**Playing with Songs Genres Classification**

|  |  |
| --- | --- |
| Ziv Levy  The Interdisciplinary Center, Herzliya  Levy.ziv@post.idc.ac.il | Shachar Langbeheim  The Interdisciplinary Center, Herzliya  email@domain |

**Abstract**

**1. Introduction**

**2. Background**

**3. Related Work(?)**

**4. Logistic Regression**

In statistics, logistic regression is a type of regression analysis used for predicting the outcome of a categorical (a variable that can take on a limited number of categories) dependent variable based on one or more predictor variables. The probabilities describing the possible outcome of a single trial are modeled, as a function of explanatory variables, using a logistic function. Logistic regression measures the relationship between a categorical dependent variable and usually a continuous independent variable (or several), by converting the dependent variable to probability scores. We will use Logistic Regressionto predict if the genre of a given song. In addition to it, we will provide confidence percentage for that decision.

The logistic regression methodology is based on a training set. A training set is a set of pre-defined cases of instances that have already been scored. In our case, we will define a set, which include a subsets of genres such that each genre will contain a group of songs belong to the certain genre.

We will use [1] as our guidelines to build a proper logistic regression function.

We will define a presence vector in which all members of that vector are the number of occurrences of each word in the song. If a certain word exists in the given inspected song we will mark the number occurrences in its index, 0 otherwise.

Let Y be the result of the algorithm, if the song is considered as belong to the genre n, then Y=n. Moreover, we want to compute the probability that Y=n with a given set of seen words X. This is giving us a probability function:  
*𝑃(𝑌=1|𝑋)*

Moreover, as part of the algorithm result, we also have the confidence level of the result, which means, we have the ability to present how accurate and sure the result of the algorithm is. This is also a great feature that might be used in a later GUI implementation of this mechanism, and of course, for further calculations and statistical measurements.

**4.1. Fields and examples of applications**

Logistic regression is used extensively in numerous disciplines, including the medical and social science fields. For example, the Trauma and Injury Severity Score (TRISS), which is widely used to predict mortality in injured patients, was originally developed by Boyd et al. using logistic regression [3]. Logistic regression might be used to predict whether a patient has a given disease (e.g. diabetes), based on observed characteristics of the patient (age, gender, body mass index, results of various blood tests, etc.). Another example might be to predict whether an American voter will vote Democratic or Republican, based on age, income, gender, race, state of residence, votes in previous elections, etc. [4]. The technique can also be used in engineering, especially for predicting the probability of failure of a given process, system or product [5]. It is also used in marketing applications such as prediction of a customer's propensity to purchase a product or cease a subscription, etc. In economics it can be used to predict the likelihood of a person's choosing to be in the labor force, and a business application would be to predict the likehood of a homeowner defaulting on a mortgage.

In each of these instances, a logistic regression model would compute the relevant odds for each predictor or interaction term, take the natural logarithm of the odds (compute the logit), perform a linear regression on the predicted values of the logit, then take the exponential function of the logit to compute the odds ratio. Conditional random fields, an extension of logistic regression to sequential data, are used in natural language processing – and this is exactly our case!

**5. Limitations**

**6. Future Work**

**7. References**

1. LIBLINEAR: A Library for Large Linear Classification. Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, Chih-Jen Lin, Department of Computer Science, National Taiwan University, Taipei, Taiwan.
2. Categorical Data Analysis (Wiley Series in Probability and Statistics), Alan Agresti, John Wiley & Sons Publications.
3. Evaluating trauma care: the TRISS method. Trauma Score and the Injury Severity Score. Boyd CR, Tolson MA, Copes WS.
4. Logistic regression analysis for experimental determination of forming limit diagrams. M. Stranoa, B.M. Colosimob a Università di Cassino, Dip. Ingegneria Industriale, Italy, Politecnico di Milano, Dip. Meccanica, via Bonardi 9, Milano, Italy
5. Logistic regression model for prediction of roof fall risks in bord and pillar workings in coal mines. [Sanjay Kumar Palei](http://www.sciencedirect.com/science/article/pii/S0925753508000118), [Samir Kumar Das](http://www.sciencedirect.com/science/article/pii/S0925753508000118), Department of Mining Engineering, Indian Institute of Technology, India.