DatA414 Tutorial 4: Unsupervised Learning

K-means clustering by hand

Suppose you are given the following dataset. You will simulate the K-means algorithm by hand.

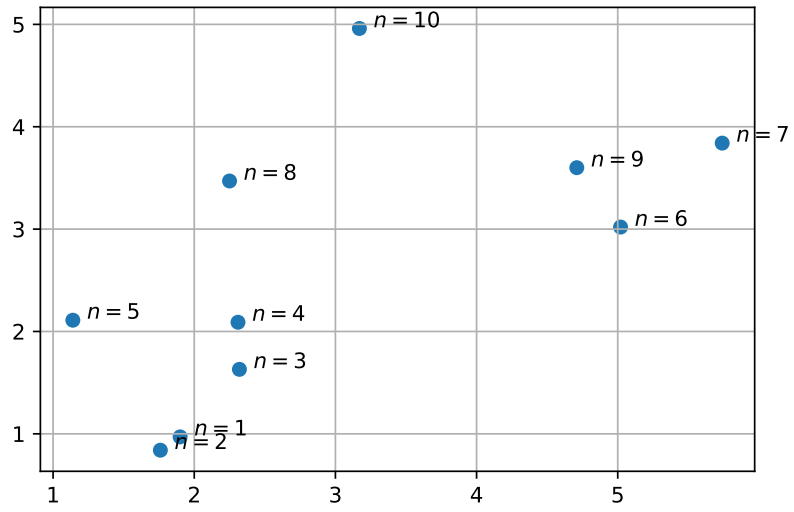


Figure 1: Dataset of ten points.

n	x_1	x_2
1	1.90	0.97
2	1.76	0.84
3	2.32	1.63
4	2.31	2.09
5	1.14	2.11
6	5.02	3.02
7	5.74	3.84
8	2.25	3.47
9	4.71	3.60
10	3.17	4.96

We will cluster this data using two clusters. Suppose the clusters are initialised with $\boldsymbol{\mu}_1 = \mathbf{x}^{(1)}$ and $\boldsymbol{\mu}_2 = \mathbf{x}^{(10)}$, i.e. the two centroids are set to the first and tenth datapoints, respectively. Now simulate the K-means algorithm with $K = 2$. After the first iteration, what are the cluster assignments? What are the cluster assignments after convergence, i.e. when the cluster assignments stop changing?

Principle components analysis

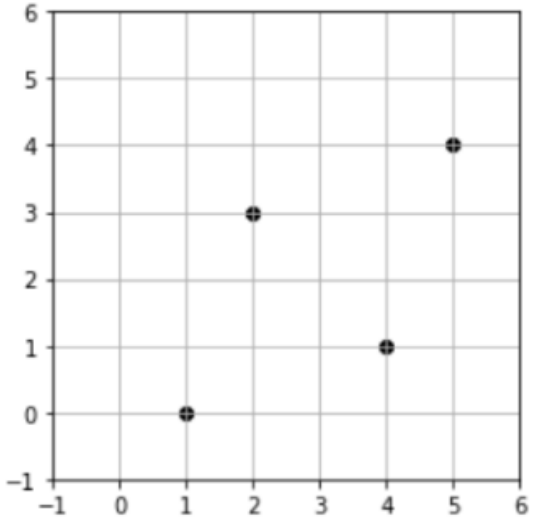
Consider the following design matrix:

$$\mathbf{X} = \begin{bmatrix} 1 & -1 \\ -1 & 1 \\ 2 & 2 \\ -2 & -2 \end{bmatrix}$$

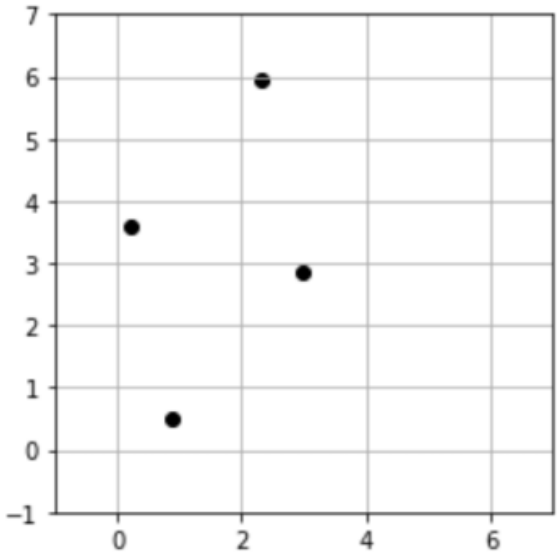
We want to represent the data in only one dimension and we will use principal components analysis (PCA).

(a) Compute the unit-length principal component directions of $\mathbf{X}$, and state which one the PCA algorithm would choose if you request just one principal component.

(b) The plot below depicts the sample points from $\mathbf{X}$. We want a one-dimensional representation of the data. Draw the principal component direction (as a line) and the projections of all four sample points onto the principal direction. Label each projected point with its principal coordinate value, where the origin's principal coordinate is zero.



(c) The plot below depicts the sample points from $\mathbf{X}$ rotated 30 degrees counterclockwise about the origin. As in part (b), identify the principal component direction that the PCA algorithm would choose and draw it (as a line) on the plot. Also draw the projections of the rotated points onto the principal direction. Label each projected point with the exact value of its principal coordinate.



PCA and SVD

(a) In your own words (with mathematics), describe the relationship between singular value decomposition (SVD) and principal components analysis (PCA).

(b) Can SVD and PCA produce the same projection result? If you think they can, under what condition are they the same? If you think they cannot, explain why.