Intelligent Data Analysis Summer 2014
C. Braune

Exercise Sheet 3

Exercise 9 Characteristic Measures

Determine for the data set used in exercise 5, that is, for

4, 3, 2, 5, 4, 6, 3, 7, 4, 1, 4, 0, 6, 4, 3, 5, 2, 3, 5, 1, 4, 4, 9, 5, 4, 3, 3, 5, 2, 4,
3, 6, 5, 2, 6, 2, 4, 5, 5, 1, 5, 4, 4, 2, 7, 1, 3, 3, 4, 7, 3, 4, 4, 6, 6, 3, 3, 2, 6, 1,

- the mode,
- the median (central value),
- the $\frac{1}{3}$-quantile,
- the mean,
- the range,
- the interquartile range,
- the variance and the standard deviation,
- the skewness and
- the kurtosis!

In addition, draw a box plot!

Exercise 10 Principal Component Analysis

Let the following two-dimensional data set be given:

x	0	1	1	2	3	3	4	5	5	6
y	0	1	2	3	2	3	4	4	6	5

From this table it is already obvious that the two attributes x and y are strongly correlated.

a) Compute the covariance matrix of the data!

b) Compute the correlation coefficient of x and y!

c) Execute a principal component analysis (using unnormalized data), that is, determine the directions of the two principal axes of the data point cloud!

Exercise 11 Expected Values of Random Variables

Show that a Poisson-distributed random variable (that is, a discrete random variable X

with the domain $\mathbb{N}_0$ and the distribution $\Lambda_X(x;\lambda) = \frac{\lambda^x}{x!}e^{-\lambda})$ has the expected value $E(X) = \lambda$ and the variance $D^2(X) = \lambda$!

Practical Exercise 1 Descriptive Analytics

Please send a copy of your KNIME workflow to `cbraune+ida2014@ovgu.de` **until May 5^{th}, 9:00 CEST, 2014.**

Using the wine data set that is available on the lecture home page create a workflow that visualizes all pairs of dimensions of the data set (scatter matrix) with points colored differently for each class.

Also create

1. boxplots for each attribute w.r.t. the whole data set
2. boxplots for each attribute w.r.t. each class alone (Hint: There is one specific node for this, but you may also use filtering nodes)
3. and any visualization of your choice (which is none of the above) that is suitable for this data set.

Explain the results and how you might distinguish between classes given the information learned from the data views.

Additional Exercise The St. Petersburg Paradox

Consider the following gamble: A coin is thrown until head turns up for the first time. A gambler wins 2 Euro, if head shows up in the first throw, 4 Euro, if heads shows up first in the second throw, and generally 2^k Euro, if head shows up first in the k-th throw of the coin. Compute the stake that make the gamble fair! (A gamble is called *fair* if the stake equals the expected winnings.) Why is the result a paradox? Why is it irrelevant is somebody actually offered such a gamble? How should one compute the expected value instead?

In an analogous way consider the following martingale system for roulette (a *martingale* is any way of playing roulette that only uses the so-called simple chances, that is, in which stakes are placed only on red/black or on even/odd): A gambler places a stake on a simple chance. If he loses, he doubles the stake and places it on the same simple chance. If he loses again, he doubles his stake again and so on. Why is this a (theoretically) certain winning system? How do the expected winnings change if he places the stake not on a simple chance, but rather plays columns (each of the number groups 1-12, 13-24 and 25-36 form a *column*) in the same way (doubling the stake on each loss)? Why do these (theoretically) certain winning systems fail in practice?