CPSC429/529: Machine Learning

Program 1: K-Nearest Neighbor Model

1 Program Description

In this programming assignment, you are to implement a general k-NN, weighted k-NN (k should be parameterized, i.e., k could be any value), using Euclidean distance.

Specifically, you will use your models to predict the level of corruption in a country based on a range of macro-economic and social features. The data is given. Below lists the list of descriptive features (Columns 2-6 in the dataset):

- LIFE EXP.: the mean life expectancy at birth
- TOP-10 INCOME , the percentage of the annual income of the country that goes to the top 10% of earners
- INFANT MORT.: the number of infant deaths per 1,000 births
- MIL. SPEND: the percentage of GDP spent on the military
- SCHOOL YEARS: the mean number years spent in school by adult females

The target feature is the Corruption Perception Index (CPI) (The last column in the dataset). The CPI measures the perceived levels of corruption in the public sector of countries and ranges from 0 (highly corrupt) to 100 (very clean).

We will use Russia as our query country for this question. The table below lists the descriptive features for Russia.

COUNTRY ID	LIFE EXP.	TOP-10 INCOME	INFANT MORT.	MIL. SPEND	SCHOOL YEARS	CPI
Russia	67.62	31.68	10.00	3.87	12.90	?

1. What value would a 3-nearest neighbor prediction model using Euclidean distance return for the CPI of Russia?
2. What value would a weighted k-NN prediction model return for the CPI of Russia? Use $k = 16$ (i.e., the full dataset) and a weighting scheme of the reciprocal of the squared Euclidean distance between the neighbor and the query.

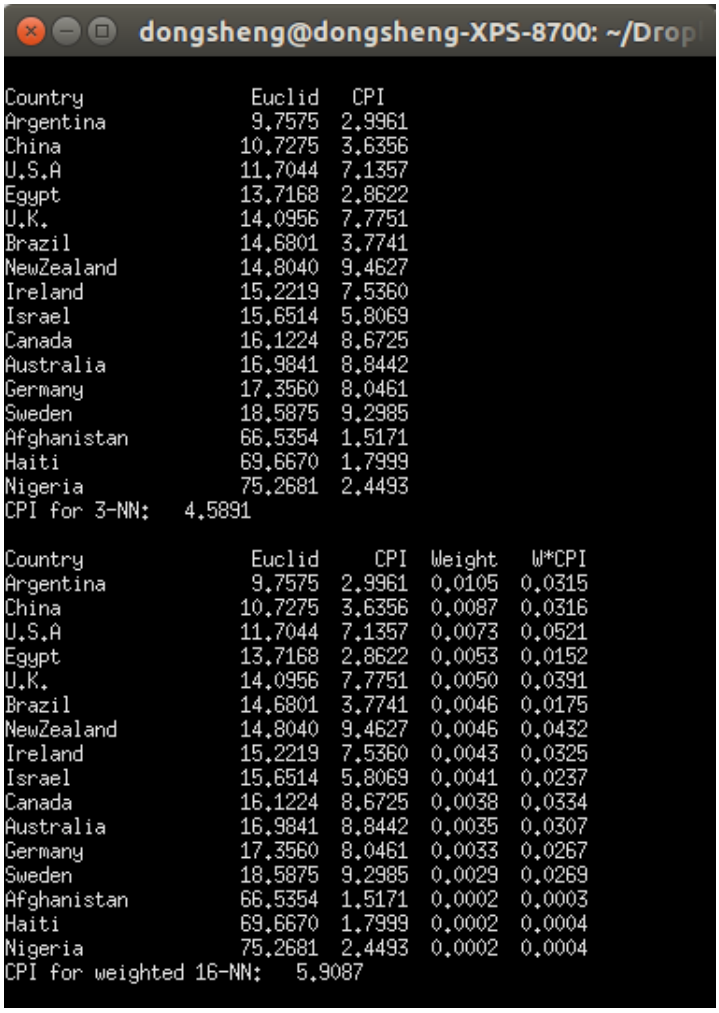

Figure 1: A screenshot of the first two outputs

3. The descriptive features in this dataset are of different types. For example, some are percentages, others are measured in years, and others are measured in counts per 1,000. We should always consider normalizing our data, but it is particularly important to do this when the descriptive features are measured in different units. What value would a 3-nearest neighbor prediction model using Euclidean distance return for the CPI of Russia when the descriptive features have been normalized using range normalization? (Hint: The normalized query is given as follows: ['Russia', 0.6099, 0.3754, 0.0948, 0.5658, 0.9058]

4. What value would a weighted k-NN prediction model—with k=16 (i.e., the full dataset) and using a weighting scheme of the reciprocal of the squared Euclidean distance between the neighbor and the query—return for the CPI of Russia when it is applied to the range-normalized data?

2 Useful Help

You should not use `scikit-learner KNN` for this program, but you are allowed to use `scikit-learner` for range normalization.

An online *Tutorial To Implement k-Nearest Neighbors in Python From Scratch* (See the link: `http://machinelearningmastery.com/tutorial-to-implement-k-nearest-neighbors-in-python-from-scratch/`) should be helpful to this program. Read the tutorial and understand how KNN can be used for predicting `Iris.data`.

You can use the code as the starting point for your program and modfiy based on that. Keep in mind there are lots of differences. Just name a few:

1. The target values of your problem are continuous, not discrete;
2. You do not need to split training and testing data;
3. You do not need to evaluate your prediction accuracy;
4. You need to normalize your data for question 3 and 4.

3 Submission

1. **Electronic submission**: Upload the following items on D2L dropbox, including:
   (a) The source code (.py code).
   (b) Program output file.
2. **Demo and submission** (Next class time after due date/time)
   (a) Demo your program.
   (b) Hand me your program outputs.