Developed country, developing country, or under developed country?

Lingfeng Cheng, Mihir Paradkar, Yuting Tian

September 24, 2016

Today, with the boom growth of technology and science, many countries are developing rapidly. However,the classification between developed countries and developing countries is facing challenges from longstanding and currently unresolved issues and new and emerging trends.

Firstly, although a few countries can be easily classified among the total 195 countries, the judgement between developed and developinig country for many other countries is still ambivalent. For example, even though UAE is categorized as a developing country, its people are already known for earning a remarkably high annual income. That is because the current criteria for quantitatively judging a specific country is Human Development Index (HDI), which only takes life expectancy, education, and personal income into consideration. However, many other features such as resource consumption, allocation, and investment scale are neglected. Moreover, the disadvantage of a dichotomous partition is that it ignores the different levels of development in the same category. For example, although both the U.S.and Italy are developed countries, the degree of development for these two countries are distinctive.

Therefore, we are trying to provide a new quantitative method to categorize every country and refine the simplistic dichotomous partition system.The dataset we chose is the World Development Indicators which is collected by the World Bank and updated quarterly. It covers a wide range of national aggregate data on topics such as agriculture, environment, education and infrastructure for every country and some aggregate regions in the world. The dataset is 245MB and is comprised of a large data file and several documentation and index files. Some of the entries for less well-known indicators are missing. Furthermore, the features are often very similarly defined and therefore highly correlated, such as GDP per capita and GNP per capita.

Given the current dataset, the proposed categorization problem can be treated as either a supervised learning or an unsupervised learning problem. From the supervised learning perspective, various features can be extracted firstly, and then,some well acknowledged developed and developing countries can be used as training data points. Therefore, a supervised learning model can be subsequently trained, which will ultimately be used to make predictions. From the unsupervised learning perspective, a clustering algorithm can be applied with each cluster assigned as a unique category of development degree. Additionally, to increase the accuracy of classification, PCA and other low-rank modeling techniques can be used to reduce the dimension of the data and impute missing values, using different regularizers to increase accuracy given a

classification model.