



MATH 3332 Data Analytic Tools
Homework 1

Due date: 28 September, 6pm, Monday

1. (a) Prove that the 1-norm defined by

$$\|\mathbf{x}\|_1 = \sum_{i=1}^n |x_i|, \quad \forall \mathbf{x} \in \mathbb{R}^n$$

is indeed a norm on \mathbb{R}^n , i.e., prove $\|\cdot\|_1$ satisfies the conditions in the definition of norms.

- (b) For any $\mathbf{A} \in \mathbb{R}^{m \times n}$, define
- $$\|\mathbf{A}\|_{2 \rightarrow 2} = \max_{\mathbf{x} \in \mathbb{R}^n, \|\mathbf{x}\|_2=1} \|\mathbf{Ax}\|_2.$$

Prove that $\|\cdot\|_{2 \rightarrow 2}$ is a norm on $\mathbb{R}^{m \times n}$.

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

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

- (b) Let $\{\mathbf{x}^{(k)}\}_{k \in \mathbb{N}}$ be a convergent sequence in V with limit $\mathbf{x} \in V$. Prove that

$$\lim_{k \rightarrow \infty} \|\mathbf{x}^{(k)}\| = \|\mathbf{x}\|.$$

(Hint: Use part (a).)

- (c) Let $\{\mathbf{x}^{(k)}\}_{k \in \mathbb{N}}$ be a sequence in V and $\mathbf{x}, \mathbf{y} \in V$. Prove that, if

$$\mathbf{x}^{(k)} \rightarrow \mathbf{x}, \quad \text{and} \quad \mathbf{x}^{(k)} \rightarrow \mathbf{y},$$

then $\mathbf{x} = \mathbf{y}$.

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

- (a) Prove that the mean of a_1, a_2, \dots, a_m minimizes

$$(a_1 - b)^2 + (a_2 - b)^2 + \dots + (a_m - b)^2$$

over all $b \in \mathbb{R}$.

- (b) Prove that a median of a_1, a_2, \dots, a_m minimizes

$$|a_1 - b| + |a_2 - b| + \dots + |a_m - b|$$

over all $b \in \mathbb{R}$.

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

- (a) Suppose the original vectors \mathbf{x}_i are nonnegative, i.e., their entries are nonnegative. Explain why the representatives \mathbf{z}_j output by the K -means/ K -medians algorithm are also nonnegative.
- (b) Suppose the original vectors \mathbf{x}_i represent proportions, i.e., their entries are nonnegative and sum to one. (This is the case when \mathbf{x}_i are word count histograms, for example.) Explain why the representatives \mathbf{z}_j output by the K -means algorithm are also represent proportions (i.e., their entries are nonnegative and sum to one), but \mathbf{z}_j be the K -medians algorithm are not.
- (c) Suppose the original vectors \mathbf{x}_i are Boolean, i.e., their entries are either 0 or 1. Give an interpretation of $(\mathbf{z}_j)_i$, 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.

2

|.

1. (a) Prove that the 1-norm defined by

$$\|\mathbf{x}\|_1 = \sum_{i=1}^n |x_i|, \quad \forall \mathbf{x} \in \mathbb{R}^n$$

is indeed a norm on \mathbb{R}^n , i.e., prove $\|\cdot\|_1$ satisfies the conditions in the definition of norms.

- (b) For any $\mathbf{A} \in \mathbb{R}^{m \times n}$, define

$$\|\mathbf{A}\|_{2 \rightarrow 2} = \max_{\mathbf{x} \in \mathbb{R}^n, \|\mathbf{x}\|_2=1} \|\mathbf{Ax}\|_2.$$

Prove that $\|\cdot\|_{2 \rightarrow 2}$ is a norm on $\mathbb{R}^{m \times n}$.

1(a). ① $\forall \mathbf{x} \in \mathbb{R}^n, \|\mathbf{x}\|_1 = \sum_{i=1}^n |x_i| \geq 0$

and $\|\mathbf{x}\|_1 = 0$ if $\mathbf{x} = 0$

② $\|\alpha \mathbf{x}\|_1 = \sum_{i=1}^n |\alpha x_i| = \sum_{i=1}^n |\alpha| |x_i| = |\alpha| \sum_{i=1}^n |x_i|$

$$\begin{aligned} \forall \mathbf{x} \in \mathbb{R}^n, \quad &= |\alpha| \|\mathbf{x}\|_1 \\ \forall \alpha \in \mathbb{R}, \quad & \end{aligned}$$

③ $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^n$

$\forall x \in \mathbb{R}^n$

$$\textcircled{3} \quad \forall x, y \in \mathbb{R}^n,$$

$$\|x+y\|_1 = \sum_{i=1}^n |x_i + y_i|$$

$$\leq \sum_{i=1}^n |x_i| + \sum_{i=1}^n |y_i|$$

$$\leq \|x\|_1 + \|y\|_1$$

$\therefore \| \cdot \|_1$ is indeed a norm.

$$\text{b) } \textcircled{1} \quad \|A\|_{2 \rightarrow 2} = \max_{x \in \mathbb{R}^n, \|x\|_2=1} \|Ax\|_2$$

$$= \sqrt{\sum_{i=1}^n \left| \left(\sum_{j=1}^m a_{ij} x_j \right) \right|^2}$$

$$\geq 0$$

and $\|A\|_{2 \rightarrow 2} = 0$ when $A=0$

$$\textcircled{2} \quad \|\alpha A\|_{2 \rightarrow 2} = \max_{x \in \mathbb{R}^n, \|x\|_2=1} \|\alpha Ax\|_2$$

$$= \sqrt{\sum_{i=1}^n \left| \alpha \sum_{j=1}^m a_{ij} x_j \right|^2}$$

$$= |\alpha| \sqrt{\sum_{i=1}^n \left| \sum_{j=1}^m a_{ij} x_j \right|^2}$$

$$= |\alpha| \|A\|_{2 \rightarrow 2}$$

$$\textcircled{3} \quad \|A+B\|_{2 \rightarrow 2} = \max_{x \in \mathbb{R}^n, \|x\|_2=1} \|(A+B)x\|_2$$

$$= \max_{x \in \mathbb{R}^n, \|x\|_2=1} \|Ax + Bx\|_2$$

$$\begin{aligned}
\textcircled{3} \quad \|A + B\|_2 &= \max_{x \in \mathbb{R}^n, \|x\|_2=1} \|Ax + Bx\|_2 \\
&= \max_{x \in \mathbb{R}^n, \|x\|_2=1} \|Ax\|_2 + \|Bx\|_2 \\
&\leq \max_{x \in \mathbb{R}^n, \|x\|_2=1} \|Ax\|_2 + \max_{x \in \mathbb{R}^n, \|x\|_2=1} \|Bx\|_2 \\
&= \max_{x \in \mathbb{R}^n, \|x\|_2=1} \|Ax\|_2 + \|B\|_2 \\
&= \|A\|_2 + \|B\|_2
\end{aligned}$$

So, $\|\cdot\|_2$ is indeed a norm.

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

(a) Prove that, for all $x, y \in V$,

$$\|x\| - \|y\| \leq \|x - y\|.$$

(b) Let $\{x^{(k)}\}_{k \in \mathbb{N}}$ be a convergent sequence in V with limit $x \in V$. Prove that

$$\lim_{k \rightarrow \infty} \|x^{(k)}\| = \|x\|.$$

(Hint: Use part (a).)

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

$$x^{(k)} \rightarrow x, \quad \text{and} \quad x^{(k)} \rightarrow y,$$

then $x = y$.

2a). From triangle inequality:

$$\begin{aligned}
\|x - y\| + \|y\| &\geq \|x - y + y\| \\
\|x - y\| &\geq \|x\| - \|y\| \quad \textcircled{1}
\end{aligned}$$

$$\begin{aligned}
\|y - x\| + \|x\| &\geq \|y - x + x\| \\
\|y - x\| &\geq \|y\| - \|x\| \quad \textcircled{2}
\end{aligned}$$

From \textcircled{1} \& \textcircled{2},

$$|\|x\| - \|y\|| \leq \|x - y\|$$

(b) Let $\{\mathbf{x}^{(k)}\}_{k \in \mathbb{N}}$ be a convergent sequence in V with limit $\mathbf{x} \in V$. Prove that

b).

$$\lim_{k \rightarrow \infty} \|\mathbf{x}^{(k)}\| = \|\mathbf{x}\|.$$

(Hint: Use part (a).)

Since $\{\mathbf{x}^{(k)}\}_{k \in \mathbb{N}}$ is a convergent sequence,

$$\therefore \lim_{k \rightarrow \infty} \|\mathbf{x}^{(k)} - \mathbf{x}\| = 0$$

Using result of (a),

$$0 \leq |\|\mathbf{x}^{(k)}\| - \|\mathbf{x}\|| \leq \|\mathbf{x}^{(k)} - \mathbf{x}\|$$

$$0 \leq \lim_{k \rightarrow \infty} |\|\mathbf{x}^{(k)}\| - \|\mathbf{x}\|| \leq \lim_{k \rightarrow \infty} \|\mathbf{x}^{(k)} - \mathbf{x}\|$$

$$0 \leq \lim_{k \rightarrow \infty} |\|\mathbf{x}^{(k)}\| - \|\mathbf{x}\|| \leq 0$$

By sandwich rule,

$$\lim_{k \rightarrow \infty} \|\mathbf{x}^{(k)}\| - \|\mathbf{x}\| = 0$$

$$\lim_{k \rightarrow \infty} \|\mathbf{x}^{(k)}\| = \|\mathbf{x}\|$$

(c) Let $\{\mathbf{x}^{(k)}\}_{k \in \mathbb{N}}$ be a sequence in V and $\mathbf{x}, \mathbf{y} \in V$. Prove that, if

c).

$$\mathbf{x}^{(k)} \rightarrow \mathbf{x}, \quad \text{and} \quad \mathbf{x}^{(k)} \rightarrow \mathbf{y},$$

then $\mathbf{x} = \mathbf{y}$.

Suppose $\mathbf{x} \neq \mathbf{y}$,

$$\text{let } \epsilon = \frac{\|\mathbf{y} - \mathbf{x}\|}{2}$$

There exist N_1 s.t. if $n > N_1$ then

$$\|\mathbf{x}^{(n)} - \mathbf{x}\| < \epsilon$$

There exist N_2 s.t. if $n > N_2$ then

$$\|\mathbf{x}^{(n)} - \mathbf{y}\| < \epsilon$$

$\therefore \max(N_1, N_2)$, if $n > N$ then

$$|x - \infty| < \epsilon$$

Let $N = \max(N_1, N_2)$, if $n > N$ then

$$|x^N - x| < \epsilon \text{ and } |x^N - y| < \epsilon$$

But then by triangle inequality,

$$|y - x| \leq |x^N - y| + |x^N - x| < \frac{2}{10} |y - x|$$

which is impossible, so $x \neq y$ is false.

$$\therefore x = y.$$

3.

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

(a) Prove that the mean of a_1, a_2, \dots, a_m minimizes

$$(a_1 - b)^2 + (a_2 - b)^2 + \dots + (a_m - b)^2$$

over all $b \in \mathbb{R}$.

$$\text{mean of } a_1, a_2, \dots, a_m = \frac{\sum_{i=1}^m a_i}{m}$$

$$\text{let } f(a) = (a_1 - b)^2 + (a_2 - b)^2 + \dots + (a_m - b)^2$$

$$f(a) = \sum_{i=1}^m (a_i - b)^2$$

$$f(a) = \sum_{i=1}^m a_i^2 - 2a_i b + b^2$$

$$f(a) = \sum_{i=1}^m a_i^2 - \sum_{i=1}^m 2a_i b + \sum_{i=1}^m b^2$$

$$f(a) = \sum_{i=1}^m a_i^2 - 2b \sum_{i=1}^m a_i + mb^2 \quad \text{--- (1)}$$

Taking derivatives with respect to a_i in (1), we have:

$$f'(a) = 2 \sum_{i=1}^m a_i - 2bm$$

Putting $b = \frac{\sum_{i=1}^m a_i}{m}$ in $f'(x)$, we have:

$$f'(a) = 2 \sum_{i=1}^m a_i - \frac{2 \sum_{i=1}^m a_i (m)}{(m)} \\ = 0$$

And checking $f''(a)$, we get

$$m > 0 \quad - \quad - \quad - \quad - \quad f \quad m > 0$$

And convex if $f''(x) > 0$

$$f''(a) = 2m > 0 \text{ for } m > 0$$

∴ From the above result, $f(a)$ is concave upward for $m > 0$ and $f(a)$ is minimized at $b = \frac{\sum_{i=1}^m a_i}{m}$.
the mean of a_1, a_2, \dots, a_m .

b7.

(b) Prove that a median of a_1, a_2, \dots, a_m minimizes

$$|a_1 - b| + |a_2 - b| + \dots + |a_m - b|$$

over all $b \in \mathbb{R}$.

Let \tilde{x} be the median of a_1, a_2, \dots, a_m .
let $m > 1$.

From the definition of median, we know that median divides a_1, a_2, \dots, a_m into two groups, namely

$$S = \{a_s\} \text{ and } L = \{a_L\} \text{ where}$$

$$S < \tilde{x} < L \text{ and } \text{num}(S) = \text{num}(L)$$

For S ,

$$f(a_s) = \sum_{s \in S} |a_s - b|$$

when $b = \tilde{x}$,

$$f(a_s) = \sum_{s \in S} (\tilde{x} - a_s)$$

$$f'(a_s) = -1 \cdot \text{num}(S)$$

For L ,

$$f(a_L) = \sum_{L \in L} |a_L - b|$$

When $b = \infty$,

$$f(\alpha) = \sum_{l \in L} (\alpha_l - \tilde{x})$$

$$f'(\alpha_l) = 1 \cdot \text{num}(L)$$

$$\therefore \frac{d}{d\alpha_l} \sum_{i=1}^m |\alpha_i - \tilde{x}| = 0$$

Since the minimum of absolute function occurs when $f'(x) = 0$ or undefined,
So when $b = \tilde{x}$, it minimizes $\sum_{i=1}^m |\alpha_i - b|$.

4

4. Suppose that the vectors x_1, \dots, x_N in \mathbb{R}^n are clustered using the K -means/ K -medians algorithm, with group representatives z_1, \dots, z_k .

- Suppose the original vectors x_i are nonnegative, i.e., their entries are nonnegative. Explain why the representatives z_j output by the K -means/ K -medians algorithm are also nonnegative.
- Suppose the original vectors x_i represent proportions, i.e., their entries are nonnegative and sum to one. (This is the case when x_i are word count histograms, for example.) Explain why the representatives z_j output by the K -means algorithm are also represent proportions (i.e., their entries are nonnegative and sum to one), but z_j be the K -medians algorithm are not.
- Suppose the original vectors x_i are Boolean, i.e., their entries are either 0 or 1. Give an interpretation of $(z_j)_i$, the i -th entry of the j group representative by K -means/ K -medians algorithms.

a). if $x_i \geq 0, 1 \leq i \leq k$

$$\text{mean} = \frac{\sum_{i=1}^k x_i}{N} \geq 0$$

$$\text{median} = \begin{cases} \frac{x_{\frac{k}{2}} + x_{\frac{k}{2}+1}}{2} & \text{if } k \text{ is even} \\ x_{\frac{k}{2}} & \text{if } k \text{ is odd} \end{cases}$$

$$> 0$$

b). If each vector sums to one,

$1 \leq \dots \leq 1$ for all k .

$$l^T x_k = 1 \text{ for all } k.$$

Then for z_j output by $k\text{-mean}$

algorithm,

$$l^T z_j = \frac{1}{|G_j|} \sum_{k \in G_j} l^T x_k = \frac{|G_j|}{|G_j|} = 1$$

For $k\text{-median}$,

$$l^T z_j = \underset{i \in G_j}{\text{median}} (l^T x_i)$$

which may not = 1.

$\therefore k\text{-mean represents proportion but}$
 $k\text{-median does not.}$

C). The i -th entry of group representative z_j generated by $k\text{-mean}$ algorithm is the fraction of the vectors in group j that have i -th entry one. If it is equal to one, all vectors in the group have i -th entry one. If it is close to one, most of vectors in the group have i -th entry one. If it is zero, then no vectors in the group have i -th entry one.

The i th entry of group representative z_j generated by $k\text{-median}$ algorithm is the majority of the vectors in group j that have i -th entry one. If it is equal to one, more than half of vectors in the group have i -th entry one. If it is 0.5, then it implies there are even members in G_j and number of vectors with value 1 equals number of vectors with value 0 in G_j . If it is zero, then less than half vectors in the group have i -th entry one.