MSBD5004 Mathematical Methods for Data Analysis
Homework 1

Due date: 30 September, 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 inequality
   $$\|\boldsymbol{x}\|_\infty \le \|\boldsymbol{x}\|_1 \le n\|\boldsymbol{x}\|_\infty, \quad \forall \boldsymbol{x} \in \mathbb{R}^n.$$
2. For any $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, we have defined
   $$\|\boldsymbol{A}\|_2 = \max_{\boldsymbol{x}\in\mathbb{R}^n,\ \|\boldsymbol{x}\|_2=1} \|\boldsymbol{A}\boldsymbol{x}\|_2$$
   (a) Prove that $\|\cdot\|_2$ is a norm on $\mathbb{R}^{m\times n}$.
   (b) 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 $\boldsymbol{x} \in \mathbb{R}^n$.
   (c) 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
   $$\sum_{i=1}^{m} |a_i - b|$$
   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 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. Suppose that the vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ in $\mathbb{R}^n$ are clustered using the K-means algorithm, with groups $G_1, ..., G_k$ and group representatives $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_k$. Prove that the sequence of objective function values
   $$J = \sum_{j=1}^{k} \sum_{i\in G_j} \|\boldsymbol{x}_i - \boldsymbol{z}_j\|_2^2$$
   produced by the K-means algorithm is non-increasing.

6. Consider the set of infinite sequences. We have defined the vector space $\ell_\infty := \{\boldsymbol{a} \mid \|\boldsymbol{a}\|_\infty < +\infty\}$ with norm $\|\cdot\|_\infty$. For each $k \geq 1$, let $\boldsymbol{a}_k = \{\underbrace{1, 1, ..., 1}_{k \text{ times}}, 0, 0, \ldots\} \in \ell_\infty$. Prove that $\boldsymbol{a}_k \not\to \boldsymbol{b} := (1, 1, 1, ...)$ in ℓ_∞ with $\|\cdot\|_\infty$.

7. Let X be a vector space with norm $\|\cdot\|$. Prove the following results.

   (a) If $\{c_k\}$ is a convergent sequence in $\mathbb{R}$ and $\{\boldsymbol{a}_k\}$ is a convergent sequence in X with limits c and $\boldsymbol{a}$ respectively, then $\{c_k \cdot \boldsymbol{a}_k\}$ is a convergent sequence in X with limit $c \cdot \boldsymbol{a}$.

   (b) If $\{\boldsymbol{a}_k\}$ and $\{\boldsymbol{b}_k\}$ are convergent sequences in X with limits $\boldsymbol{a}$ and $\boldsymbol{b}$ respectively, then $\{\boldsymbol{a}_k + \boldsymbol{b}_k\}$ is a convergent sequence in X with limit $\boldsymbol{a} + \boldsymbol{b}$.