ФКН НИУ ВШЭ Москва

Бакалавриат

Миркин Б.Г.

Основные методы анализа данных

Задания по Проекту 2017-18

FCS NRU HSE Moscow

BSc Programme

B. Mirkin

Core data analysis

Homework assignments 2017-18

Please take this seriously. Not only because the mark for the Homework is 40% of the final mark, but, mostly, because this gives you a chance to get hold over key data analysis skills practically.

Your report of the homework must reach Instructor at [bmirkin@hse.ru](mailto:bmirkin@hse.ru) by the end of 3 December (till morning of 4 December). Reports submitted after this deadline but before the end of 10 December will be penalized by 20% off the mark. No reports accepted after 10 December.

Please be reminded that your report must begin with a title page with the name of the team and its members.

For computations, you may use any computing environment including the most popular MatLab, R, Python, etc. You may write your own codes or use those provided by the environments. You must make an exact reference to what a specific tool was used, provide its code and specify the parameters of your application. The less the comments you do to the solutions, the greater the penalty to be imposed on your mark. An assignment may be not graded at all (mark=0) If no comments are provided.

The language is not important. Of course good English is taken notice of and supported.

Assignment 1

**1.** Each to join a team of one or two; the team **finds a meaningful dataset** of their liking **on the internet**:

* Number of entities not less than 70
* Number of features not less than 7
* **No missings**
* **No Irivine ML depository**

**The dataset is to be approved by me on 15/09/2017.**

**2. Start writing a team’s report file**

* Project title page
* Section 1.
  + Explanation of the choice of the dataset
  + Information of the dataset: features, number of entities, source address, examples of problems

Assignment 2

* Reasonably select several features in your dataset
* Apply K-Means at two or three different K
* Take a clustering of your liking and interpret all the clusters by comparing their centers with grand mean
* Answer this question in your report: is it an interesting partition?

Assignment 3

* Reasonably select two clusters in the clustering you dealt with in HW2
* Take one of the features at one of the clusters and validate its within-cluster mean using bootstrap
* Take one more cluster and compare the within-cluster means of the feature by using bootstrap

Assignment 4

* Выбрать или создать один-два номинальных признака на ваших данных.
* Построить таблицу (таблицы) сопряжённости этого признака (этих признаков) с разбиением по методу к-средних.
* Рассчитать таблицу коэффициентов Кетле и прокомментировать выделяющиеся значения.
* Рассчитать и проинтерпретировать сводный коэффициент Кетле и убедиться, что он совпадает с коэффициентом Пирсона фи-квадрат.
* Рассчитать, при каком количестве объектов гипотеза независимости разбиения и признака должна быть отвергнута с 95% уровнем доверия. Прокомментировать.

Assignment 5

* In your data set, select 3-6 features related to the same aspect and explain your choice
* Visualize the data with these features using SVD with two versions of normalization: (a) over ranges and (b) over standard deviations.
* At these visualizations, use a distinct shape/color for points representing a pre-specified by you group of objects.
* Also, apply the conventional PCA for the visualization and see if there is any difference.
* Comment on which of the normalizations is better and why.

Assignment 6

* Find two features in your dataset with more or less “linear-like” scatterplot.
* Display the scatter-plot.
* Build a linear regression of one of them over the other. Make a comment on the meaning of the slope.
* Find the correlation and determinacy coefficients, and comment on the meaning of the latter.
* Make a prediction of the target value, given two or three predictor’ values; make a comment
* Compare the relative absolute error of the regression on all points of your set and the determinacy coefficient and make comments