

Homework 1

IE 7275 Data Mining in Engineering

Note: Read the book chapter “R Graphics.pdf” posted on Blackboard. Practice example problems given in the book chapter.

Problem 1 (Forest Fires) [40 points]

The file `forestfires.xlsx` includes data from Cortez and Morais (2007). 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) \times 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. Number of instances and attributes are 517 and 13 respectively.

Attribute Information:

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

First load the file `forestfires.csv`, next perform the following tasks for the data:

- Plot `area` vs. `temp`, `area` vs. `month`, `area` vs. `DC`, `area` vs. `RH` for January through December combined in 1 graph. Hint: Place `area` on Y axis and use 2x2 matrix to place the plots adjacent to each other.
- Plot the histogram of `wind` speed (km/h).
- Compute the summery statistics (min, 1Q, mean, median, 3Q, max,) of part b.
- Add a density line to the histogram in part b.
- Plot the density function of months. Use different colors in the graph to interpret your result clearly. [Hint: use `qplot(geom=density)`]
- Plot the scatter matrix for `temp`, `RH`, `DC` and `DMC`. How would you interpret the result in terms of correlation among these data?
- Create boxplot for `wind`, `ISI` and `DC`. Are there any anomalies/outliers? Interpret your result.
- Create the histogram of `DMC`. Create the histogram of log of `DMC`. Compare the result and explain your answer.

Problem 2 (Tweeter Accounts) [40 points]

Twitter is a social news website. It can be viewed as a hybrid of email, instant messaging and sms messaging all rolled into one neat and simple package. It's a new and easy way to discover the latest news related to subjects you care about.

This is the data set crawled on July, 2009. BlogCatalog is a social blog directory website. This contains the friendship network crawled. For easier understanding, all the contents and variables are organized in CSV file format.

First load the file `M01_quasi_twitter.csv`, next perform the following tasks:

- How are the data distributed for `friend_count` variable?
- Compute the summery statistics (min, 1Q, mean, median, 3Q, max) on `friend_count`.
- How is the data quality in `friend_count` variable? Interpret your answer
- Produce a 3D scatter plot with highlighting to impression the depth for variables below on `M01_quasi_twitter.csv` dataset. `created_at_year`, `education`, `age`. Put the name of the scatter plot "3D scatter plot".
- Consider 650, 1000, 900, 300 and 14900 tweeter accounts are in `UK`, `Canada`, `India`, `Australia` and `US` respectively. Plot the percentage Pie chart includes percentage amount and country name adjacent to it, and also plot 3D pie chart for those countries along with the percentage pie chart. Hint: Use $C=(1, 2)$ matrix form to plot the charts together.
- Create kernel density plot of `created_at_year` variable and interpret the result.

Problem 3 (Insurance Claims) [20 points]

Consider that we need to rate a product based on four different aspects Sustainability, Carbon footprint, weight and required power to be built. Those variables are gathered into `raw_data.csv` spreadsheet in columns A, B, C and D respectively.

First load the file `raw_data.csv`, next perform the following tasks:

- Standardize the data and create new dataset with standardized data and name it `Ndata`.
- Create the boxplot of all the variables in their original form.
- Create boxplot of all the variables in their standardized form.
- Compare the result of part b and part c; interpret your answer.
- Prepare scatter plot of variables A and B. How are the data correlated in these variables? Interpret your answer.

Files Included in the Folder:

`Homework 1.pdf`
`R Graphics.pdf`
`forestfires.csv`
`M01_quasi_twitter.csv`
`raw_data.csv`