MATH 3332 Data Analytic Tools
Homework 1

Due date: 28 September, 6pm, Monday

1. (a) Prove that the 1-norm defined by
$$\|\boldsymbol{x}\|_1 = \sum_{i=1}^{n} |x_i|, \qquad \forall \boldsymbol{x} \in \mathbb{R}^n$$
is indeed a norm on $\mathbb{R}^n$, i.e., prove $\|\cdot\|_1$ satisfies the conditions in the definition of norms.

   (b) For any $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, define
$$\|\boldsymbol{A}\|_{2\to 2} = \max_{\boldsymbol{x}\in\mathbb{R}^n,\ \|\boldsymbol{x}\|_2=1} \|\boldsymbol{A}\boldsymbol{x}\|_2.$$
Prove that $\|\cdot\|_{2\to 2}$ is a norm on $\mathbb{R}^{m\times n}$.

2. Let $(V, \|\cdot\|)$ be a normed vector space.

   (a) Prove that, for all $\boldsymbol{x}, \boldsymbol{y} \in V$,
$$|\|\boldsymbol{x}\| - \|\boldsymbol{y}\|| \le \|\boldsymbol{x} - \boldsymbol{y}\|.$$

   (b) Let $\{\boldsymbol{x}^{(k)}\}_{k\in\mathbb{N}}$ be a convergent sequence in V with limit $\boldsymbol{x} \in V$. Prove that
$$\lim_{k\to\infty} \|\boldsymbol{x}^{(k)}\| = \|\boldsymbol{x}\|.$$
(*Hint: Use part (a).*)

   (c) Let $\{\boldsymbol{x}^{(k)}\}_{k\in\mathbb{N}}$ be a sequence in V and $\boldsymbol{x}, \boldsymbol{y} \in V$. Prove that, if
$$\boldsymbol{x}^{(k)} \to \boldsymbol{x}, \quad \text{and} \quad \boldsymbol{x}^{(k)} \to \boldsymbol{y},$$
then $\boldsymbol{x} = \boldsymbol{y}$.

3. Let $a_1, a_2, \ldots, a_m$ be m given real numbers.

   (a) Prove that the mean of $a_1, a_2, \ldots, a_m$ minimizes
$$(a_1 - b)^2 + (a_2 - b)^2 + \ldots + (a_m - b)^2$$
over all $b \in \mathbb{R}$.

   (b) Prove that a median of $a_1, a_2, \ldots, a_m$ minimizes
$$|a_1 - b| + |a_2 - b| + \ldots + |a_m - b|$$
over all $b \in \mathbb{R}$.

4. Suppose that the vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ in $\mathbb{R}^n$ are clustered using the K-means/K-medians algorithm, with group representatives $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_k$.

(a) Suppose the original vectors $\boldsymbol{x}_i$ are nonnegative, i.e., their entries are nonnegative. Explain why the representatives $\boldsymbol{z}_j$ output by the K-means/K-medians algorithm are also nonnegative.

(b) Suppose the original vectors $\boldsymbol{x}_i$ represent proportions, i.e., their entries are nonnegative and sum to one. (This is the case when $\boldsymbol{x}_i$ are word count histograms, for example.) Explain why the representatives $\boldsymbol{z}_j$ output by the K-means algorithm are also represent proportions (i.e., their entries are nonnegative and sum to one), but $\boldsymbol{z}_j$ be the K-medians algorithm are not.

(c) Suppose the original vectors $\boldsymbol{x}_i$ are Boolean, i.e., their entries are either 0 or 1. Give an interpretation of $(\boldsymbol{z}_j)_i$, the i-th entry of the j group representative by K-means/K-medians algorithms.

5. *(You don't need to answer anything for this question.)* An interactive demonstration of K-means algorithm can be found at `http://alekseynp.com/viz/k-means.html`, where the K-means algorithm is also called *Lloyd's algorithm*. Generate data by "random clustered", and choose the same number of clusters in "Data Generation" and "K-means". You will see that the K-means algorithm converges to a correct clustering in most of the test examples. There do exist some test examples for which the K-means algorithm converges to a wrong clustering.