Foundations of data science, summer 2020
JONATHAN LENNARTZ, MICHAEL NÜSKEN, ANNIKA TARNOWSKI

7. Exercise sheet
Hand in solutions until Thursday, 4 June 2020, 12:00

Exercise 7.1 (Toy Example of SVD). (15 points)

You can use the numpy command `numpy.linalg.svd()` for this exercise.

(i) Take $A = \begin{bmatrix} 1 & 2 \\ 3 & 1 \\ 2 & 3 \end{bmatrix}$ (from the course). $\boxed{6}$

(a) Use python to compute its SVD and its 1-truncated SVD.

(b) Plot the rows of the data matrix A and the rows of its 1-truncated SVD A_1 as points.

(c) Compare to the first Lemma about the k-truncated SVD.

Lemma. *The rows of A_k are the projections of the rows of A to the best-fit k-subspace $\mathcal{V}_k$ spanned by the first k singular vectors of A.*

Hint: You may want to repeat the previous with a few other matrices A.

(ii) Write python routines: $\boxed{3}$

(a) For a given set of points and a line, the sum of the squared distances from the points to the line.

(b) For a given set of points, the best-fit 1- and 2-subspace to approximate the points.

(iii) Apply the two routines to the data sets given in the file `07-2d.csv` $\boxed{4}$ you find in sciebo.

(a) Plot the points with the corresponding best-fit line so we can visually check the correctness of your routine.

(b) Compare the sum of the squared distances of the best-fit line to the sum of the squared distances of the line given by $y = \frac{2}{3}x$.

(iv) Plot the data set and best-fit 2-subspace given in `07-3d.csv`. $\boxed{2}$

(v) Repeat some of the previous with a random data sets containing 500 points...