CSINTSY MCO3: Machine Learning

Members:

Asturiano, Christian Emmanuel S.
Cheng, Samuel Vincent T.
Custer, Mark John T.
De Ramos, Ghrazielle Rei A.

Submitted to:

Sir Thomas James Tiam-Lee

April 2023

I. Introduction

As humans, differentiating between the two sexes comes very naturally to us, but how can a machine identify the differences in appearance between males and females? Given that humans have incredibly many characteristics, it is very difficult to identify a formal set of rules for determining the gender of a person understandable by a computer. Machine learning has the special ability to allow computers to generate these rules on their own, given a sufficient data set. This makes it the most appropriate tool for the task.

It has been scientifically shown that a convenient way to recognize the difference between gender or sex is by using physical appearance as a basis. According to Hobgood (2015), The length of the hair may be a factor that identifies sex, long hair for females and short hair for males. Males commonly have broader shoulders than females. Males have significantly longer facial hairs than females. Also, males conventionally have wider and higher foreheads than females. Females also have plumper lips than males. The length between the lips to the nose differs between sex, where female lips are much closer to the nose than with males.

This project seeks to determine a person's gender based on their physical characteristics. This will be done by training 2 machine learning models using the data acquired from the data set. Determining a person's gender may be a subproblem in other, more important computational problems. Some examples are identifying a person from an image/video feed, analyzing a person's behavior, and automatic diagnosis of medical conditions.

II. The Dataset

https://www.kaggle.com/datasets/elakiricoder/gender-classification-dataset

The dataset contains seven features and one label. 2 features are numerical while the rest are binary. The label is the gender of a person. Each feature represents a physical characteristic of a person. The features are the following, the possible values of binary features are enumerated beside the name:

1. Hair length (long or not)
2. Forehead width in cm
3. Forehead height in cm
4. Nose width (wide or not)
5. Nose length (long or not)
6. Lip thinness (thin or not)

7. Distance between the nose and lips (long or not)

III. Methodology

The machine learning framework that was used for this project is scikit-learn. It allowed us to easily build and train our chosen machine learning models. It also provides a range of useful tools for data preprocessing, feature selection, and performance evaluation. We also used another library statsmodels solely for the purpose of automatically performing the Z-test.

Before applying the machine learning algorithms, some data exploration was done to determine which of the features in the data set are significant. Out of the 7 features in the data set, only 4 were fed into the machine learning models. We began the exploration by visualizing the data. After graphing each feature with respect to gender (see the jupyter notebook file) and looking at the data, we got the impression that some of the features appear to differ very significantly by gender.

We ran several statistical tests to confirm our intuition. All tests were done with a confidence interval of 95% For each of the 2 numerical features, we performed a z-test for difference in means, which returns whether the feature differs significantly by gender, i.e. whether the feature is related to gender. For each of the binary features, we used a chi-square test of independence, which determines whether there is a relationship between the feature and gender. The results of the tests revealed that only nose width, nose length, lip thickness, and distance between nose and lip were significantly related to gender.

The task of determining a person's gender based on their physical characteristics can be reframed as a classification problem: "Classify a person as male or female based on their characteristics". We chose logistic regression and Naive Bayes because they are both well-suited for binary classification tasks.

Logistic regression is a widely used statistical method for predicting binary outcomes. It works by modeling the relationship between the dependent variable (in this case, gender), and one or more independent variables (the selected features). Logistic regression predictions produce a probability value that represents the likelihood of a given input belonging to a particular category.

Similarly, Naive Bayes is also a probabilistic algorithm that models the probability of an input belonging to a particular class based on the probability of its features given that class. Both logistic regression and Naive Bayes are relatively simple and easy to implement, which is also

one of the reasons why we chose them. Additionally, they are both relatively efficient algorithms compared to other models and are commonly used as a standard.

IV. Results and Analysis

To test the models, we used k-fold cross-validation with k arbitrarily set to 5. To reduce the chance of overfitting, we opted for k-fold cross-validation as opposed to the basic method of splitting the data into one training set and one test set. The metrics used to evaluate the model's performances were accuracy, precision, and recall. The overall performance of a model evaluated with a particular metric was computed as the average of that metric among all the sub-tests conducted by k-fold cross-validation.

Accuracy was a reasonable choice for an evaluation metric since there is little class imbalance in the dataset (2500 males and 2501 females). The logistic regression model and the naive models had accuracies of 95.86% and 95.90%, respectively. Because the 2 models performed so similarly, we decided to use precision and recall to see whether their performances would be so similar again, and they were (see the jupyter notebook file for details).

We think we can attribute the strong performances of the models to two things. First is the fact that the data set contains features that are clearly statistically related to the target label. Second is the fact that we purposely only included these types of features. As such, the models were able to easily identify and learn the patterns and relationships between the features and the target label, resulting in a strong performance.

V. Conclusions and Recommendations

This project trained 2 machine learning models for the task of determining a person's gender based on several of their physical features. The models were trained on a data set acquired from kaggle, using the sci-kit-learn framework. The models performed well on all metrics used to evaluate their quality.

Since the quality of the models depends on the quality of the data set used, the project can be improved by improving the dataset. The set we used was relatively small, only having about 5000 entries. It would be better if a data set with more entries was used to feed the models. Another area in which the data set is somewhat lacking is in the number of features. There may be features that are very good predictors of gender that are not present in this data set. Given that a person has very many visible physical features, it should be possible to acquire enough data to expand the data set in this way.

VI. References

Issadeen, J. (2020). *Gender Classification Dataset*. (n.d.). Kaggle. Retrieved April 17, 2023, from https://www.kaggle.com/datasets/elakiricoder/gender-classification-dataset

Hobgood, T. (2015, August 17). *What Makes a Person's Face Look Masculine or Feminine?*. Hobgood Facial Plastic Surgery. Retrieved April 17, 2023, from https://www.toddhobgood.com/blog/what-makes-a-persons-face-look-masculine-or-feminine/

VII. Contributions of Each Members

ASTURIANO, Christian Emmanuel S.
Implemented the exploration of the data, and machine learning algorithms with scikit-learn in Jupyter Notebook.

CHENG, Samuel Vincent T.
Proposed and guided the portion of the data set exploration involving statistical analysis. Wrote parts 3 and 4 of the documentation.

CUSTER, Mark John T.
Worked on the documentation introduction, dataset description, and conclusions and recommendations.

DE RAMOS, Ghrazielle Rei A.
Worked on the documentation introduction, dataset description, and conclusions and recommendations