UNNC COMP3055: Machine Learning Coursework

Kezar Lake

Jae sung Park

Shyjp1@nottingham.edu.cn

20121762

26-12-20

School of Computer Science University of Nottingham Ningbo China

Environment

All codes are implemented in Python Scikit Learn with Jupyter Notebook. The attached zip file is containing both .ipynb and .py files.

Dataset

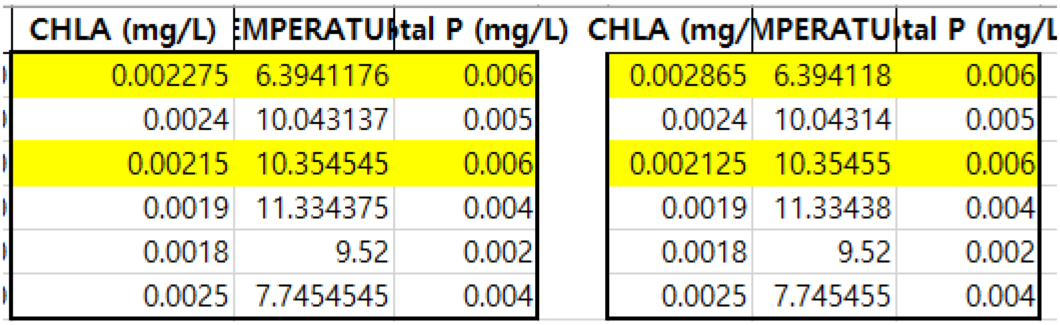
The Kezar Lake dataset has been selected and it consists of six variable features and 74 rows on the CHLA sheet, 2194 rows on Temperature, and 88 rows on Total P. The reason why I first chose this data is that it contained over 2,000 rows on Temperature. I found out there are similar numbers of rows on both CHLA and Total P among other datasets, so I decided to use Kezar Lake.

TASK1_Imputation

Firstly, there are not many data on the same depth level on the Kezar Lake dataset. I believed firsthand that if I were forced to choose one depth level and apply mean imputation on empty cells, the result might be worse. It was an adventure to choose this own method. The detailed explanation is on 'description.pdf'. Basically, Get the mean for depths and targeted items (CHLA, TEMPERATURE, TOTAL P) each month and make it into one row. After pre-processing the data, the mean is imputed on the unknown with the previous mean of those each month. But then, after realized the mean of depths and mean-imputed depths are useless on this specific coursework, I manually set depths as 0. I was tried to use every single data that I had.

Mean Imputation vs. KNN Imputation

The result seems similar as shown in Figure 1. The yellow-colored boxes are the ones imputed.



CHLA (mg/L)	EMPERATU	tal P (mg/L)
0.002275	6.3941176	0.006
0.0024	10.043137	0.005
0.00215	10.354545	0.006
0.0019	11.334375	0.004
0.0018	9.52	0.002
0.0025	7.7454545	0.004

CHLA (mg/	MPERATU	tal P (mg/L
0.002865	6.394118	0.006
0.0024	10.04314	0.005
0.002125	10.35455	0.006
0.0019	11.33438	0.004
0.0018	9.52	0.002
0.0025	7.745455	0.004

Figure 1. KNN Imputation(LEFT) vs. Mean Imputation(RIGHT)

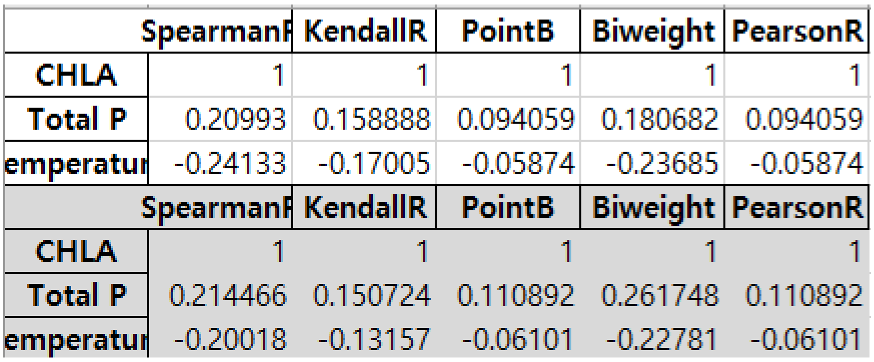
On KNN imputation, picking 'n_neighbors' as 4 brought the most similar results as result by mean imputation. Unlike own expectation that since a smaller number of closer data would bring an accurate result, decreasing number below 4 brought unusual result (for instance, 6.394 to 1.345).

	SpearmanR with Mean	SpearmanR with KNN
CHLA	1	1
Total P	0.20993	0.214466
Temperature	-0.24133	-0.20018

Figure2. Correlation between Mean vs. KNN on SpearmanR

Figure 2 is also showing a similar result on the correlation among targeted items and the result seems unsatisfying. Assuming it is because Mean and KNN imputation are processed on the data that is already mean on all different depths.

TASK2_5 Methods



	Spearman	KendallR	PointB	Biweight	PearsonR
CHLA	1	1	1	1	1
Total P	0.20993	0.158888	0.094059	0.180682	0.094059
emperatur	-0.24133	-0.17005	-0.05874	-0.23685	-0.05874
	Spearman	KendallR	PointB	Biweight	PearsonR
CHLA	1	1	1	1	1
Total P	0.214466	0.150724	0.110892	0.261748	0.110892
emperatur	-0.20018	-0.13157	-0.06101	-0.22781	-0.06101

Figure 3. Correlation among three targeted items on different methods

Spearman, Kendall, Pointbiserial, Biweight, and Pearson correlation testing methods are used to compare each result. It generally shows that as Chlorophyll a increases, total phosphorus increases. On the other hand, temperature decreases as CHLA increases. One interesting find is Pointbiserial and Pearson brought the same result. It is because a point-biserial correlation is simply the correlation between one dichotomous variable and one continuous variable which is the same as computing Pearson correlation on the same condition [1].

Among the five methods, the Biweight method showed the highest correlation. By weight, midcorrelation is median-based, rather than mean-based. Meaning, it is less sensitive to outliers. Since my testing dataset is containing all different depths, there must be more outliers than any other testing dataset that only used one specific depth.

Reflection

Overall, the method that pre-processed the data brought me the unsatisfying result. I was expecting the mean of more data on each month would bring a better result, but it was not. It is a shame that I could not manage to compare results between the same whole process on one depth level vs. the current result.

Reference

[1] Jason. Unkown. Point-biserial correlation, Phi, & Cramer's V. Retrieved from http://web.pdx.edu/~newsomj/pa551/lectur15.htm#:~:text=A%20point%2Dbiserial%20correlation%20is,variable%20and%20one%20continuous%20variable.&text=So%20computing%20the%20special%20point,and%20the%20other%20is%20continuous.