

Knowledge Discovery in Databases
SS 2016

Exercise 2: Data Representation and Data Reduction

Regarding tutorials on 27.04.-29.04.2016.

For all programming assignments in this course we recommend the usage of *python 2.7*. We will also use *numpy*, *pandas* and *matplotlib* libraries. Solutions will be provided as *jupyter notebook* files. If you are new to python, the following links might be useful for you:

python tutorial (not obligatory for this exercise):

http://www.scipy-lectures.org/intro/language/python_language.html

numpy tutorial (not obligatory for this exercise):

<http://www.scipy-lectures.org/intro/numpy/index.html>

pandas tutorial:

<http://pandas.pydata.org/pandas-docs/stable/10min.html>

matplotlib in pandas:

<http://pandas.pydata.org/pandas-docs/stable/visualization.html>

There are several python distributions which already include popular libraries for data analysis, e.g. Anaconda:

<https://www.continuum.io/downloads>

Exercise 2-1 Dataset Exploration

In this exercise we will explore the Lending Club Loan dataset. Lending Club is a US peer-to-peer lending company. The dataset describes loans issued by the company and contains informations about borrowers. The subsets for different time periods are available on the Lending Club website:

<http://www.lendingclub.com/info/download-data.action>

Feel free to use any of these subsets. In the exercise we will use the complete dataset of the year 2015.

There is also a data dictionary explaining the description of dataset attributes. It can also be downloaded on the website.

To get first impressions of the dataset we try the following:

(a) Data loading

First, check data characteristics. What is the data format, is metadata available? Is there a possibility to load the data directly or is a transformation necessary?

(b) First exploration

Now, after data is loaded, what does it look like? How many instances are there, how many features? To get a first feeling about the dataset, print a few instances and take a closer look at them.

(c) Loan status

The feature *loan_status* describes current loan status for each loan. Now it is the time to gain first insights. What do you think, is Lending club successful? Would you recommend to invest in it?

(d) Bad & Good loans

For further analysis we want to compare different characteristics of successful and unsuccessful loans. Select two samples of the same size: one with already repaid (*loan_status* 'Fully Paid') and one with charged off loans (*loan_status* 'Charged Off').

(e) Per state distribution

The dataset provides the home state information for each borrower. In the lecture you've learned about different types of data. What is the data type of this feature?

Now, if we want to compare creditworthiness for the residents of different states, what would be a suitable visualization? Visualize the per state distributions for good and bad loans. What do you think, is this feature valuable?

(f) Other interesting features

What about *home_ownership*, *emp_length*, *purpose* features? How do the values look like? What do you think, the borrower with which characteristics would repay credit with highest probability? Visualize features separately and prove your intuition.

(g) Loan Grade

The *grade* feature describes the grade assigned to the loan by the Lending club. This grade describes Lending club's estimation of repayment probability and therefore determines the loan interest rate. Prove this estimation. How good is it?

(h) Credit costs to income

The *dti* feature describes borrowers' costs of previously taken credits relative to income. What data type is it? What are the possible ways to describe the distribution?

Let's say we want to order the values and visualize minimum and maximum values. In addition, consider smallest quarter, half and three quarters of values. Select a suitable visualization and compare distributions of both samples.

(i) Interest rate

Does the credit interest rate correlate with the probability of repayment? Compare distributions of interest rates of both samples. For each sample compute also a **mean** of interest rate. Does it differ from the **median**. If yes, why? What do you think, which measure is more meaningful?

(j) Income

What about income, does a person's income say anything about creditworthiness? Compare distributions of *annual_inc* of both samples.

(k) Income discretization

Comparison of continuous values can be tedious. Sometimes it is much easier to discretize and then to compare. Think of meaningful income ranges, name it and assign income a label to each loan. Compare the distributions of values for both samples.

(l) Limits comparison

Let's continue with the whole dataset. The *total_bc_limit* feature shows the limit of all borrowers' bank-cards. *total_rev_hi_lim* states the limit issued by the lending club. Visualize the coherence between two features. Is there a correlation?

(m) Credit limit to annual income

Now we want to see whether there is correlation between borrowers' income and credit limit. Make a visualization give a statement.

(n) Interest rate to income

What about interest rate? Is there a correlation to annual income?