MSBD5004 Mathematical Methods for Data Analysis
Homework 1

Due date: 13 March, Friday

1. Consider the vector space $\mathbb{R}^n$.
   (a) Check that $\|\boldsymbol{x}\|_\infty = \max_{1\le i\le n} |x_i|$ is indeed a norm on $\mathbb{R}^n$.
   (b) Prove that: for any $\boldsymbol{x} \in \mathbb{R}^n$,
   $$\|\boldsymbol{x}\|_\infty = \lim_{p\to\infty} \|\boldsymbol{x}\|_p.$$
   (c) Prove the equivalence
   $$\|\boldsymbol{x}\|_\infty \le \|\boldsymbol{x}\|_1 \le n\|\boldsymbol{x}\|_\infty, \quad \forall \boldsymbol{x} \in \mathbb{R}^n.$$

2. For any $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, we have defined
   $$\|\boldsymbol{A}\|_2 = \sup_{\boldsymbol{x}\in\mathbb{R}^n,\ \boldsymbol{x}\neq\boldsymbol{0}} \frac{\|\boldsymbol{A}\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2}.$$
   (a) Prove that
   $$\|\boldsymbol{A}\|_2 = \max_{\boldsymbol{x}\in\mathbb{R}^n,\ \|\boldsymbol{x}\|_2=1} \|\boldsymbol{A}\boldsymbol{x}\|_2$$
   (b) Prove that $\|\cdot\|_2$ is a norm on $\mathbb{R}^{m\times n}$.
   (c) Prove that $\|\boldsymbol{A}\boldsymbol{x}\|_2 \le \|\boldsymbol{A}\|_2\|\boldsymbol{x}\|_2$ for any $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ and $\boldsymbol{x} \in \mathbb{R}^n$.
   (d) Prove that $\|\boldsymbol{A}\boldsymbol{B}\|_2 \le \|\boldsymbol{A}\|_2\|\boldsymbol{B}\|_2$ for all $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ and $\boldsymbol{B} \in \mathbb{R}^{n\times p}$.

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

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.