Machine Learning Group 70 Project Proposal

Team members:

NAME	ID
YIFAN PEI	20303063
MIN WU	20302508

Project Title:

How the proportion of different kinds of food intake(to get nutrition like fat, protein, and calories) influences obesity rate?

What problem are you tackling?

From what the percentage of a person/group intake from different food in their daily life to get the least nutrition supply(e.g. fat, protein, calories), we can predict the probability of getting obesity.

What data will you use and how will you collect it?

We will use 3 datasets including fat,protein,calories supply percentage from common food of different countries, e.g. in dataset Fat_Supply_Quantity_Data.CSV, the Afghanistani fat intake from Alcoholic Beverages(0%), Animal Products (21.6397%),Animal fats(6.2224%) and ...the obesity rate of Afghanistan is 4.5%. We collect the data from each country as a row because the food structure varies from different countries, and what people eat everyday is kind of similar of the same country, so we use the food supply of the same country as a row.

The data is from the US open data portal.

What machine learning techniques are you planning to apply or improve upon?

Since our mission can be generalized to a kind of prediction task, we would use linear model like Linear Regression, Ridge , lasso , elastic net , which implement three different ways to constrain the weights, by modifying the penalty and other parameters to find the best model.

Since the amount of features is relatively huge(23) not only does this make training extremely slow, it can also make it much harder to find a good solution, and the two main approaches to reducing dimensionality: projection and Manifold Learning, Principal

Component Analysis (PCA) is by far the most popular dimensionality reduction algorithm and by doing that we can visualize the dataset and fit it with the different models.

What experiments are you planning to run? How do you plan to evaluate your machine learning algorithm?

After pre-processing the input data, firstly reduce the dimension of the features , then we will have a look at the hyperparameters of the model and make some comparison to choose the best hyperparameters. For example, if we use Lasso, we need to make experiments to find the best C while with different degrees of polyfeatures. And we would use dummy model or just a line parallel axis as the baseline model.

For evaluation, we would use:

1)accuracy of each model

2)cross-validation to get the mean score and mean squared error with standard deviation, plot them separately in comparison

3) r2_score: $$R^2 = 1 - \frac{\sum_i \left(\hat{y}^{(i)} - y^{(i)}\right)^2}{\sum_i \left(\overline{y} - y^{(i)}\right)^2}$$ R Squared is a measurement that tells you to what extent the proportion of variance in the dependent variable is explained by the variance in the independent variables. In simpler terms, while the coefficients estimate trends, R-squared represents the scatter around the line of best fit.