University of Nottingham

UK | CHINA | MALAYSIA

Machine Learning Coursework

Dataset: China Lake

Qichen An

20032255

Task 1: Complete the missing data

a. Data Pre-processing

Some steps are taken as follows:

Firstly, three columns are inserted to divide the data of date into year, month and day.

Date	Year	Month	Day
1998/5/13	1998	5	13
5/13/1998	1998	5	13
1998/5/27	1998	5	27
5/27/1998	1998	5	27
6/8/1998	1998	6	8
1998/6/24	1998	6	24
6/24/1998	1998	6	24

Then, according to the requirement, the data except from May to October was deleted.

Year	Month
1998	5
1998	6
1998	7
1998	8
1998	9
1998	10
1999	5
1999	6
1999	7
1999	8
1999	9
1999	10

Afterwards, three sheets are merged into one single sheet.

MIDAS	Lake	Town	Station	Date	Year	Month	Day	Depth	CHLA (mg/L)	TEMPERATURE (Centrigrade)	Total P (mg/L)
5448	China Lak	China, Va	1	1998/5/13	1998	5	13	7	0.00263		
5448	China Lak	China, Va	1	5/13/1998	1998	5	13	7		11	
5448	China Lak	China, Va	1	1998/5/27	1998	5	27	7	0.00494		
5448	China Lak	China, Va	1	5/27/1998	1998	5	27	7		15.6	
5448	China Lak	China, Va	1	6/8/1998	1998	6	8	7		15.4	
5448	China Lak	China, Va	1	1998/6/24	1998	6	24	7	0.00489		
5448	China Lak	China, Va	1	6/24/1998	1998	6	24	7		17.7	
5448	China Lak	China, Va	1	1998/6/24	1998	6	24	7			0.013
5448	China Lak	China, Va	1	1998/6/24	1998	6	24	7			0.014
5448	China Lak	China, Va	1	1998/7/8	1998	7	8	7	0.00352		
5448	China Lak	China, Va	1	7/8/1998	1998	7	8	7		19.2	
5448	China Lak	China, Va	1	1998/7/8	1998	7	8	7			0.012
5448	China Lak	China, Va	1	1998/7/8	1998	7	8	7			0.015
5448	China Lak	China, Va	1	1998/7/21	1998	7	21	7	0.00919		
5448	China Lak	China, Va	1	7/21/1998	1998	7	21	7		19.6	
5448	China Lak	China, Va	1	1998/7/21	1998	7	21	7			0.014
5448	China Lak	China, Va	1	1998/7/21	1998	7	21	7			0.011
5448	China Lak	China, Va	1	1998/8/4	1998	8	4	7	0.01531		
5448	China Lak	China, Va	1	8/4/1998	1998	8	4	7		21.5	
5448	China Lak	China, Va	1	1998/8/4	1998	8	4	7			0.016
5448	China Lak	China, Va	1	1998/8/4	1998	8	4	7			0.018
5448	China Lak	China, Va	1	8/19/1998	1998	8	19	7		22.7	

After merging all data into one sheet, there are still something need to do before it's ready to handle in Python program – discarding noisy and missing data.

Data of station (station 2, which only appears related CHLA).

MIDAS	Lake	Town	Station
5448	China Lake	China, Vas	2
5448	China Lake	China, Vas	2
5448	China Lake	China, Vas	2
5448	China Lake	China, Vas	2
5448	China Lake	China, Vas	2
5448	China Lake	China, Vas	2
5448	China Lake	China, Vas	2

Data of depth. All data without depth equal to 7 are removed as there is very few data without depth equal to 7.

MIDAS	Lake	Town(s)	Station	Date	Depth
5448	China Lake	China, Vas	1	1998/5/27	25
5448	China Lake	China, Vas	1	1998/6/8	8
5448	China Lake	China, Vas	1	1998/6/8	8
5448	China Lake	China, Vas	1	1998/6/8	16
5448	China Lake	China, Vas	1	1998/6/8	24
5448	China Lake	China, Vas	1	1998/6/24	7
5448	China Lake	China, Vas	1	1998/6/24	7
5448	China Lake	China, Vas	1	1998/6/24	15
5448	China Lake	China, Vas	1	1998/6/24	25
5448	China Lake	China, Vas	1	1998/7/8	7
5448	China Lake	China, Vas	1	1998/7/8	7
5448	China Lake	China, Vas	1	1998/7/8	16
5448	China Lake	China, Vas	1	1998/7/8	23
5448	China Lake	China, Vas	1	1998/7/21	7
5448	China Lake	China, Vas	1	1998/7/21	7

Data of some years. As there are very few data of year before 1998 and year 2004 and year 2016, these years' data are removed.

2	5448	China Lake	China, Vas	1	2003/10/28	14
3	5448	China Lake	China, Vas	1	2003/10/28	21
4	5448	China Lake	China, Vas	1	2004/4/30	6
5	5448	China Lake	China, Vas	1	2004/4/30	12
6	5448	China Lake	China, Vas	1	2004/4/30	21
7	5448	China Lake	China, Vas	1	2004/5/14	7
8	5448	China Lake	China, Vas	1	2004/5/14	8
9	5448	China Lake	China, Vas	1	2004/5/14	14
10	5448	China Lake	China, Vas	1	2004/5/14	21
11	5448	China Lake	China, Vas	1	2004/6/3	7
12	5448	China Lake	China, Vas	1	2004/6/3	7

Then, process the data in Python. During the process, some principles are followed: if in one month, there are one day or

some days containing all of CHLA, TEMPERATURE and TOTAL P, only these days are considered to be on behalf of this month. Otherwise, there is no day containing all three metrics, then all data of this month will be considered to be on behalf of this month.

```
year_ori = year = sheet1.cell_value(1, 5)
month_ori = month = sheet1.cell_value(1, 6)
day_ori = day = sheet1.cell_value(1, 7)

#每月数据
chla=tem=total=0
amount_chla = amount_tem = amount_total = 0

#每天数据
sub_chla = sub_tem = sub_total = 0
subAm_chla = subAm_tem = subAm_total = 0

#如果存在某天同时包含三项度量数据时的数据
exist_chla = exist_tem = exist_total = 0
existAm = 0
```

```
220                 sheet2.write(j, 2, chla / amount_chla)
221             if amount_tem != 0:
222                 sheet2.write(j, 3, tem / amount_tem)
223             if amount_total != 0:
224                 sheet2.write(j, 4, total / amount_total)
225         else:
226             sheet2.write(j, 2, exist_chla / existAm)
227             sheet2.write(j, 3, exist_tem / existAm)
228             sheet2.write(j, 4, exist_total / existAm)
229
230         f.save('Lake_v2.xls')
```

Finally, in order to proceed more smoothly for the next step, three months of containing no metrics (1998.10, 2006.9, 2012.8) are added into the sheet (Lake_v2).

48	2006	9			
49	2006	10	0.014	15.2	0.02

After pre-processing, the sheet ‘Lake_v2’ in the file ‘Lake_v2’ has been created.

b. Method 1: Mean Value

According to the file after pre-processing, there are no more than 100 lines of data samples. So mean value method of completing missing data in this part are mainly based on manual work and some built-in functions in excel. For the metrics from July to June, the value is calculated from the average of the two months near the month. For example, the data of June is calculated as the average of data of May and July. For May and October, the data of June and November respectively are used to fill in the missing part.

	A	B	C	D	E
1	Year	Month	CHLA	TEMPER	TotalP
2	1998	5	0.003785	13.3	0.0135
3	1998	6	0.00489	17.7	0.0135
4	1998	7	0.006355	19.4	0.013
5	1998	8	0.01531	21.5	0.017
6	1998	9	0.0251	20.8	0.01925
7	1998	10	0.0251	20.8	0.01925
8	1999	5	0.03042	14.8	0.022
9	1999	6	0.037445	17.2	0.0195
10	1999	7	0.00825	20.3	0.0185
11	1999	8	0.01551	20.03333	0.019667
12	1999	9	0.01551	18.8	0.019
13	1999	10	0.01551	14.3	0.02
14	2000	5	0.007	13.5	0.024
15	2000	6	0.0046	18.1	0.016
16	2000	7	0.00805	19.6	0.013
17	2000	8	0.0207	20.5	0.0165
18	2000	9	0.0176	19.1	0.0165
19	2000	10	0.0291	13	0.021
20	2001	5	0.0051	11.9	0.018
21	2001	6	0.0051	12.1	0.018
22	2001	7	0.0065	18.76667	0.014333
23	2001	8	0.01765	20.15	0.01475
24	2001	9	0.0127	18.8	0.017
25	2001	10	0.0127	16.1	0.017
26	2002	5	0.0058	9.7	0.0185
27	2002	6	0.0031	15.45	0.01425
28	2002	7	0.0043	21.7	0.01325
29	2002	8	0.0173	21	0.017

c. Method 2: Random Forests

The code of method Random Forests are in the file ‘randomForest.py’. And the result is generated in the file ‘Lake_v3.xls’. Then copy these data into sheet Random Forest of file ‘Lake_v2.xls’.

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean/average prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set. Random forests generally outperform decision trees, but their accuracy is lower than gradient boosted trees. However, data characteristics can affect their performance.

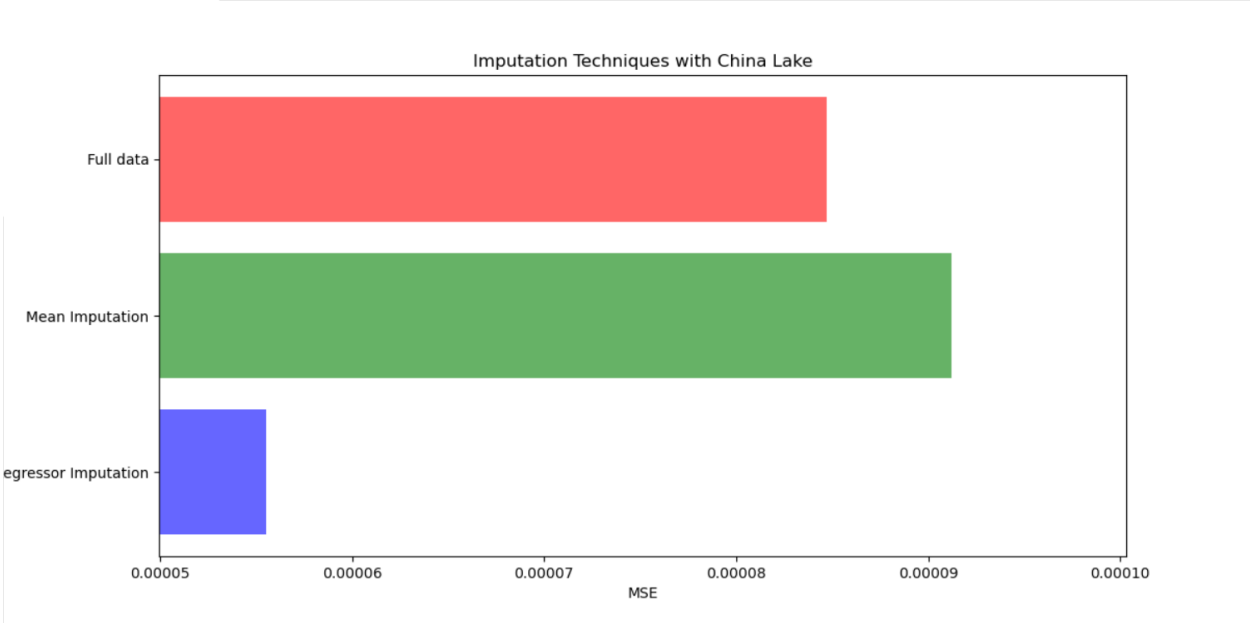
	A	B	C	D	E
1	Year	Month	CHLA	TEMPERATURE	TotalP
2	1998	5	0.003785	13.3	0.020100083
3	1998	6	0.00489	17.7	0.0135
4	1998	7	0.006355	19.4	0.013
5	1998	8	0.01531	21.5	0.017
6	1998	9	0.0251	20.8	0.01925
7	1998	10	0.025906935	14.0819	0.0202955
8	1999	5	0.03042	14.8	0.022
9	1999	6	0.037445	17.2	0.0195
10	1999	7	0.00625	20.3	0.0185
11	1999	8	0.01551	20.03333333	0.019666667
12	1999	9	0.020899778	18.8	0.019
13	1999	10	0.025740842	14.3	0.02
14	2000	5	0.007	13.5	0.024
15	2000	6	0.0046	18.1	0.016
16	2000	7	0.00805	19.6	0.013
17	2000	8	0.0207	20.5	0.0165
18	2000	9	0.0176	19.1	0.0165
19	2000	10	0.0291	13	0.021
20	2001	5	0.006542313	11.9	0.021131042
21	2001	6	0.0051	12.1	0.018
22	2001	7	0.0065	18.76666667	0.014333333
23	2001	8	0.01765	20.15	0.01475
24	2001	9	0.0127	18.8	0.017
25	2001	10	0.022625575	15.1	0.020816875
26	2002	5	0.0058	9.7	0.0185
27	2002	6	0.0031	15.45	0.01425
28	2002	7	0.0043	21.7	0.01325
29	2002	8	0.0173	21	0.017

d. Analysis

The comparation and analysis utilize the measure Mean squared error (MSE).

In statistics, MSE of an estimator (of a procedure for estimating an unobserved quantity) measures the average of the squares of the errors—that is, the average squared difference between the estimated values and the actual value. MSE is a risk function, corresponding to the expected value of the squared error loss. The fact that MSE is almost always strictly positive (and not zero) is because of randomness or because the estimator does not account for information that could produce a more accurate estimate.

The MSE is a measure of the quality of an estimator—it is always non-negative, and values closer to zero are better.



It can be seen from the results that the result for the random forest is the best, followed by the original data itself.

Task 2

a. Method

There are many kinds of methods to measure the correlation between variants. In this project, five methods of pearson, spearman, kendall, bicor and skipped are used with the package ‘pingouin’. ’pingouin’ provides functions to implement these correlation measures directly. into the excel file together to show their correlation. The code of these methods is in the file ‘correlation.py’. And the result is generated in the file ‘correlation.xls’. Then copy these data into sheet Correlation of file ‘Lake_v2.xls’.

```
data = pd.read_excel(io, sheet_name=2, usecols=[2,3,4])
data.head()
print(len(data))
for i in range(len(data)):
    print(data.loc[i])

corr=pg.pairwise_corr(data, method='pearson')
spearman_corr=pg.pairwise_corr(data, method='spearman')
kendall_corr=pg.pairwise_corr(data, method='kendall')
bicor_corr=pg.pairwise_corr(data, method='bicor')
skipped_corr=pg.pairwise_corr(data, method='skipped')
corr=corr.append(spearman_corr)
corr=corr.append(kendall_corr)
corr=corr.append(bicor_corr)
corr=corr.append(skipped_corr)

corr.to_excel('correlation.xls')
```

b. Analysis

X	Y	method	n	r
CHLA	TEMPERA	pearson	90	0.348164
CHLA	TotalP	pearson	90	0.402555
CHLA	TEMPERA	spearman	90	0.315987
CHLA	TotalP	spearman	90	0.514534
CHLA	TEMPERA	kendall	90	0.213955
CHLA	TotalP	kendall	90	0.360364
CHLA	TEMPERA	bicor	90	0.306763
CHLA	TotalP	bicor	90	0.467023
CHLA	TEMPERA	skipped	90	0.315987
CHLA	TotalP	skipped	90	0.527157

B1. In statistics, the Pearson correlation coefficient, also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or the bivariate correlation, is a statistic that measures linear correlation between two variables X and Y. It has a value between +1 and −1. A value of +1 is total positive linear correlation, 0 is no linear correlation, and −1 is total negative linear correlation. Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations.

B2. In statistics, Spearman's rank correlation coefficient or Spearman's is a nonparametric measure of rank correlation (statistical dependence between the rankings of two variables). It assesses how well the relationship between two variables can be described using a monotonic function. The Spearman correlation between two variables is equal to the Pearson correlation between the rank values of those two variables; while Pearson's correlation assesses linear relationships, Spearman's correlation assesses monotonic

relationships (whether linear or not). If there are no repeated data values, a perfect Spearman correlation of +1 or −1 occurs when each of the variables is a perfect monotone function of the other.

B3. In statistics, the Kendall rank correlation coefficient, commonly referred to as Kendall's coefficient, is a statistic used to measure the ordinal association between two measured quantities. It is a non-parametric hypothesis test for statistical dependence based on the τ coefficient. It is a measure of rank correlation: the similarity of the orderings of the data when ranked by each of the quantities. Intuitively, the Kendall correlation between two variables will be high when observations have a similar (or identical for a correlation of 1) rank (i.e. relative position label of the observations within the variable: 1st, 2nd, 3rd, etc.) between the two variables, and low when observations have a dissimilar (or fully different for a correlation of −1) rank between the two variables.

B4: In statistics, biweight midcorrelation (also called bicor) is a measure of similarity between samples. It is median-based, rather than mean-based, thus is less sensitive to outliers, and can be a robust alternative to other similarity metrics, such as Pearson correlation or mutual information.

B5: Numerous robust correlation coefficients have been proposed that deal with outliers among the marginal distributions, but these methods do not take into account the overall structure of the data in terms of dealing with outliers. A skipped correlation addresses this concern and methods for testing the hypothesis that this correlation is zero have been

studied. However, there are serious limitations associated with one of these methods and extant studies regarding an alternative percentile bootstrap method do not address practical concerns reviewed.