Advanced Topics in Algebra – Lab#5

dr Michał Jarema

Topic: dimension reduction

1. Read this:

https://stats.stackexchange.com/a/140579/28666

https://stats.stackexchange.com/q/134283

2.Calculate covariance matrix

Remember: variance is the "average squared distance from the mean"

$$\text{Var}\, X = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n-1}$$

There were n observations of a single variable X.
Now we observe d different variables simultaneously.
Our data matrix $M_{n\times \mathrm{d}}$ has d columns (for each variable) and n rows (for each observation). We will construct CoVariance matrix with dimension $d \times d$, because it has an entry for each pair of variables. First, from each column of M subtract its mean – the resulting centered matrix is denoted by $C_{n\times \mathrm{d}}$. Then the $d \times\ d$ covariance matrix is $(C^T)_{\mathrm{d}\times\mathrm{n}} * \mathrm{C}_{\mathrm{n}\times\mathrm{d}}$ divided by $n-1$.

Construct datatable as shown on the lab.

Calculate covariance as shown above.

Compare with function cov().

Analyze the values of covariances and draw conclusions.

Calculate eigenvectors and eigenvalues of Covariance.

Compare them wir the result of function prcomp().

3. Work in pairs and discuss possible strategies to reduce dimensionality of the following problems. In other words, how can we get rid of many unnecessary variables and extract only useful information? We will discuss your ideas on the next lab.

A) For each client we have the total amount of bank account (negative means debit) daily for the past 10 years (i.e. 3650 variables). We want to decide, to which clients we should offer a loan.

B) You have high-res scanned rgb images of handwritten digits. (google MNIST dataset) Which unnecessary information can be dropped before passing the data to a classification algorithm?

C) We have audio recording of the beginning of "We will rock you". (44.1 kHz stereo) We partition it into 0.1s time intervals (8820 variables each). For each time interval we want to detect if there is the clap sound or silence.
Extract one more feature from the data: how you can detect whether the clap is low pitch or high pitch?

D) You have hi-res rgb bird's eye picture of fire in a forrest. (flames and green trees are visible). Partition it into squares 50x50 pixels. (each has 3*50*50 variables). Detect if there is fire or not in each square.

Reduce dimensions further to 2 coordinates of the center of burning region, where a helicopter should flush water.