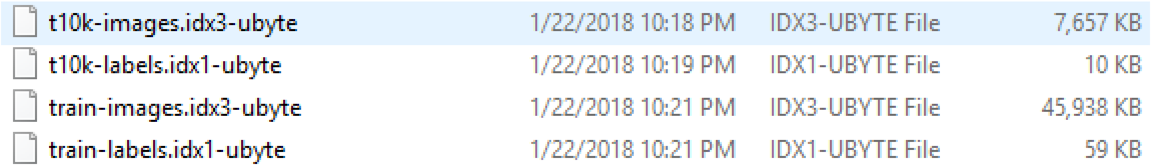
Assignment 1 Report by Muxin Liang

Problem 1. MNIST DataSet

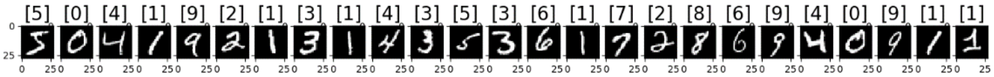
a) **Download Data**



t10k-images.idx3-ubyte	1/22/2018 10:18 PM	IDX3-UBYTE File	7,657 KB
t10k-labels.idx1-ubyte	1/22/2018 10:19 PM	IDX1-UBYTE File	10 KB
train-images.idx3-ubyte	1/22/2018 10:21 PM	IDX3-UBYTE File	45,938 KB
train-labels.idx1-ubyte	1/22/2018 10:21 PM	IDX1-UBYTE File	59 KB

b) **Visualization**

- Plot 25 images: I think I can recognize all the 9's in this dataset, though they look a little bit different in writing styles.



- Plot random 15 images: I guessed these 15 random images and checked with the labels and they are all correct

0
20
0 25 0 25 0 25 0 25 0 25 0 25 0 25 0 25 0 25 0 25 0 25 0 25 0 25 0 25 0 25

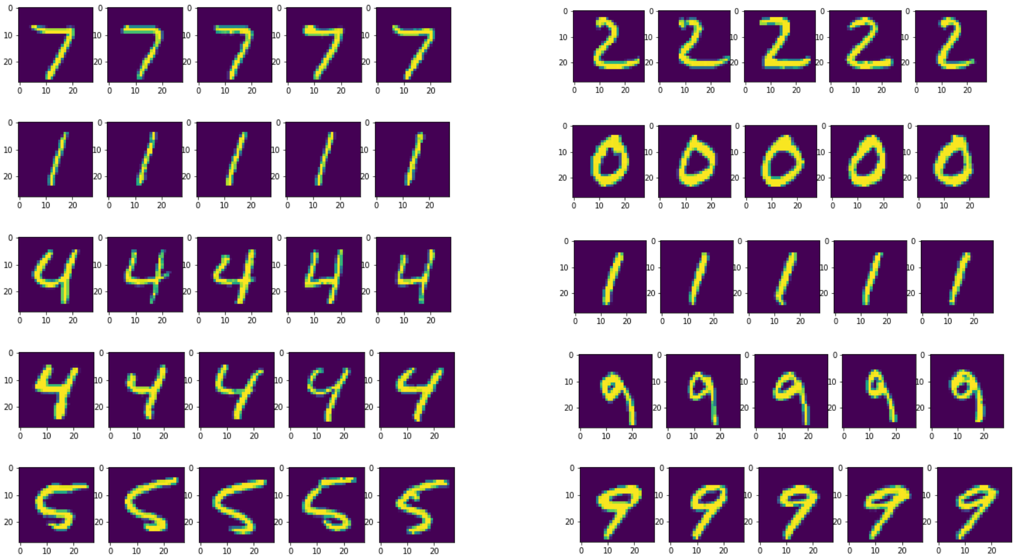
- Explore:
  Does not look alike: There are 5 "1"'s in the first figure, and we can notice that the 1st "1" and the 2nd "1" slash to the different side, and the last "1" have a different writing style

  Look alike: Take some notice on the first number (5) and the 7th number (3), they are different numbers, but they look similar in some degrees.

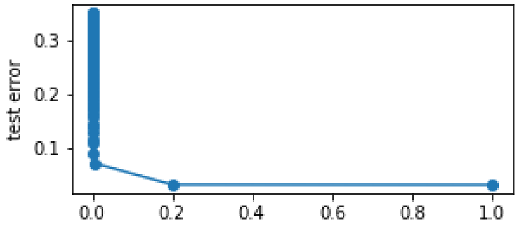
c) Classification using KNN on mnist

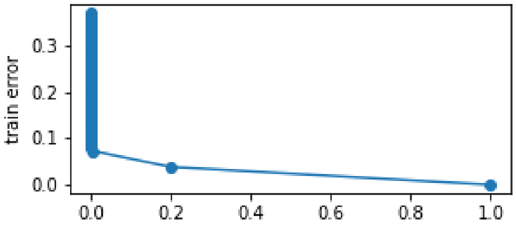
- What is the nearest neighbor?
  The nearest neighbor is the sample in the trainset that has the least distance to the given sample

- Find 5 nearest neighbor
  I used scikit-learn packages for applying nearest neighbor, the following plots are the 5-nearest neighbors for the first 10 samples in testing data from train data. (The codes are in the jupyter notebook)

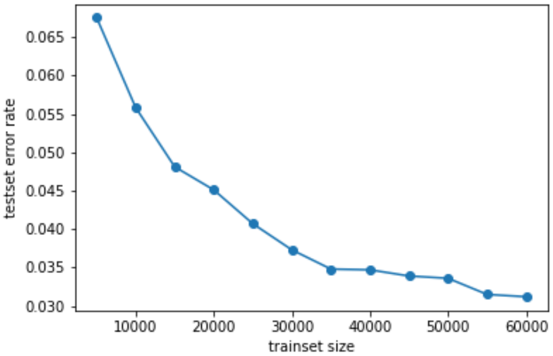


- The plotted result for train and test on k
  I tried all the prediction on different k, and I found that the lowest error rate occurs when k =1 for this query. The errors are increasing together with the increment of k. I guess it is the problem of over fitting. According to some research, the best k should be in the range of (0,100). You can see the test error list and train error list in the Jupyter Notebook. Because I made the k into subsets and I printed them in the notebook so I calculated all the recorded manually afterwards. However, k=1 is not the highest accuracy point. For example, when k = 5, accuracy reaches at almost 96%. Therefore, in the following questions, I will choose k = 5 as my parameter for KNN.



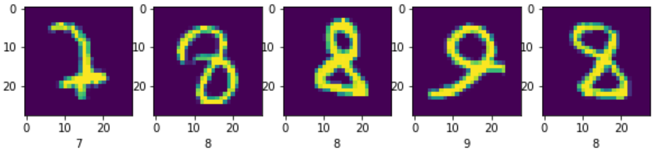


- The plotted result for varied sizes of train and testset (K = 5), I assuming it asks for test error rate (I did not find the definition about best error rate). As can be seen, there is a significant reduction in error with the increment of train set size.

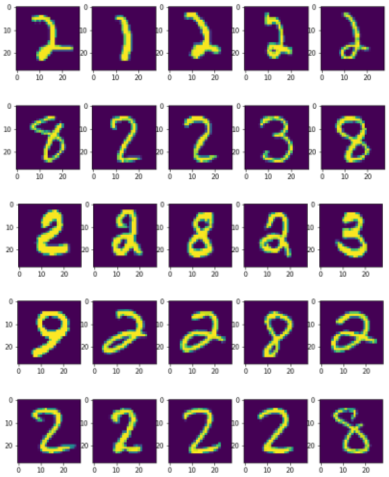


d) Plot for some misclassified samples

Here, I searched for some samples that are wrongly predicted to “2”, the characters below show their original values on test_labels. And I will search for their nearest



And for these figures, their nearest neighbors are:



e) Using different distance:

And the result is below (The train set size is 10000 and test set size is 5000):

Distance Type	Error Rate (%)
Euclidean	7.78
Manhattan	9.02
Chebyshev	91.88
log10(p) = 0.1	8.5
log10(p) = 0.2	8.08
log10(p) = 0.3	7.8
log10(p) = 0.4	7.54

log10(p) = 0.5	7.32
log10(p) = 0.6	6.96
log10(p) = 0.7	6.78
log10(p) = 0.8	6.56
log10(p) = 0.9	6.58
log10(p) = 1.0	6.58

For Mahalanobis and Hausdroff Distance, I tried to compute the distance metrics, but I failed in transform the input data into metrics. I hope I can get the answers about how to do that after the assignment.

f) Using difference polling decisions

I used inversed distance polling as another polling way: which means every neighbor is multiplied by a weighted parameter determined from the distance to the sample.

And the result is below (The train set size is 10000 and test set size is 5000):

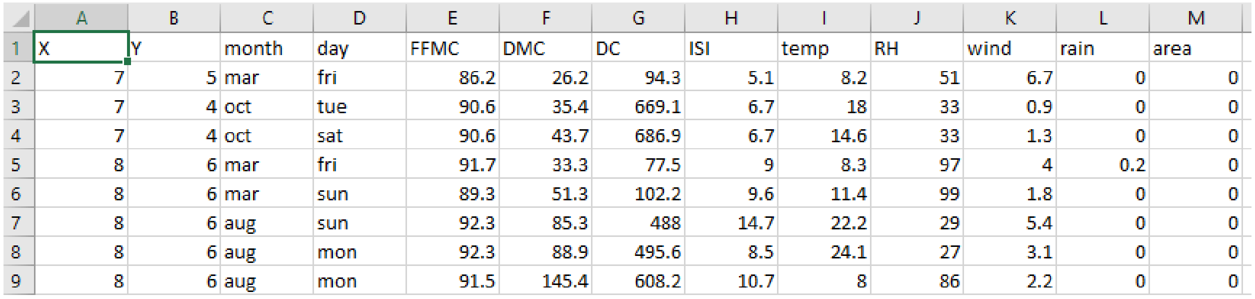
	Test Error for Majority Polling (%)	Test Error for Inversed Distance Polling (%)
Euclidean	7.78	7.28
Manhattan	9.02	8.42
Chebyshev	91.88	91.8

As can be seen from the chart, there is a little bit increment in accuracy when change polling decisions to inversed distance. And the best distance is Euclidean, which gives a 7.28% error rate.

g) Lowest error rate is 0%, it occurs when I used trainset itself as the testset for k=1, Euclidean distance. However, this result is trivial because under k=1, the nearest neighbor of a sample is itself. The best accuracy for testing set is when k = 5, using 60000 samples in trainset and 10000 samples in test set. Its accuracy reaches 96.88% (Error rate: 3.12%)

Problem 2. Forest Fire Data

a) Downloaded Dataset



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	X	Y	month	day	FFMC	DMC	DC	ISI	temp	RH	wind	rain	area
2	7	5	mar	fri	86.2	26.2	94.3	5.1	8.2	51	6.7	0	0
3	7	4	oct	tue	90.6	35.4	669.1	6.7	18	33	0.9	0	0
4	7	4	oct	sat	90.6	43.7	686.9	6.7	14.6	33	1.3	0	0
5	8	6	mar	fri	91.7	33.3	77.5	9	8.3	97	4	0.2	0
6	8	6	mar	sun	89.3	51.3	102.2	9.6	11.4	99	1.8	0	0
7	8	6	aug	sun	92.3	85.3	488	14.7	22.2	29	5.4	0	0
8	8	6	aug	mon	92.3	88.9	495.6	8.5	24.1	27	3.1	0	0
9	8	6	aug	mon	91.5	145.4	608.2	10.7	8	86	2.2	0	0

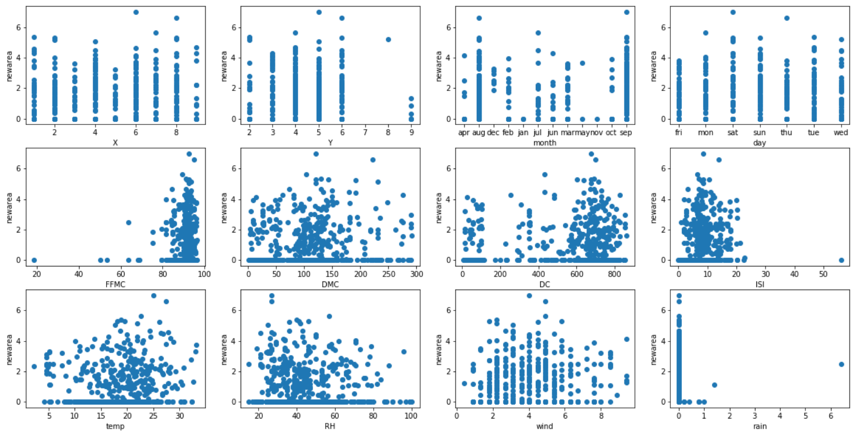
b) Exploring the data:

- There are 13 columns and 517 rows in this data set. Every row is an individual instance, and every column represents the value of a variable (column M is the output variable and other columns are the input variables).

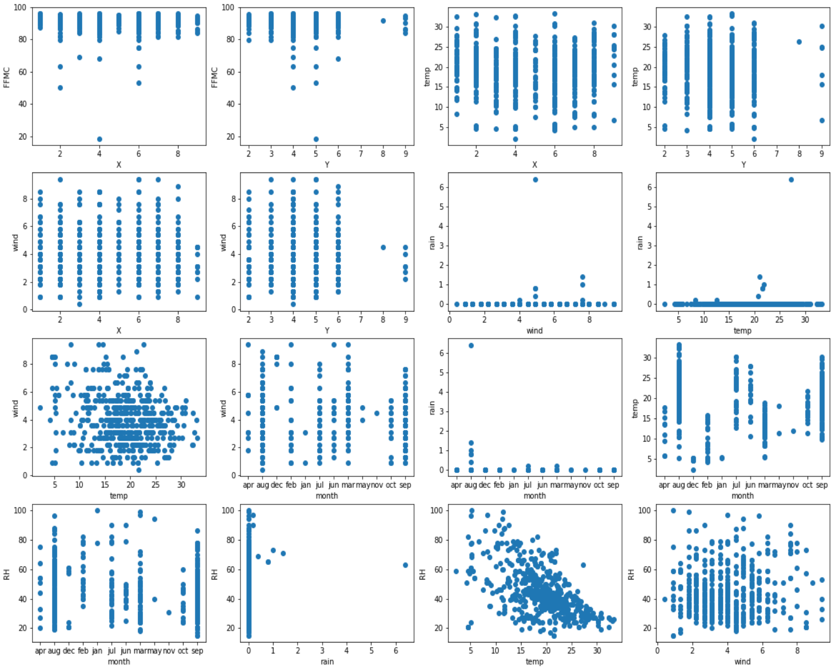
From the webpage, we can find the meaning of every column:

1. X - x-axis spatial coordinate within the Montesinho park map: 1 to 9
2. Y - y-axis spatial coordinate within the Montesinho park map: 2 to 9
3. month - month of the year: 'jan' to 'dec'
4. day - day of the week: 'mon' to 'sun'
5. FFMC - FFMC index from the FWI system: 18.7 to 96.20
6. DMC - DMC index from the FWI system: 1.1 to 291.3
7. DC - DC index from the FWI system: 7.9 to 860.6
8. ISI - ISI index from the FWI system: 0.0 to 56.10
9. temp - temperature in Celsius degrees: 2.2 to 33.30
10. RH - relative humidity in %: 15.0 to 100
11. wind - wind speed in km/h: 0.40 to 9.40
12. rain - outside rain in mm/m2 : 0.0 to 6.4
13. area - the burned area of the forest (in ha): 0.00 to 1090.84

- Because the output data(area) Y is very skewed (range from 0 to 1000+), using transformation $Y_1 = \ln(1+Y)$ will tense the output (generally in a range of [0, 7]), making it easier to manipulate. I called Y1 “newarea” in my data frame.
- The pairwise scatterplot of predictors is shown below:
  Some findings: for some predictors, Y1 is normally distributed (month, day), which means there is no significant influence from predictors to Y1. However, for some parameters, such as rain. Most of fire happens when there is no rain (sounds reasonable). This also applies for parameters FFMC and ISI.



- 16 pairwise scatter plot of predictors are shown below:
  The pairs I chose were: [("X","FFMC"), ("Y", "FFMC"), ("X" , "temp"),("Y", "temp"), ("X", "wind"), ("Y", "wind"), ("wind”, “rain"), ("temp", "rain"), ("temp", "wind"), ("month", "wind"), ("month", "rain"),("month", "temp"), ("month", "RH"), ("rain", "RH"), ("temp", "RH"), ("wind", "RH")]
  For example, I can find the range of the temperature and humidity corresponding to every month. And I also find that rain is more likely to happen when temperature is high. And most of the rain happens in August.



- The figure below shows the values asked:

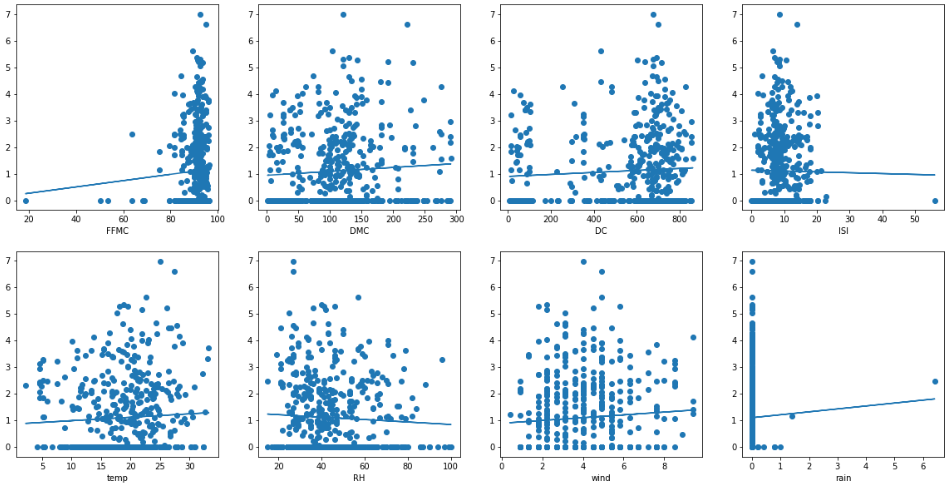
Variable	Mean	Median	Range	1st Quantile	3rd Quantile	Interquantile Ranges
X	4.669246	4	8	3	7	4
Y	4.299807	4	7	4	5	1
month	7.475822	8	11	7	9	2
day	4.259188	5	6	2	6	4
FFMC	90.64468	91.6	77.5	90.2	92.9	2.7
DMC	110.8723	108.3	290.2	68.6	142.4	73.8
DC	547.94	664.2	852.7	437.7	713.9	276.2
ISI	9.021663	8.4	56.1	6.5	10.8	4.3
temp	18.88917	19.3	31.1	15.5	22.8	7.3
RH	44.2882	42	85	33	53	20
wind	4.017602	4	9	2.7	4.9	2.2
rain	0.021663	0	6.4	0	0	0
area	12.84729	0.52	1090.84	0	6.57	6.57
newarea	1.111026	0.41871	6.99562	0	2.024193067	2.024193067

c) Linear Regression

X, Y, Month and Day are qualitative data and other are quantitative data, by looking at OLS report's p-values, none of those p-value for quantitative data is small enough to reject null hypothesis.

For qualitative data, I find that X = 3 has a p-value for 0.029 and month = December has a p-value for 0.024. They are possible to have linear relationship towards Y1.

I plotted the result of those quantitative data against Y1, it shows that there is no significant linear relationship between them and Y1 separately. I take X, Y, month, day as categorical data. (X, Y are not continuous variables and I think every number has some special meaning in geography, so I choose them as categorical)



All the regressions for quantitative variables should be rejected because none of the single parameter shows p-value that small enough to be keep as linear fit.

d) Multivariable regression:

By generating the report from multiple variable fitting from OLS and take X, Y, month, day as categorical data. Detailed report can be seen from Jupyter notebook.

The following are the predictors that can reject null hypothesis:

X = 3, p-value = 0.015, coefficient = -0.7342;

X = 9 p-value = 0.008, coefficient = 1.5756;

Y = 4 p-value = 0.06, coefficient = 0.5071;

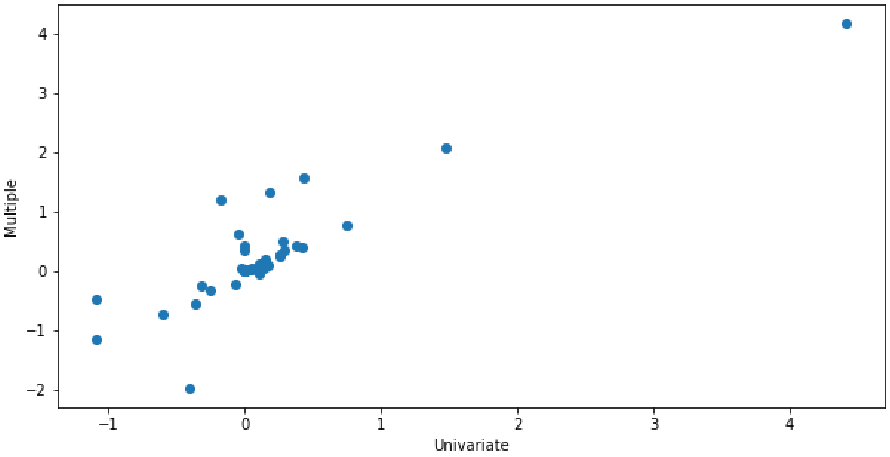
Y = 8 p-value = 0.004, coefficient = 4.1815;

Y = 9 p-value = 0.016, coefficient = -1.9851;

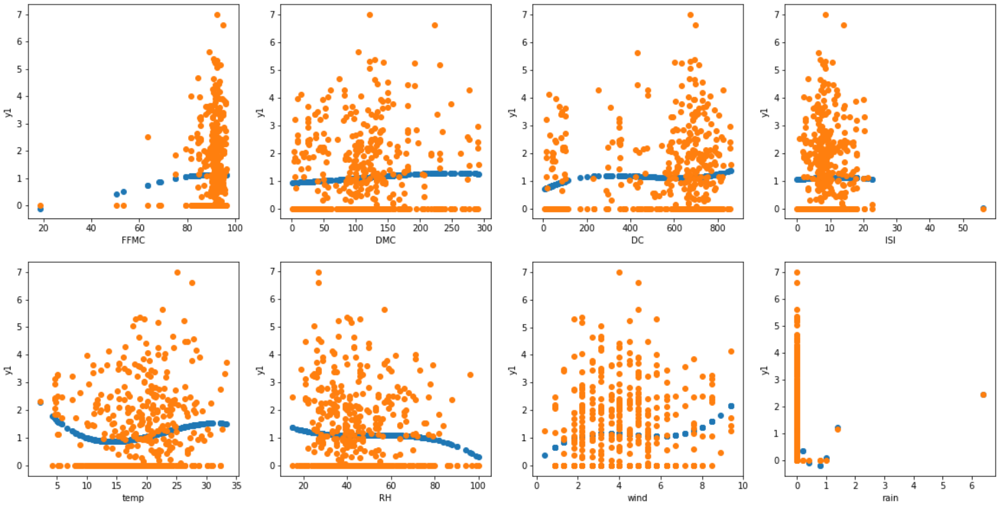
month =Dec p-value = 0.010, coefficient = 2.0877;

and quantitative variables: DMC p-value = 0.017, coefficient = 0.0045;

e) According to the report, there are total of 39 variables' coefficients in regression and we plot them together with univariate regression on x-axis and multiple regression on y-axis:



f) I used Curve-fit to get the coefficient of the 8 quantitative variables. The blue points are fitted curve, and the orange points are original data points. And I also did OLS analysis on those curves. Detailed report can be generated from Jupyter Notebook.



For $Y_1 = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \varepsilon$, I found some variables can reject null hypothesis based on the reports.

temp: with p-value for parameter from β_1 *to* β_3: 0.018, 0.043, 0.107, and forms a regression formula of $Y_1 = 2.9565 - 0.3461X + 0.0177X^2 - 0.0003\,X^3 + \varepsilon$

wind: with p-value for parameter from β_1 *to* β_3: 0.058, 0.055, 0.045, and forms a regression formula of $Y_1 = 0.1055 - 0.7721X - 0.1787X^2 + 0.0127\,X^3 + \varepsilon$

g) Association in interaction:

There is a total of 66 possible interactions for 12 predictors and I check against all of them (Code in Jupyter notebook):

And I checked among all Interactions' p-value, and those predictors provides a little bit significance in interactions:

X==5 and DC, coefficient = -0.0036 p-value = 0.0039
Y==4 and wind, coefficient = 0.2516 p-value = 0.034
Y==8 and wind, coefficient = 0.9554 p-value = 0.001
month = may and temp, coefficient = 0.6857 and p-value = 0.028
month = mar and wind, coefficient =-0.5823 and p-value = 0.014
month = may and wind, coefficient =-4.5927 and p-value = 0.036
DMC and temp, coefficient = 0.0004 and p-value = 0.028
DC and temp, coefficient =0.0001 and p-value = 0.042
temp and wind, coefficient = -0.0134 and p-value = 0.015

I keep some of those coefficient, together with the coefficient in the previous practice to achieve my final regression

h) My own regression model (After choosing minimum p-values and some selections):

$$\mathrm{Y1} = \mathrm{DMC} + \,wind + wind^2 + \,wind^3 + \mathrm{DMC} * \mathrm{temp} + \mathrm{DC} * \mathrm{temp} + \mathrm{DC} + \mathrm{temp} * \mathrm{wind}$$

And I tried to fit coefficients for this model:

The fit result was not that satisfying because all the p-value was too big for rejecting null hypothesis. However, to finish this problem, I give my model with coefficient here:

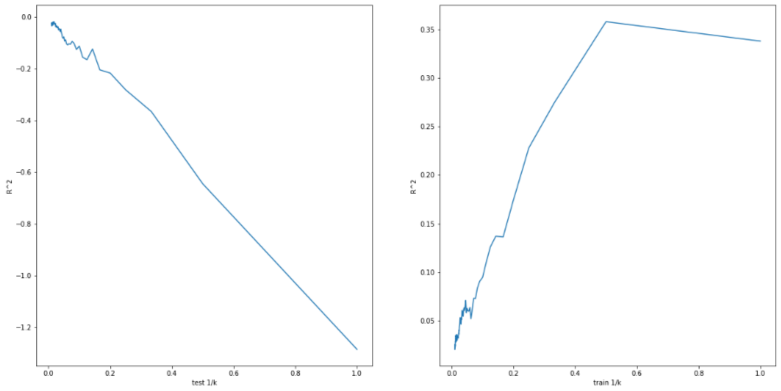
$$\mathrm{Y1} = -0.0073\mathrm{DMC} + \,0.9902wind - 0.1555wind^2 + \,0.0096wind^3 + 0.0005\mathrm{DMC} * \mathrm{temp} + 0.0013\mathrm{DC} - 5.33^{-15}\mathrm{DC} * \mathrm{temp} - 0.0132\mathrm{temp} * \mathrm{wind} + \varepsilon$$

Then I apply this to test set to calculate error on test set and got RSE value 1.401. (Calculation in Jupyter Notebook).

i) KNN regression: I used KNN regressor for this question. I tested k from 1 to 100 and check the R^2 value to find out whether it is a good fit. Since the R^2coefficient in sklearn is defined as (1 - u/v), where u is the residual sum of squares ((y_true - y_pred) ** 2).sum() and v is the total sum of squares ((y_true - y_true.mean()) ** 2).sum(). It may possible reach a negative value (which means the model is arbitrarily worse). And R^2 value that closer to 1 means a better fit.

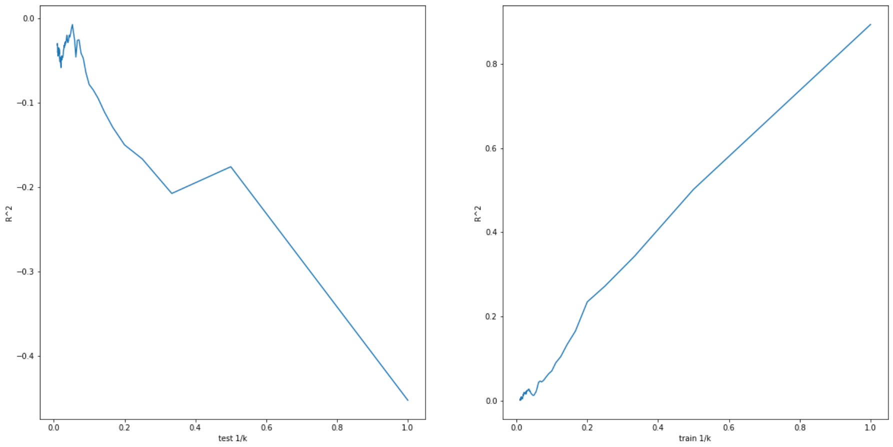
The parameters for KNN regression: minkowski distance, algorithm = auto, test size = 0.3 of all data.

- First 4 predictors: to evaluate parameter month and day, I used sklearn's label encoder method. And based on the generated result and my observation, all the estimated R^2 is under 0. Which means the model is not valuable no matter how I change K value. And the plotted result for R^2 with train set and test set is below:

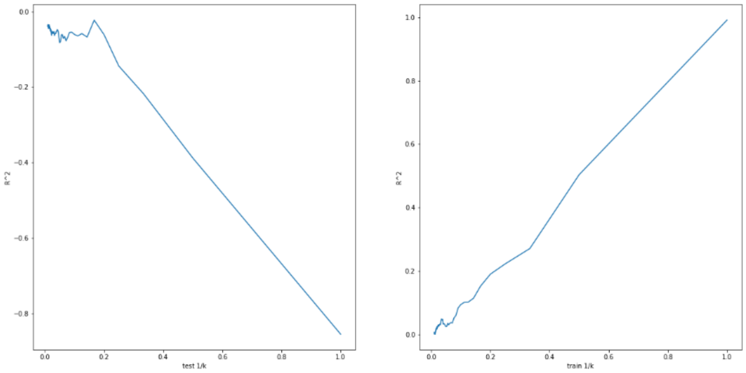


When using train set for testing, the value of $R^2$was above 0, but when using test set, $R^2$was below 0. We can say that these 4 predictors were not good for KNN regression.

- Last 4 predictors: This time I got an excellent value of R^2 for train set. It is above 1 which means there is small errors between the predicted value and true value of Y1. R^2value against 1/k graphs are shown below, the best $R^2$is about 0.93 when k = 1. However, the value of $R^2$is still below 0 under test sets testing.



- Predictors 1,2, 9, 10, 11, this result is not good ether. For testing, R^2 is below 0, we have to conclude that these predictors were not good enough for KNN regression.



j) In conclusion, KNN did a better job in regression when taking correct features, though none of them gives out a reasonable regression for test set. For linear regression, things get difficult when you have a lot of combined categorial and quantitative predictors. It is hard to find out linear patterns when dealing with so many parameters. I do not think the linear model I derived in this assignment is the best model, so I also cannot say that KNN provides a better fit generally. I really hope I can get more technique in feature selection for linear regression so that I can do better in the next regression work.