HW1 ANOVA

Problem description:

The study of **Social groups** and the **collective behaviors** of their members are hot topic not only in sociology, but also in computer science. In this homework, we attempt to perceive the semantics of social groups from collective social and behavioral information. Given the categories of the social groups and some features of collective social and behavioral information, our final goal is to test whether these features can distinguish the categories.

Data

All the data is stored in one file, named **data.xlsx.**

The dataset describes the online group collected from QQ. We select 2040 online groups with corresponding information in 14 columns (denoted Col[1-14]):

Col[1-2]: online group name, group category. As you know, each QQ group has a group name to describe the semantics of the group. For both privacy and intuition, some characters of the names are masked by '*'. The descriptions of category are shown below in Table 1:

Table 1. Category description

Category	Theme	No.
1	Online Game	484
2	School Alumni	300
3	House & Living	196
4	Stock Market	425
5	Organizations & Industry	635

Col[3-14]: 12 dimension features, they are group size, message number, friendship relational density, sex ration, average age, variance of age, geographical area, mobile conversation ratio, conversation number, no-response conversation ratio, night conversation ratio, images ratio.

Experiments

1. (**5 points**) Recall and write down the assumptions which one-way ANOVA are based on.
2. (**5 points**) Focus on two columns: Category (Col[2]) and Average Age (Col[7]). Taking feature Average Age as an example, we want to measure whether the average age varied significantly across the categories. Clearly state the null (H0) and the alternative (H1) hypotheses for this task.
3. Use your favorite statistics analysis software, like Matlab, R, Excel, SPSS or …
    a) (**5 points**) Draw the empirical probability density function of Col[7], i.e. the empirical pdf of average age. Does the data in this dimension follow Gaussian distribution? Test normality of Col[7].

b) (**5 points**) In Col[7], there are 5 components divided by category labels. We denote the data in Col[7] with category i (where i = 1,…,5) as Col[7| categoty=i]. Test the normality of each components and test the homogeneity of variances.

c) (**15 points**) Do the one-way ANOVA test for Col[7] with categories in Col[2]. Write down your conclusion, supporting statistics, and visualize your data which inspire the process.

4. (**10 points**) Choose another 3 columns, draw the empirical pdf of each feature columns and test which column follows these assumptions in question 1? How about their corresponding log transformation?
5. How to do one-way ANOVA with the non-normal data?

   a) (**10 points**) Find and list the possible solutions set.

   b) (**15 points**) Do the one-way ANOVA on the 3 columns you choose. Do these feature columns vary significantly? Visualize the results.
6. (**20 points**) Redo the ANOVA test in question 3 c) by sampling 10% data (i.e. around 200 groups). Repeat 10 times and compute the mean and standard deviation of the supporting statistics (F value). Compare at least two sampling strategies. Which sampling method is more stable? How are the results compared to the results without sampling? Why?
7. (**10 points**) Choose any two categories, and classify them by logistical regression, or you can try multi-label classification on all categories.

Compress the experiments report and the necessary codes into one file. For more details and beyond, please refer to: http://cuip.thumedialab.com/papers/Cui-group.pdf