VE472 — Methods and tools for big data

Homework 5

Manuel — UM-JI (Summer 2022)

Reminders

- Write in a neat and legible handwriting or use LaTeX
- Clearly explain the reasoning process
- Write in a complete style (subject, verb, and object)
- Be critical on your results

Ex. 1 — *Numerical stability*

In the lectures we mentioned numerical stability (4.13|4.137) and related that problem to the precision of the used data types. We now want to experience it using MATLAB.

1. In the big data context, how beneficial would it be to increase precision? For instance what would be the gain of using double instead of float, or move on with multi-precision?
2. Generate 100 random 1000×100 matrices X and measure the total time needed for MATLAB to compute:
   a) The SVD of X;
   b) The SVD of $X^\top$;
   c) The eigenvalues of $XX^\top$;
   d) The eigenvalues of $X^\top X$;
3. Let $X = \begin{pmatrix} -9 & 11 & -21 & 63 & -252 \\ 70 & -69 & 141 & -421 & 1684 \\ -575 & 575 & -1149 & 3451 & -13801 \\ 3891 & -3891 & 7782 & -23345 & 93365 \\ 1024 & -1024 & 2048 & -6144 & 24572 \end{pmatrix}$.
   a) Use MATLAB to determine the eigenvalues of $X + \delta X$, where δX represents a small random perturbation on X. Study the variations over 1000 tests.
   b) Use MATLAB to determine the singular values of $X + \delta X$, where δX represents a small random perturbation on X. Study the variations over 1000 tests.

   Hint: use the MATLAB `eps` *command.*
4. In light of the lectures content, explain your observations.

Ex. 2 — *Simple SVD calculations*

Without using a computer, find the singular values of the matrix

$$X = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ 9 & 0 & 1 & 2 \end{pmatrix}.$$

Detail your steps.

Ex. 3 — *PCA in Spark*

After a computer crash Krystor realises that his filsesystem was damaged and some of his files corrupted. After much work he could partially recover two files (`sensors1.csv` and `sensors2.csv`)[1]. Besides, he guesses that at least one of them corresponds to the output of 1001 sensors in the electric circuit of Reapor Rich's new cinema. Unfortunately as the column headers have disappeared in the crash, he neither

[1] Download them from the data server in the archive `cinema_sensors.tar.gz`.

knows what sensor corresponds to what column, nor whether the two files are related or not.

If his assumption is correct, then at least one of the two files hides some underlying patterns and its last columns y corresponds to the hourly electric consumption.

1. Explain how PCA can be of any help to Krystor?
2. How many columns of `sensors1.csv` are necessary to explain 90% of the data? Let n be that number.
3. Construct the linear model $y = \sum_{i=1}^{n} \beta_i p_i + x_0 + \varepsilon$, where, $\varepsilon \sim \text{Normal}(0, 1)$, x_0 is the intercept at $x = 0$, and $(p_i)_{1 \leq i \leq n}$ are the first n principal components of `sensors1.csv`.
4. Help Krystor determine whether `sensors2.csv` also contains the output of the sensors in the electric circuit of Reapor Rich's new cinema.