Introduction to Machine Learning and Data Mining
Forest fires

Student names and ids: Anders H. Opstrup (s160148); Gu Jinshan (s161944); Huayu Zheng(s126077)

Report: 1

1 Description of the data set

This section of the report will shed some light on basic information about the dataset used in the project.

Problem of interest

The data is about forest fires. The goal of the data is to try to predict forest fries in an attempt to prevent casualties and property damages.

Place of data

The data is obtained from this link: http://archive.ics.uci.edu/ml/datasets/Forest+Fires

Previously work on the data set

P. Cortez and A. Morais. A Data Mining Approach to Predict Forest Fires using Meteorological Data. In Proceedings of the 13th EPIA 2007 - Portuguese Conference on Artificial Intelligence, December, 2007. (http://www.dsi.uminho.pt/ pcortez/fires.pdf)

In the above reference, the output "area" was first transformed with a ln(x+1) function. Then, several Data Mining methods were applied. After fitting the models, the outputs were post-processed with the inverse of the ln(x+1) transform. Four different input setups were used. The experiments were conducted using a 10-fold (cross-validation) x 30 runs. Two regression metrics were measured: MAD and RMSE. A Gaussian support vector machine (SVM) fed with only 4 direct weather conditions (temp, RH, wind and rain) obtained the best MAD value: 12.71 +- 0.01 (mean and confidence interval within 95% using a t-student distribution). The best RMSE was attained by the naive mean predictor. An analysis to the regression error curve (REC) shows that the SVM model predicts more examples within a lower admitted error. In effect, the SVM model predicts better small fires, which are the majority.

Primary machine learning modeling aim

Detection and test of outlier methods and try different regression methods and look at the correlation between the temperature, wind, rain and the burn area.

2 A detailed explanation of the attributes of the data

In this section the reader will gain further knowledge about the dataset, a short explanation about the different attributes can be found in this part of the report.

Attribute information

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 - The Fine Fuel Moisture Code (FFMC) is a numeric rating of the moisture content of litter and other cured fine fuels. This code is an indicator of the relative ease of ignition and the flammability of fine fuel
6. DMC - The Duff Moisture Code (DMC) is a numeric rating of the average moisture content of loosely compacted organic layers of moderate depth. This code gives an indication of fuel consumption in moderate duff layers and medium-size woody material.
7. DC - The Drought Code (DC) is a numeric rating of the average moisture content of deep, compact organic layers. This code is a useful indicator of seasonal drought effects on forest fuels and the amount of smoldering in deep duff layers and large logs.
8. ISI - The Initial Spread Index (ISI) is a numeric rating of the expected rate of fire spread. It combines the effects of wind and the FFMC on rate of spread without the influence of variable quantities of fuel.
9. temperature in Celsius degrees: 2.2 to 33.30
10. RH - relative humidity in
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

Describe if the attributes are discrete/continous, Nominal/Ordinal/Interval/Ratio.

1. X - Discrete, Nominal
2. Y - Discrete, Nominal
3. month - Discrete, Ordinal
4. day - Discrete, Ordinal
5. FFMC - Continous, Interval

6. DMC - Continous, Interval
7. DC - Continous, Interval
8. ISI - Continous, Interval
9. temp - Continous, Interval
10. RH - Continous, Ratio
11. wind - Continous, Ratio
12. rain - Continous, Ratio
13. area - Continous, Ratio

Data issues

After a further investigation of the dataset

Describe the basic summary statistics of the attributes.

Statistics	FFMC	DMC	DC	ISI	temp	RH	wind	rain	area
Mean	90.64468	110.8723	547.94	9.021663	18.88917	44.2882	4.017602	0.021663	12.84729
Median	91.6	108.3	664.2	8.4	19.3	42	4	0	0.52
Variance	30.41268	4094.018	61417.81	20.74862	33.65168	265.7448	3.20381	0.087422	4044.226
Standard deviation	5.51477	63.98451	247.8262	4.555065	5.801007	16.30168	1.789919	0.295673	63.59423

3 Data visualization

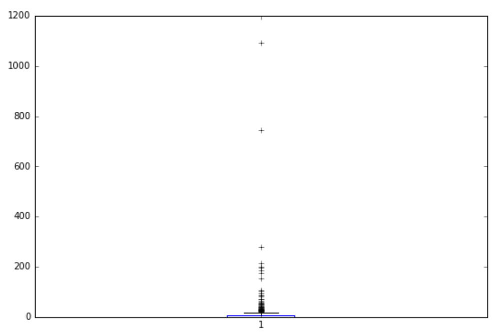
The boxplot analyse method have been used for detecting outliers in the data set. In the boxplots below we can see that there are some outliers and some attributes where the dataset is not very detailed. The area attribute needs some stemming for correcting the data, due to a lot of zero areas. The FFMC attribute and ISI also needs stemming for cleaning up the data.

Do the attributes appear to be normal distributed

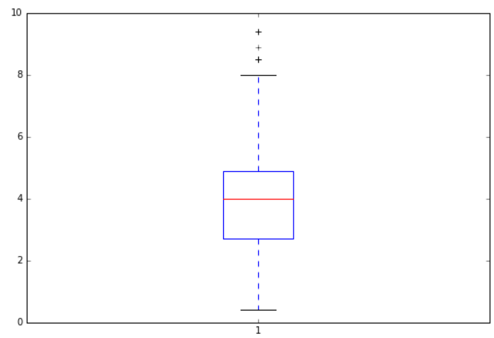
We generate the histogram for 9 attributes to find whether the attributes is normal distributed. We did not analysis the attributes X, Y, month and day since it is meaningless draw the histogram for them. From the histogram we can find that the FFMC, ISI and temp attributes seems to be like normal distributed. But others are not.

Variables correlation

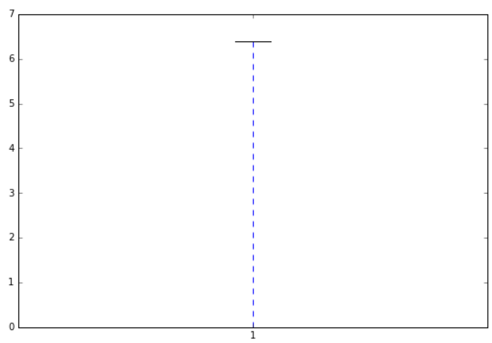
After calculated the $\hat{COR}$ of two variables, we found those two couples of variables are possibly correlated. The absolute value of $\hat{COR}$ are above 0.5. The two scatter plots below show the data distribution.



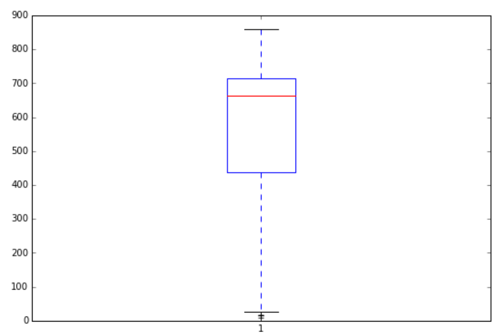
(a) area



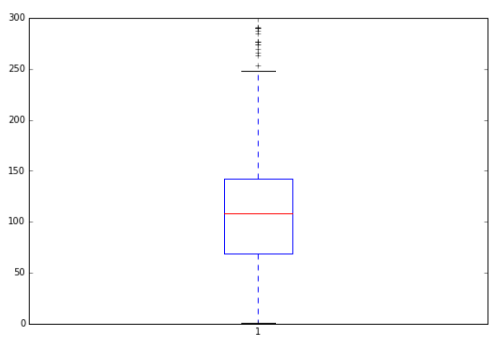
(b) wind



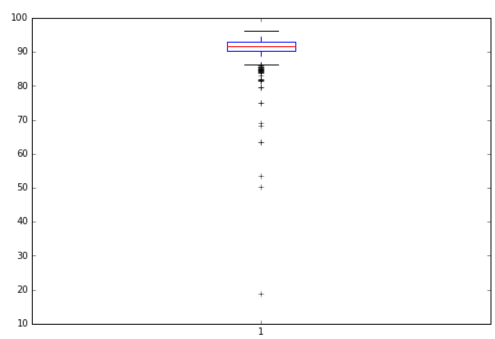
(c) rain



(d) DC

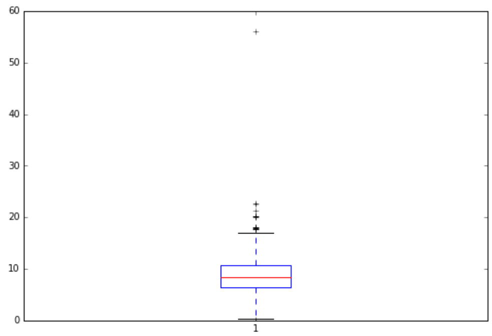


(e) DMC

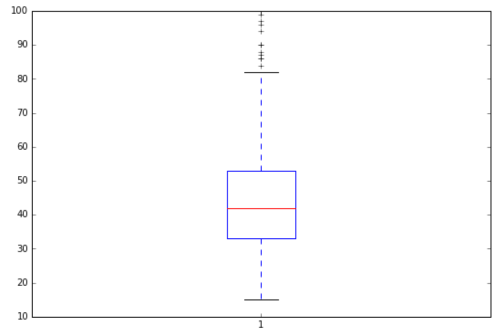


(f) FFMC

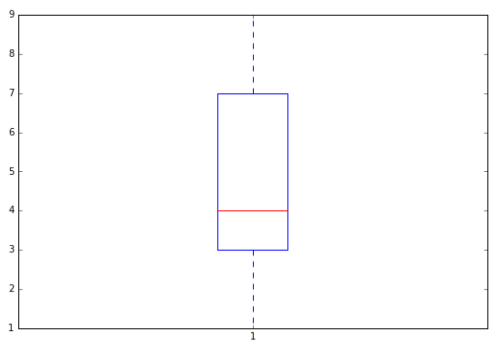
Figure 1: Boxplot diagrams



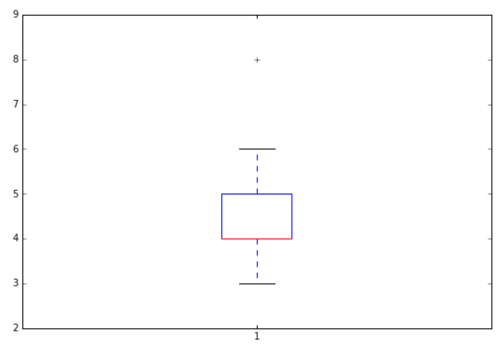
(a) ISI



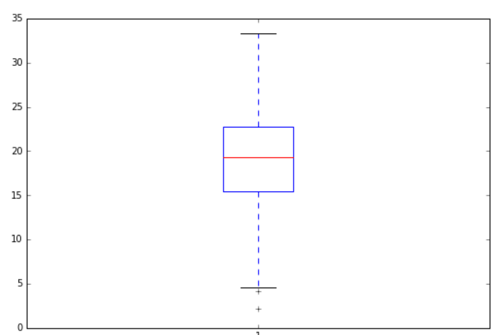
(b) RH



(c) X



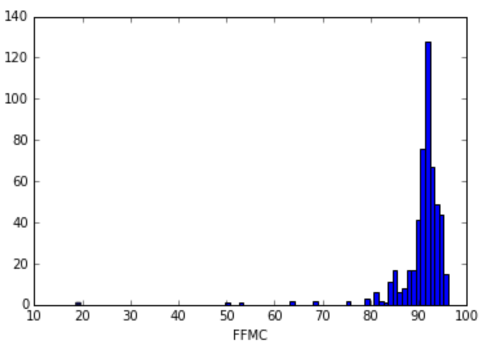
(d) Y



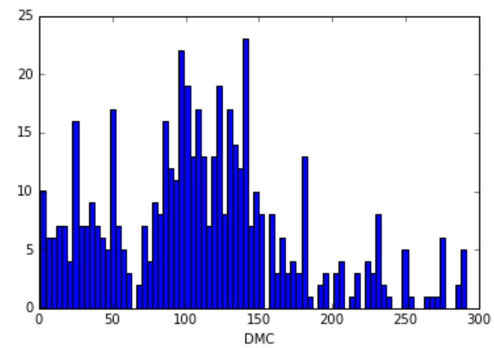
(e) temp

Figure 2: Boxplot diagrams

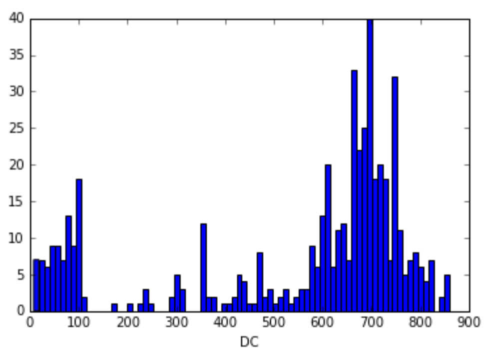
The machine learning modeling aim appears to be quite feasible for the data set, but the set needs some thoroughly clean up, as seen in the boxplots, before some techniques can be used.



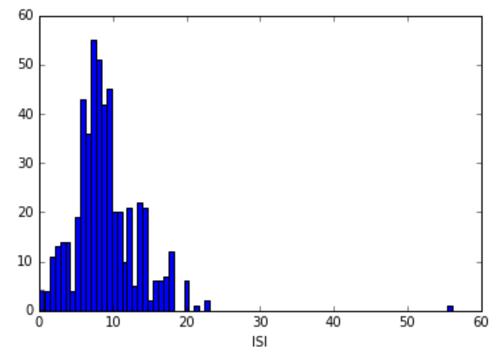
(a) FFMC



(b) DMC

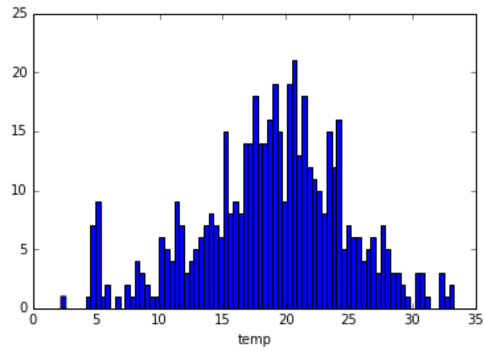


(c) DC

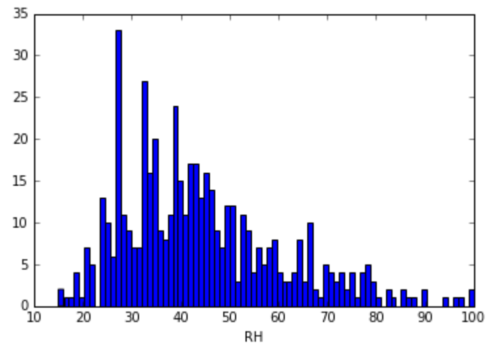


(d)ISI

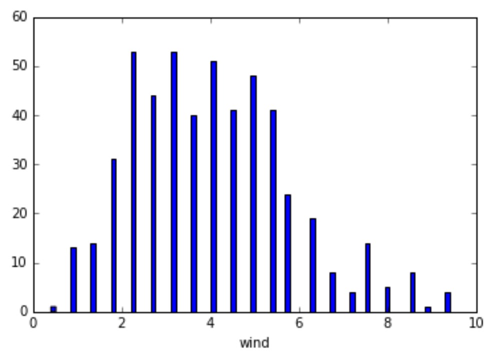
Figure 3: Histogram



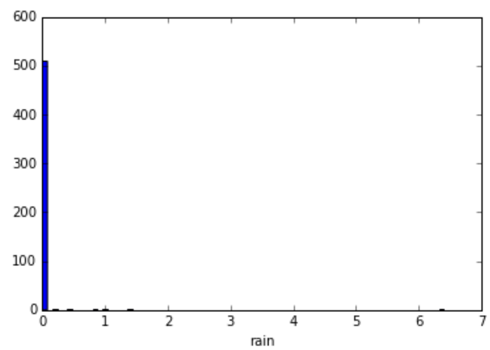
(e) temp



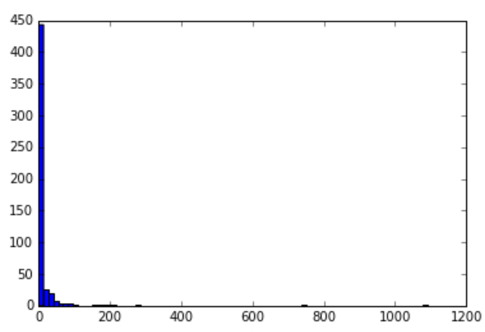
(f) RH



(g) wind

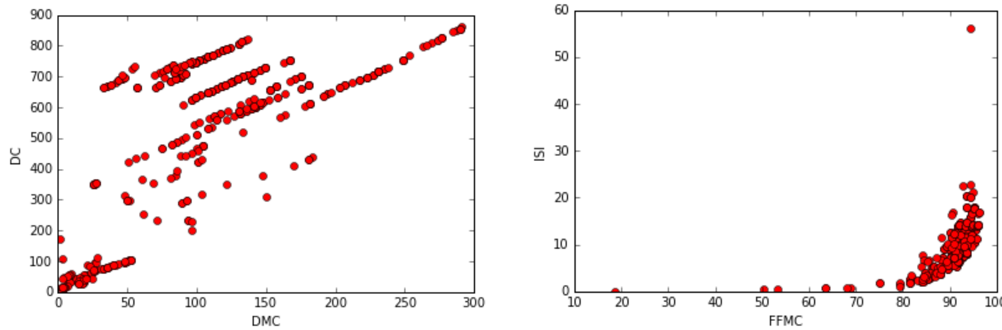


(h) rain



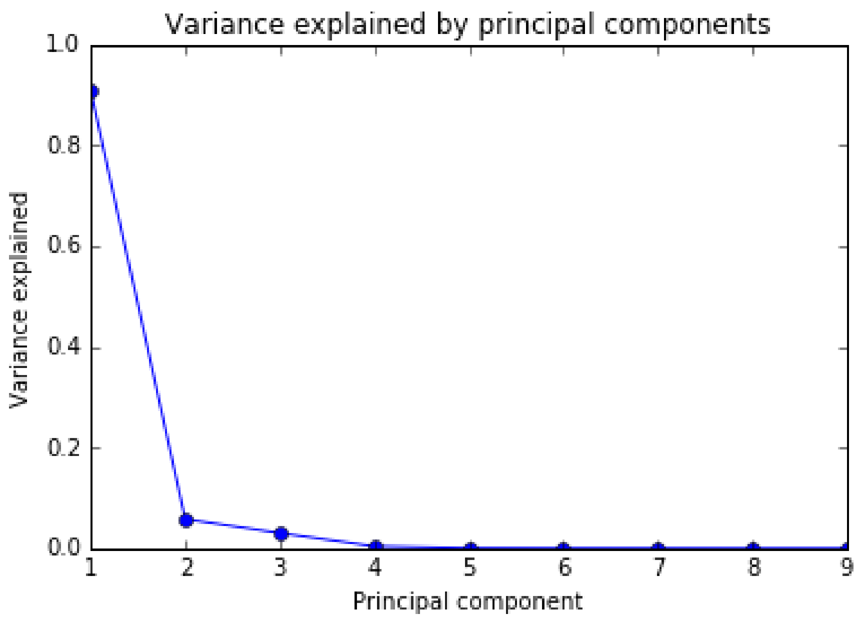
(i) area

Figure 4: Histogram



(a) CÔR=0.682 (b) CÔR=0.531

Figure 5: Scatter Plot



PCA

In order to do the PCA analyse some class had to be made for the dataset. The months was chosen as classes for trying to project the data in aspect of a year. By looking at (figure 3) the reader will see that around 90% of the information is in component 1, of the PCA model. By this we can conclude that most of the variance happens in the first component and there by is the most important component.

4 Discussion

After analysing the dataset, the group is a bit concerned if the dataset is big enough. With only around 500 records and 11 attributes, the dataset is a bit small. The boxplot shows that there is a lot of outliers in the dataset and no concrete data class where to be found in the dataset, which made the pc-analyse very hard. It means we should exclude the outliers after we move on the next step on our dataset.Also,some of the attributes are kind of skewed, we used some transform to pre-process them.Among all the attributes, FFMC, ISI and temp are considering normal distributed, which can be esaily to analyze in the further work. After calculated the $\hat{COR}$ value, we found out that the correltation between attributes are not that abvious. Only two coulpe of attributes are kind of correltated.It is also the aim of our machine learning task, to find out more correlations during our study.

Before visualizate the dataset, we made assumption that area and tempreture were highly correltive with the forest fire.Though the tendency isn't that obivous, they will still be the attributes we are more interested in the further study, since our orignal aim of learning this dataset is help to prevent forest fire.Besides, the previously work on the dataset give us an idea to combine different attribute as new element to analyze.