Week 9 - Class Worksheet

Transform: Data Transformation and Data Reduction

Dr. Anil Dolgun

Last updated: 18 June, 2020

Required Packages

The following packages and the function will be required or may come in handy.

```
library(readr)
library(dplyr)
library(forecast)
library(infotheo)

minmaxnormalise <- function(x){(x- min(x)) /(max(x)-min(x))}
```

Exercises

US Candy Production Data

The following exercises 1-4 will be based on US Candy production candy_production.csv (../data/candy_production.csv) data set from Kaggle https://www.kaggle.com/rtatman/us-candy-production-by-month (https://www.kaggle.com/rtatman/us-candy-production-by-month). Variables are self explanatory however it is expected to do checks on the type of the data and using the suitable transformations if necessary. Here is a quick look of the candy data:

observation_date	production
1972-01-01	85.6945
1972-02-01	71.8200
1972-03-01	66.0229
1972-04-01	64.5645
1972-05-01	65.0100
1972-06-01	67.6467

1 **Data Transformation:** Use `hist()` to check the shape of the distribution of production variable in `candy` data set. Apply data transformation via mathematical operations such as log base 10, log base e, square root and reciprocal transformations. Apply Box - Cox transformation. After you applied transformations, use `hist()` and check shape of the distribution for each transformation.

2 **Data Normalisation:** Apply mean - centering and scale by the standard deviations without centering to the production variable in `candy` data set. Use `hist()` to check the shape of the distribution for both normalisations you applied.

3 **z Score Standardisation and Min- Max Normalisation:** Apply z-score standardisation and min-max normalisation to the production variable in `candy` data set. Use `hist()` to check the shape of the distribution for both transformations you applied.

4 **Binning (a.k.a. Discretisation):** Use equal width (distance) binning and equal depth (frequency) binning to the production variable in `candy` data set. Check the head of the first 15 observations for both transformations.

Ozone Data

The following exercises 5-9 will be based on ozone.csv (../data/ozone.csv) data set which is taken from http://rstatistics.net/wp-content/uploads/2015/09/ozone.csv (http://rstatistics.net/wp-content/uploads/2015/09/ozone.csv) containing 366 observations and 13 variables. Variables are self explanatory however it is expected to do checks on the type of the data and using the suitable transformations if necessary.

Here is a quick look of the ozone data:

Month	Day_of_month	Day_of_week	ozone_reading	pressure_height	Wind_speed	Humidity	Temperature_Sandburg
1	1	4	3.01	5480	8	20	NA
1	2	5	3.20	5660	6	NA	38
1	3	6	2.70	5710	4	28	40
1	4	7	5.18	5700	3	37	45
1	5	1	5.34	5760	3	51	54
1	6	2	5.77	5720	4	69	35

Temperature_ElMonte	Inversion_base_height	Pressure_gradient	Inversion_temperature	Visibility
NA	5000	-15	30.56	200
NA	NA	-14	NA	300
NA	2693	-25	47.66	250
NA	590	-24	55.04	100
45.32	1450	25	57.02	60
49.64	1568	15	53.78	60

5 **Data Transformation via Mathematical Operations:** Subset variables `ozone_reading`, `pressure_height`, `Pressure_gradient`, `Visibility`, `Inversion_temperature` from `ozone` data set and name it ozone_sub. Use `hist()` to check the shape of the distribution for all the variables. Apply log base 10, log base e and square root transformations to the variables. `sapply()` function will come in handy to transform all the variables at once. Check the shape of the distribution of the variables using `hist()`.

6 **Centering and Scaling:** Apply mean-centering to `ozone_sub` data frame using `apply()` function. Check the shape of the distribution of the variables using `hist()`.

7	**Min- Max Normalisation:** Use min-max normalisation to the `ozone_sub` data frame. If you are getting NAs explain why. Take the appropriate action to fix the problem and apply the normalisation again. Use `hist()` to check the shape of the distributions of the variables.
8	**Binning:** Use `ozone_reading` variable from `ozone` dataset and apply equal width (distance) and equal depth (frequency) binning. Compare the variable before and after binning. To do so use `cbind()` and show 15 observations from the outputs.
9	**Data Challenge**: Use `ozone_sub` data frame and apply Box Cox transformation using `apply()` function. Show the shape of the distibution of the variables using `hist()` . See if you can use a loop for histograms.
10	**Bonus Exercise**: Select `ozone_reading` , `pressure_height` , `Inversion_temperature` variables from ozone data set. Apply z-score standardisation using `scales()` and `scores()` functions. Then compare the results of these two functions to see if you get the same results. Don't forget to deal with NA values.
11	**Data Reduction:** Explore the `who` dataset under the `tidyr` package. What would be the benefit of reducing the dimensions on this dataset? Post your answers on the discussion board.

Finished?

If you have finished the above tasks, work through the weekly list of tasks posted on the Canvas announcement page.

Return to Course Website (../index.html)