that reason some numerical iterative solver must be used to find the optimal $\mathbf{w}$. One of the most popular methods to find the optimal solution is gradient descent (GD) method. In GD the weights $\mathbf{w}$ are updated at each iteration $\mathrm{k}+1$ according to equation

$$\mathbf{w}^{(k+1)} = \mathbf{w}^{(k)} - \alpha\nabla\mathcal{E}(\mathbf{w}^{(k)}). \quad (3)$$

To be able to use the GD method we need know the gradients $\nabla\mathcal{E}(\mathbf{w}^{(k)})$. By marking $t^i = \max(0,y^i)$ we can write the equation (2) as

$$\begin{aligned}\mathcal{E}((\mathbf{x},y);\mathbf{w}) = \min_{\mathbf{w}\in\mathbf{R}^2} -\sum_{i}^{N}&(t^i\ln(p_{y^i})\\ &+(1-t^i)\ln(1-p_{y^i})). \quad (4)\end{aligned}$$

The derivative of p_{y^i} with respect to w_j can be written as $\frac{\partial p_{y^i}}{\partial w_j} = \frac{\partial p_{y^i}}{\partial(\mathbf{wx})}\frac{\partial(\mathbf{wx})}{\partial w_j} = p_{y^i}(1-p_{y^i})x_j^i$. Thus taking derivative of equation (4) with respect to w_j we get for $\nabla\mathcal{E}(w_j^k)$.

$$\begin{aligned}\nabla\mathcal{E}(w_j^k) &= -\sum_i\Big(t^i\frac{1}{p_{y^i}}(1-p_{y^i})p_{y^i}x_j^i\\ &\quad+(1-t^i)\frac{1}{1-p_{y^i}}(-1)p_{y^i}(1-p_{y^i})x_j^i\Big)\\ &= -\sum_i\left(t^i - t^ip_{y^i} - p_{y^i} + t^ip_{y^i}\right)x_j^i\\ &= -\sum_i(t^i-p_{y^i})x_j^i. \quad (5)\end{aligned}$$

Now the $\nabla\mathcal{E}(w_j^k)$'s can be written in vector form as

$$\nabla\mathcal{E}(\mathbf{w}^k) = -\left(\hat{\mathbf{y}} - \frac{1}{1+\exp(-(\mathbf{w}^k)^TX^T)}\right)^T X, \quad (6)$$

where $\hat{\mathbf{y}} = \max(\mathbf{0},\mathbf{y})$ is a vector of length N for which the i-th element is zero if $y_i = -1$ and one if $y_i = 1$. The matrix X is a matrix which rows are the feature

vectors $\mathbf{x}^i$. Thus we can solve the minimization problem (4) iteratively using the GD method (3) and the gradients (6). The parameters $\mathbf{w}$ are solved for each class i using the training data. The category i of a song in test-set is then the one with the largest probability $p(y_i)$.

2. Bayes classifier

For a Bayes classifier we assume that the distribution of the feature vector of a music signal with respect to label y_i is a Gaussian distribution

$$p(\mathbf{x}|y_i;\mathbf{m}_i,\mathbf{C}_i) = \frac{1}{\sqrt{\det\{2\pi\mathbf{C}_i\}}} e^{-(1/2)(\mathbf{x}-\mathbf{m}_i)^T\mathbf{C}_i^{-1}(\mathbf{x}-\mathbf{m}_i)}. \tag{7}$$

Using the Baye's theorem the posterior probability $p(y_i|\mathbf{x};\mathbf{m}_i,\mathbf{C}_i)$ can be written as

$$p(y_i|\mathbf{x};\mathbf{m}_i,\mathbf{C}_i) = \frac{p(y_i)p(\mathbf{x}|y_i;\mathbf{m}_i,\mathbf{C}_i)}{p(\mathbf{x})}, \tag{8}$$

where the $p(\mathbf{x})$ is a normalization constant and can be omitted. To be able to use the equation (8) we need to find optimal values for parameters $p(y_i)$, $\mathbf{m}_i$ and $\mathbf{C}_i$. The prior $p(y_i)$ can be simply estimated as the fraction of labels y_i among all labels. Because the samples $\mathbf{x}_i^{(t)}, y_i^{(t)}$ are independent the parameters $\mathbf{m}_i$ and $\mathbf{C}_i$ and can be obtained by maximizing the respective log-likelihood function with respect to the parameters. The log-likelihood can be written as

$$\begin{aligned}\mathcal{L}_{y=i} &= \sum_{y^t=i} \log(P(\mathbf{x}^t|y^t=i;\mathbf{m}_i,\mathbf{C}_i)) \\ &= \sum_{y^t=i}(-\frac{1}{2}\log(2\pi^n) - \frac{1}{2}\log(\det(\mathbf{C}_i)) \\ &\quad -\frac{1}{2}(\mathbf{x}^t-\mathbf{m}_i)^T\mathbf{C}_i^{-1}(\mathbf{x}^t-\mathbf{m}_i)).\end{aligned} \tag{9}$$

Now the optimal $\mathbf{m}_i$ is obtained by setting the derivative of $\mathcal{L}_{y=i}$ with respect to $\mathbf{m}_i$ to zero, *i.e.*,

$$\frac{\partial\mathcal{L}_{y=i}}{\partial\mathbf{m}_i} = \mathbf{C}_i^{-1}\sum_{y^t=i}(\mathbf{x}^t-\mathbf{m}_i). \tag{10}$$

Setting eq. (10) to zero and solving for $\mathbf{m}_i$ gives $\mathbf{m}_i = \frac{\sum_{y^t=i}\mathbf{x}^t}{N}$. Similarly the $\mathbf{C}_i$ is obtained by setting the derivative of $\mathcal{L}_{y=i}$ with respect to $\mathbf{C}_i$ to zero giving

$$\begin{aligned}\frac{\partial\mathcal{L}_{y=i}}{\partial\mathbf{C}_i} &= \sum_{y^t=1}(-\frac{1}{2}\mathbf{C}_i^{(-1)} \\ &\quad +\frac{1}{2}\mathbf{C}_i^{(-1)}(\mathbf{x}^t-\mathbf{m}_i)(\mathbf{x}^t-\mathbf{m}_i)^T\mathbf{C}_i^{-1}).\end{aligned} \tag{11}$$

Setting eq. (11) to zero and solving for $\mathbf{C}_i$ gives $\mathbf{C}_i = \frac{1}{N}\sum_{y^t=i}(\mathbf{x}^t-\mathbf{m}_i)(\mathbf{x}^t-\mathbf{m}_i)^T$. In equations (10) and (11) the identities $\frac{\partial(\mathbf{x}-\mathbf{s})^T\mathbf{W}(\mathbf{x}-\mathbf{s})}{\partial\mathbf{s}} = 2\mathbf{W}(\mathbf{x}-\mathbf{s})$, $\frac{\partial\ln|\det(\mathbf{X})|}{\partial\mathbf{X}} = (\mathbf{X}^T)^{-1}$ and $\frac{\partial\mathbf{a}^T\mathbf{X}^{-1}\mathbf{b}}{\partial\mathbf{X}} = -\mathbf{X}^{-T}\mathbf{a}\mathbf{b}^T\mathbf{X}^{-T}$ from matrix cook-book[2] were used. In classification the parameters $p(y_i)$, $\mathbf{m}_i$ and $\mathbf{C}_i$ are optimized for all ten classes using the training data. The song is then classified to certain category i that maximizes the posterior probability (8), *i.e.*,

$$i = \underset{i}{\operatorname{argmax}}\, p(y_i|\mathbf{x};\mathbf{m}_i,\mathbf{C}_i). \tag{12}$$

3. Principal component analysis

Let $\mathbf{X}$ be matrix which rows are the feature vectors, *i.e.*, the dimension of the matrix is $n \times p$ where n is the number of samples and p is the length of the feature vector. The sample covariance matrix $\mathbf{C}$ is then obtained as $\mathbf{C} = \mathbf{X}^T\mathbf{X}/(n-1)$ which can be diagonalized as

$$\mathbf{C} = \mathbf{V}\mathbf{L}\mathbf{V}^T, \tag{13}$$

where $\mathbf{V}$ are the eigenvectors of $\mathbf{C}$ and are called the principal axes. The matrix $\mathbf{X}$ can be decomposed using singular-value decomposition as $\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^T$, where $\mathbf{S}$ is a diagonal matrix containing the singular values of $\mathbf{X}$. Now the matrix $\mathbf{C}$ can be written as

$$\mathbf{C} = \mathbf{V}\frac{\mathbf{S}^2}{n-1}\mathbf{V}^T. \tag{14}$$

Comparing equation (14) to (13) shows that the principal axes are the same as the right singular vectors of $\mathbf{X}$. Now the principal components of $\mathbf{X}$ are obtained as $\mathbf{X}\mathbf{V} = \mathbf{U}\mathbf{S}$. Now we can include only feature vectors that have impact by excluding components that corresponds to singular values below some threshold. In this work we included singular values of which sum contained 90% of the total sum of the singular values and excluded rest.

III. RESULTS

Both the logistic-regression classifier and the Bayes classifier where trained using the training data set. After that the trained classifiers were applied to test data set and the accuracy and the logarithmic-loss of the classifiers were evaluated. In addition to test data set an external feature set with unknown labels were classified and the accuracy and the logarithmic-loss of classifiers were evaluated on an external server and results were compared to those obtained from the used test data set. The accuracy of the classifier was simply evaluated as the fraction of correct labels with respect to the total number of labels $accuracy = \frac{|y_{true}=y_{predicted}|}{N}$. The logarithmic loss was evaluated as $log-loss = 1/N\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}log(p_{ij})$ where y_{ij} is and indicator function evaluated to 1 if sample i was labeled to class j and zero otherwise and the p_{ij} is the corresponding class probability. The use logarithm is the base 10 logarithm.

TABLE I: Confusion matrix corresponding to classification obtained using logistic regression. The column direction indicates the true value and the row direction is the predicted value. The labels 1 . . . 10 are the ten music genres specified in section II A.

	1	2	3	4	5	6	7	8	9	10
1	652	35	7	9	4	10	1	3	2	3
2	57	130	9	4	2	1	0	2	2	0
3	14	9	82	4	2	1	0	0	1	0
4	25	3	0	43	0	2	0	1	0	2
5	38	4	1	5	11	1	1	0	3	0
6	37	6	12	8	4	25	0	0	0	0
7	34	6	1	2	2	4	3	1	0	0
8	50	0	0	1	2	1	0	9	0	0
9	2	5	0	2	1	0	0	8	0	27
10	21	0	0	6	1	2	0	1	0	3

A. Logistic-regression

The accuracy obtained using the logistic-regression and all elements of the feature vectors was 0.66 and the corresponding logarithmic-loss was 0.27. The accuracy and the logarithmic loss evaluated on the external server were 0.65 and 0.178 respectively. We also tested to use principal-component analysis to reduce the number of features. The corresponding accuracy and logarithmic-loss were 0.64 and 0.30 respectively being somewhat worse than without feature extraction. However it is good to notice that with feature extraction the computation was significantly faster. In table I the confusion matrix corresponding to full feature vector classification is shown.

B. Bayes classifier

For Bayes-classifier the sample covariance matrix C_i defined in section II B 2 becomes singular if all elements in the feature vector are used and for that reason classification was only performed alongside with the feature extraction. In feature extraction we selected features corresponding to singular values that counted 80% of the total sum of singular values and excluded rest. The obtained accuracy was 0.53 and the corresponding logarithmic-loss was 0.33. The accuracy and the logarithmic loss evaluated on the external server were 0.32 and 1.17 respectively significantly lower performance than with the test data. In table II the confusion matrix corresponding to Bayes-classifier classification is shown.

IV. CONCLUSIONS

In this work we used logistic-regression and Bayes-classifier to classify songs to different genres based on the music signal's characteristics. For logistic regression the obtained accuracy for test set was 0.67 and logistic-loss 0.27. For the external data set used the obtained accuracy and logistic-loss were 0.65 and 0.178 respectively in the case of logistic-regression. For the Bayes-classifier the obtained accuracy and logistic-loss were 0.53 and 0.33 for the test-data and for external data set 0.32 and 1.17 respectively. According to obtained results both classifiers performed clearly better than random guess, but remained far from perfect classification. From the two classifiers used the logistic-regression classifier performed clearly better. The logistic classifier also generalized much better to completely new data giving nearly equal performance for test data set and external data set.

TABLE II: Confusion matrix corresponding to classification obtained using Bayes-classifier. The column direction indicates the true value and the row direction is the predicted value. The labels 1 . . . 10 are the ten music genres specified in section II A.

	1	2	3	4	5	6	7	8	9	10
1	544	172	0	7	0	2	0	1	0	0
2	27	174	3	2	0	0	0	1	0	0
3	9	53	51	0	0	0	0	0	0	0
4	20	48	0	6	0	1	0	1	0	0
5	31	32	0	0	0	0	0	1	0	0
6	28	57	5	0	0	2	0	0	0	0
7	21	31	1	0	0	0	0	0	0	0
8	53	8	0	1	0	0	0	1	0	0
9	6	18	3	0	0	0	0	0	0	27
10	15	18	0	0	0	0	0	1	0	0

[1] A. Klapuri, *Introduction to Music Transcription* (Springer, New York, 2006).
[2] K. B. Petersen and P. M. S., *The Matrix Cookbook* (2012).