Naive Bayes Name-Gender Classifier

Hamza Sadaat, Usman Ali Awan, Omer Shoukat

Abstract:

A name is one of many features that a language has which functions as identifier of an individual. We could observe patterns emerge from names that could lead us to classify individuals. For example, we could tell whether a person is male or female just by knowing his/her names. Another classification that might be able to deduced from names are race, nationality, social class, etc. In this project, we focus on how to recognize gender solely from the knowledge of names. We employ two well-known machine learning classification algorithms, the Multinomial Naive Bayes (NB) and Random Forrest (RF), and compare both performances.

I. INTRODUCTION:

In machine learning, a Bayes classifier is a simple probabilistic classifier, which is based on applying Bayes' theorem. The feature model used by a naive Bayes classifier makes strong independence assumptions. NB is a family of simple probabilistic classifiers based on applying Bayes' Theorem with a naive assumption that each features/predictor are independent to each other. NB classifier shave worked quite well in many real-world situations. They require small amount if training data to estimate the necessary parameters, making NB classifiers can be extremely fast compared to other methods. We will see how this algorithm perform to classify gender in the next section.

II. Dataset used:

The dataset used for training and testing is dataset contains more than 50000 Pakistani names.
https://www.nrscotland.gov.uk/statistics-and-data/statistics/statistics-by-theme/vital-events/names/babies-first-names/babies-first-names-summary-records-comma-separated-value-csv-format.

III. CONCLUSION:

This project takes the gender names dataset from Kaggle (sourced nrs records) and uses it as training data for Naive Bayes classifier that probabilistically classifies lists of names into a likely split of the genders within by determining patterns within gendered names.