ACM/IDS 104 APPLIED LINEAR ALGEBRA
PRACTICE PROBLEMS FOR LECTURE 5

KONSTANTIN M. ZUEV

Each lecture is accompanied by two practice problems: a somewhat easier, more practical *Problem A*, and a more difficult, more conceptual *Problem B*. The main goal of the practice problems is threefold: to help you better understand the material covered in the corresponding lecture, to help you prepare to solve problems in problem sets and exams, and to accommodate the diversity of students' math backgrounds by providing both easier and more challenging problems. These problems are for self-practice: they will not be graded, and the solutions—posted on Piazza—also illustrate the expected level of rigor for problem sets and exams.

Problem 5A. K-Means Algorithm: Toy Example
Whenever you learn (or develop) a new algorithm, it is always very useful to explore how the algorithm works in simple, particular cases where the solution to the problem the algorithm is intended to solve is known. This illuminates the algorithm, makes it more intuitively clear, and serves as a common-sense check.

Suppose we want to partition the 12 blue points shown in Fig. 1 into $K = 4$ clusters. Four of the points, marked with red crosses, are the initial representatives of the clusters. Perform the k-means algorithm by hand and determine the resulting clusters.

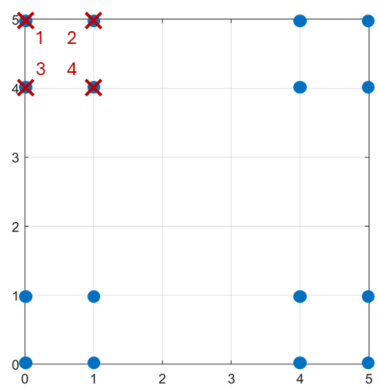

Figure 1. A cloud of (blue) points and four cluster representatives (red crosses).

Problem 5B. Convergence of the K-Means Algorithm
Suppose that, given n points $v_1, \dots, v_n$ in $\mathbb{R}^d$, we want to partition the points into K disjoint clusters using the K-means algorithm. Let's recall how the K-means algorithm works.

(1) Choose cluster representatives $r_1^{(0)}, \dots, r_K^{(0)}$ at random among $v_1, \dots, v_n$.
(2) Compute the first clustering $c^{(1)} = (c_1^{(1)}, \dots, c_n^{(1)})$ by assigning v_i to cluster number $c_i^{(1)}$, where

$$c_i^{(1)} = \arg \min_{k=1,\dots,K} \|v_i - r_k^{(0)}\|, \quad i = 1, \dots, n. \tag{1}$$

(3) Update representatives as follows:

$$r_k^{(1)} = \frac{1}{n_k} \sum_{i:c_i^{(1)}=k} v_i, \quad k = 1, \dots, K, \tag{2}$$

where n_k is the number of points in cluster k.

Then compute the second clustering $c^{(2)}$ using the updated representatives, and continue iterating Steps (2) and (3) until two consecutive clusterings are identical, that is, until $c^{(m)} = c^{(m-1)}$ for some m.

Prove the K-Means algorithm always converges in a finite number of steps.

E-mail address: `kostia@caltech.edu`.