MSBD5004 Mathematical Methods for Data Analysis

Homework 1

Due date: Thursday, September 26

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 the inequality
   $$\|\boldsymbol{x}\|_\infty \le \|\boldsymbol{x}\|_1 \le n\|\boldsymbol{x}\|_\infty, \quad \forall \boldsymbol{x} \in \mathbb{R}^n.$$

   (c) Prove the inequality
   $$|\boldsymbol{x}^T\boldsymbol{y}| \le \|\boldsymbol{x}\|_1\|\boldsymbol{y}\|_\infty, \quad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n.$$

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. For any $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, we define the Frobenius norm $\|\boldsymbol{A}\|_F = \left(\sum_{i=1}^m \sum_{j=1}^n a_{i,j}^2\right)^{1/2}$. Prove that
   $$\|\boldsymbol{A}\|_2 \le \|\boldsymbol{A}\|_F \le \sqrt{n}\|\boldsymbol{A}\|_2.$$

4. A magic square $\boldsymbol{M}_n$ is a $n \times n$ matrix containing the integers from 1 to n^2 whose row and column sums are all the same. For example:
   $$\begin{bmatrix} 16 & 2 & 3 & 13 \\ 5 & 11 & 10 & 8 \\ 9 & 7 & 6 & 12 \\ 4 & 14 & 15 & 1 \end{bmatrix}.$$

   This magic square appears in the Renaissance engraving *Melencolia I* by the German painter, engraver, and amateur mathematician Albrecht Dürer (1471–1528).

   Let a_n denote the magic constant of $\boldsymbol{M}_n$, so that $a_n = n(n^2+1)/2$. Let $\boldsymbol{d}$ denote a vector in $\mathbb{R}^n$ with each element equal to 1.

   (a) Determine $\boldsymbol{M}_n\boldsymbol{d}$ and $\boldsymbol{d}^T\boldsymbol{M}_n$. Conclude that a_n is an eigenvalue of $\boldsymbol{M}_n$.

   (b) Show that the row and column sums of $\boldsymbol{M}_n^2$ are all the same.

   (c) Determine $\|\boldsymbol{M}_n\|_2$.

5. 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 $b \in \mathbb{R}$. *(As we discussed in the lecture, this result is crucial for deriving the K-medians algorithm in clustering.)*

6. 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.