582631 Introduction to Machine Learning, Fall 2016
Exercise set VI
Model solutions

1.

(a) The objective is to minimize a function

$$f(x') = \sum_{i=1}^{n} (x_i - x')^2.$$

Let's first differentiate this function with respect to x':

$$f'(x') = -2\sum_{i=1}^{n} (x_i - x') = -2\left(\sum_{i=1}^{n} x_i - nx'\right).$$

and then find the critical points of the function by solving where this derivative is zero:

$$\begin{aligned} f'(x') &= -2\left(\sum_{i=1}^{n} x_i - nx'\right) = 0 \\ x' &= \frac{1}{n}\sum_{i=1}^{n} x_i. \end{aligned}$$

Type of this critical point can be found by considering the second derivative of $f(\mathbf{x}')$. Because the second derivative

$$f''(x') = 2n$$

is always positive, $f(x')$ is a convex function, and so

$$x^* = \frac{1}{n}\sum_{i=1}^{n} x_i$$

is its global minimum point. This could also be noticed without computing the second derivative from the fact that $f(x')$ is a parabola that opens upwards; its only critical point is then its global minimum point.

(b) Now the function to minimize is

$$\begin{aligned} f(\mathbf{x}') &= \sum_{i=1}^{n} ||\mathbf{x}_i - \mathbf{x}'||_2^2 = \sum_{i=1}^{n}\sum_{j=1}^{p} (x_{ij} - x'_j)^2 \\ &= \sum_{i=1}^{n} (x_{i1} - x'_1)^2 + \cdots + \sum_{i=1}^{n} (x_{ip} - x'_p)^2. \end{aligned}$$

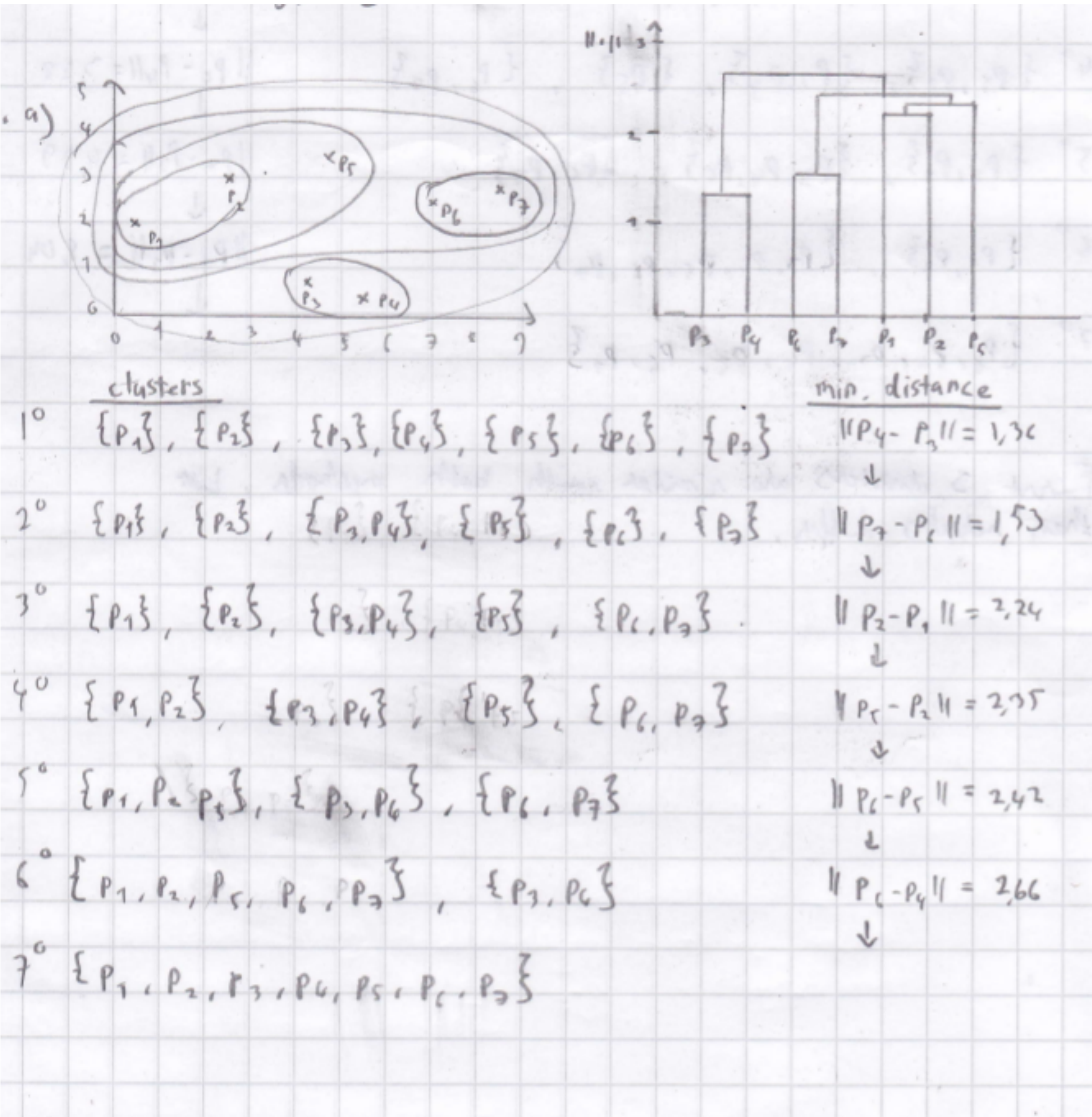
Because the value of j:th term of the sum depends only on the j:th component of $\mathbf{x}'$, the minimum value of this function can be found by minimizing each of the components separately:

$$\mathbf{x}^* := \arg\min_{\mathbf{x}'} f(\mathbf{x}') = \left(\arg\min_{x'_1} \sum_{i=1}^{n} (x_{i1} - x'_1)^2, \ldots, \arg\min_{x'_p} \sum_{i=1}^{n} (x_{ip} - x'_p)^2\right)$$

But from the first part of the exercise we observe that this is just

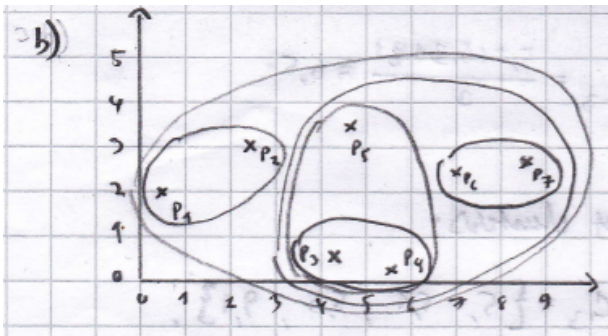
$$\mathbf{x}^* = \left(\frac{1}{n} \sum_{i=1}^{n} x_{i1}, \ldots, \frac{1}{n} \sum_{i=1}^{n} x_{ip} \right) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i.$$

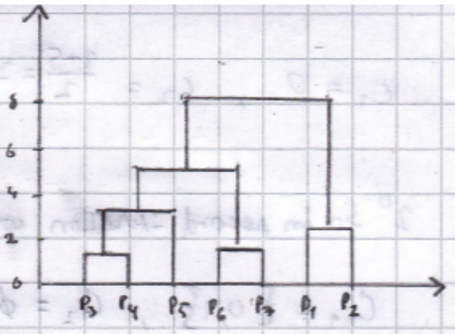
2.



	clusters	min. distance
1°	$\{P_1\}, \{P_2\}, \{P_3\}, \{P_4\}, \{P_5\}, \{P_6\}, \{P_7\}$	$\|P_4 - P_3\| = 1{,}36$ ↓
2°	$\{P_1\}, \{P_2\}, \{P_3, P_4\}, \{P_5\}, \{P_6\}, \{P_7\}$	$\|P_7 - P_6\| = 1{,}53$ ↓
3°	$\{P_1\}, \{P_2\}, \{P_3, P_4\}, \{P_5\}, \{P_6, P_7\}$	$\|P_2 - P_1\| = 2{,}24$ ↓
4°	$\{P_1, P_2\}, \{P_3, P_4\}, \{P_5\}, \{P_6, P_7\}$	$\|P_5 - P_2\| = 2{,}35$ ↓
5°	$\{P_1, P_2, P_5\}, \{P_3, P_4\}, \{P_6, P_7\}$	$\|P_6 - P_5\| = 2{,}42$ ↓
6°	$\{P_1, P_2, P_5, P_6, P_7\}, \{P_3, P_4\}$	$\|P_6 - P_4\| = 2{,}66$ ↓
7°	$\{P_1, P_2, P_3, P_4, P_5, P_6, P_7\}$	

b)





	clusters	shortest max distance
1°	$\{p_1\}, \{p_2\}, \{p_3\}, \{p_4\}, \{p_5\}, \{p_6\}, \{p_7\}$	$\|p_4 - p_3\| = 1,36$ ↓
2°	$\{p_1\}, \{p_2\}, \{p_3, p_4\}, \{p_5\}, \{p_6\}, \{p_7\}$	$\|p_7 - p_6\| = 1,53$ ↓
3°	$\{p_1\}, \{p_2\}, \{p_3, p_4\}, \{p_5\}, \{p_6, p_7\}$	$\|p_2 - p_1\| = 2,24$ ↓
4°	$\{p_1, p_2\}, \{p_3, p_4\}, \{p_5\}, \{p_6, p_7\}$	$\|p_5 - p_4\| = 3,28$ ↓
5°	$\{p_1, p_2\}, \{p_3, p_4, p_5\}, \{p_6, p_7\}$	$\|p_7 - p_5\| = 4,79$ ↓
6°	$\{p_1, p_2\}, \{p_3, p_4, p_5, p_6, p_7\}$	$\|p_7 - p_1\| = 8,04$ ↓
7°	$\{p_1, p_2, p_3, p_4, p_5, p_6, p_7\}$	

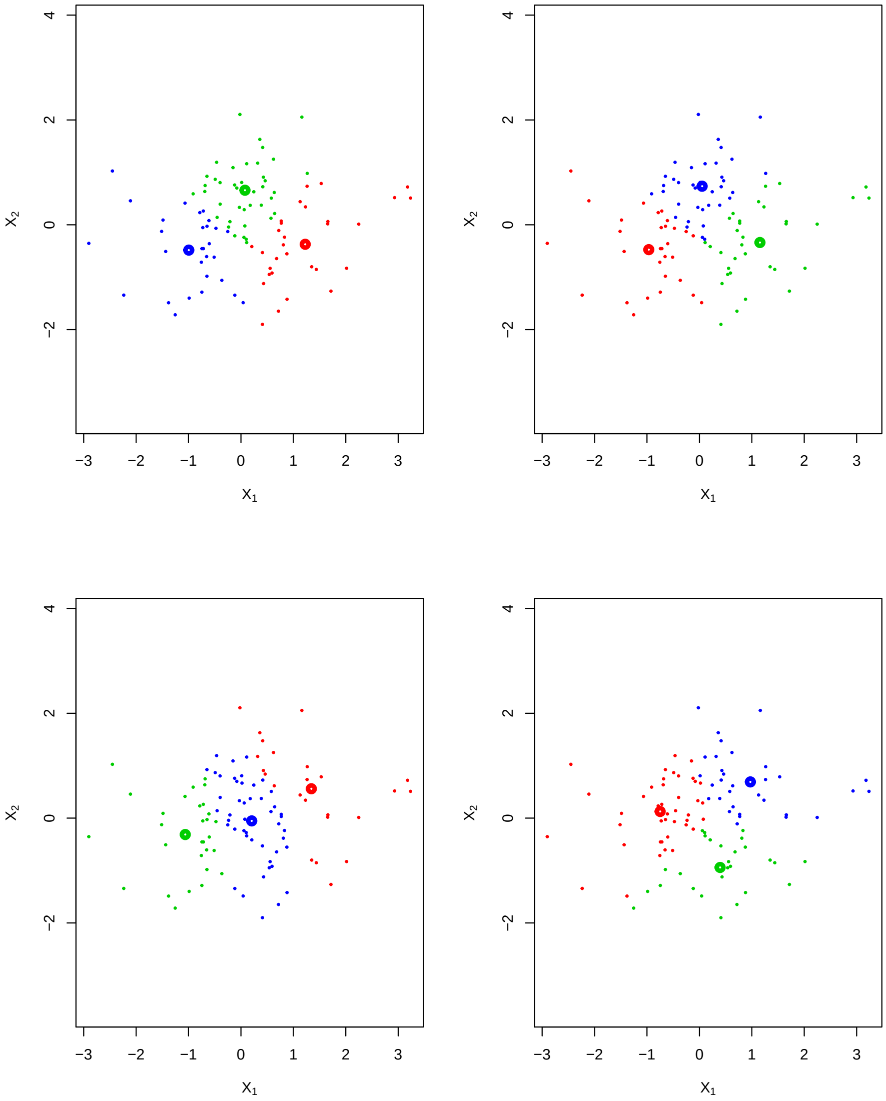
3.

(a) In Figure 1 one some results of the k-means algorithm on the same test data set with K = 3. Cluster centerpoints are marked as circles. Each time the algorithm converged into a different local minimum.

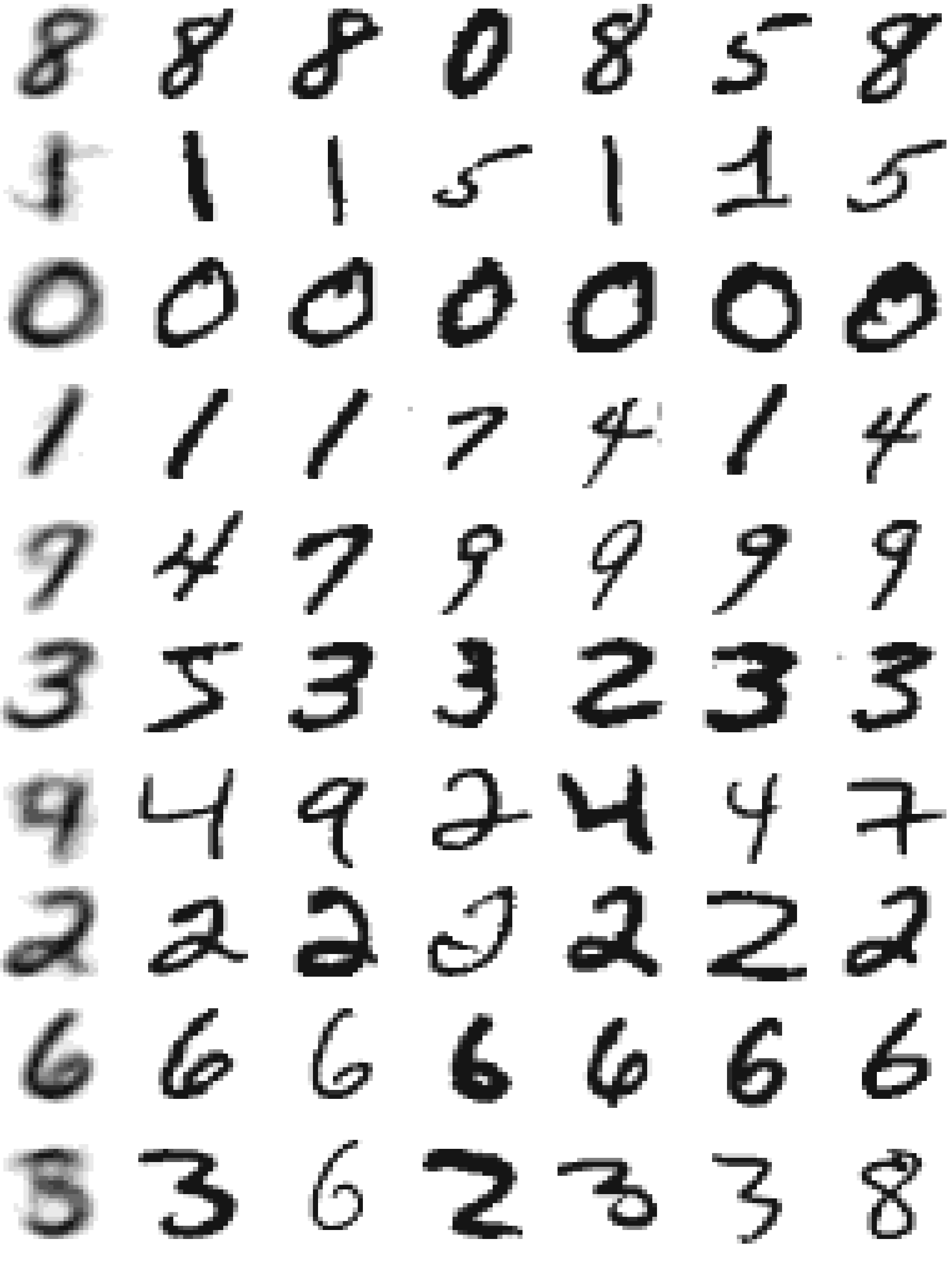
(b) In Figure 2 are shown both the prototypes and the examples of the digits belonging to the cluster.

Some of the clusters are more pure, for example the third cluster seems to consist of zeros, and the ninth cluster seems to consist of sixes, whereas most of the clusters consist of two or more digits. However, even in the mixed clusters the shapes of the digits resemble each other, with the exception of the second cluster, which is a mix of 1 and 5.

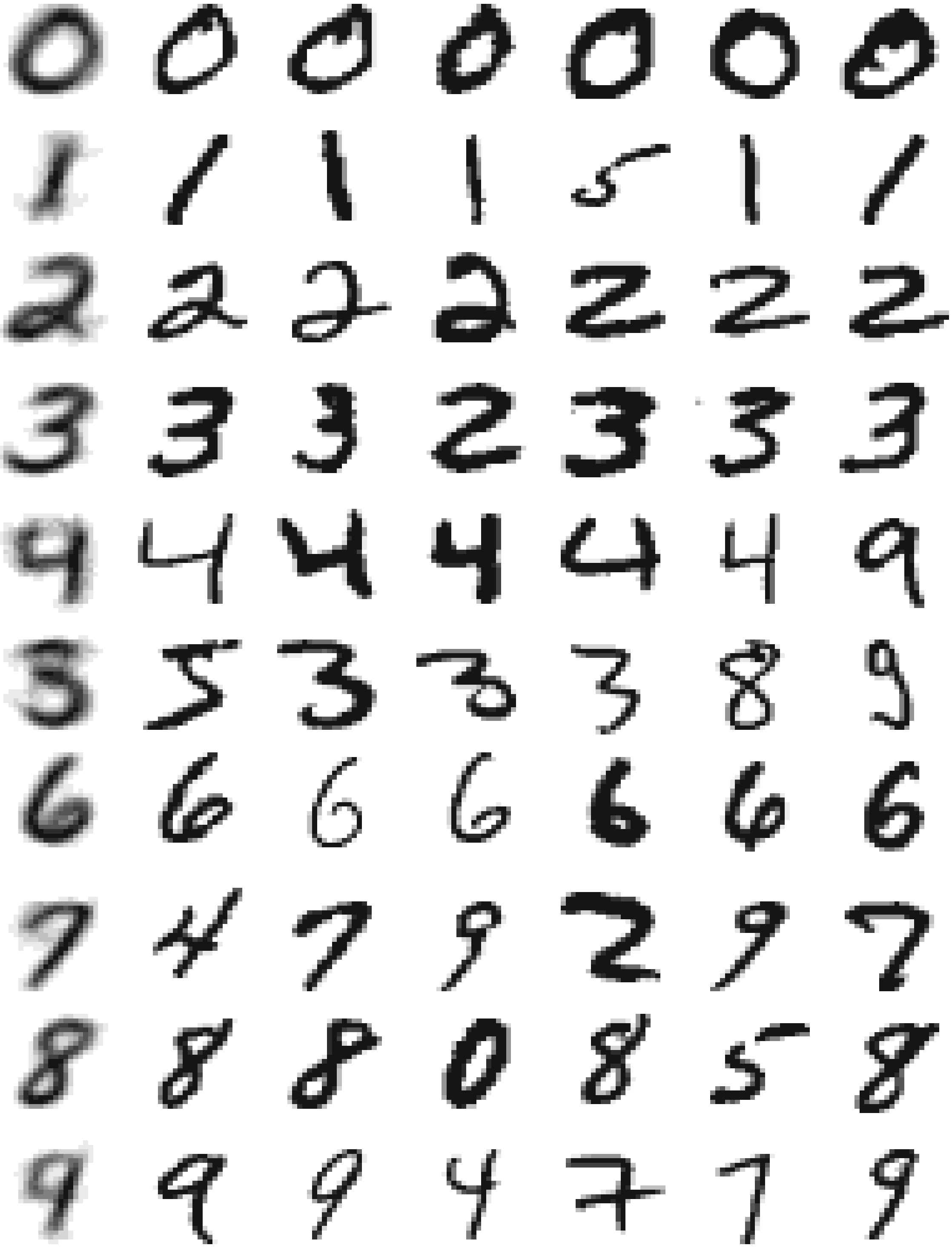
(c) As can be seen from Figure 3, now each cluster prototype resembles their initial values. Also the clusters seem to be more pure, and the algorithm converges faster (6 vs. 15 iterations). Although we 'cheated' here because we knew the true class labels, this demonstrates that k-means is sensitive to the choice of initial values, and its performance can be improved by choosing good initial values.



Kuva 1: Results of k-means algorithm with K = 3.



Kuva 2: Cluster prototypes and examples of the digits of the cluster for the mnist data set.



Kuva 3: Cluster prototypes and examples of the digits of the cluster for the mnist data set with centerpoints set to unique digits.