HW1

2020年9月22日　　下午 05:45

HW1

MATH 3332 Data Analytic Tools
Homework 1

Due date: 28 September, 6pm, Monday

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

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

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.

1.

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

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

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 $\mathbb{R}^{m\times n}$.

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

and $\|x\|_1 = 0$ if $x=0$

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

$\forall x\in\mathbb{R}^n$, $\forall \alpha\in\mathbb{R}$ $= |\alpha|\|x\|_1$

③ $\forall x,y \in \mathbb{R}^n$,

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

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

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

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

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

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

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

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

$$\geq 0$$

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

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

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

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

$$= |\alpha| \, \|A\|_{2\to 2}$$

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

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

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

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

$$= \|A\|_{2\to 2} + \|B\|_{2\to 2}$$

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

2. 2.

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

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

$$|\|\boldsymbol{x}\| - \|\boldsymbol{y}\|| \leq \|\boldsymbol{x} - \boldsymbol{y}\|.$$

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

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

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

$$|\|\boldsymbol{x}\| - \|\boldsymbol{y}\|| \leq \|\boldsymbol{x} - \boldsymbol{y}\|.$$

(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 $\boldsymbol{x} = \boldsymbol{y}$.

2a). From triangle inequality:

$$\|x-y\| + \|y\| \geq \|x-y+y\|$$
$$\|x-y\| \geq \|x\| - \|y\| \quad -①$$
$$\|y-x\| + \|x\| \geq \|y-x+x\|$$
$$\|y-x\| + \|x\| \geq \|y\|$$
$$\|y-x\| \geq \|y\| - \|x\| \quad -②$$

From ① & ②,

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

b). (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).*)

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

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

Using result of (a),

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

$$0 \leq \lim_{k\to\infty} \left| \|x^{(k)}\| - \|x\| \right| \leq \lim_{k\to\infty} \|x^{(k)} - x\|$$

$$0 \leq \left| \lim_{k\to\infty} \|x^{(k)}\| - \|x\| \right| \leq 0$$

By sandwich rule,

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

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

c). (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 $\boldsymbol{x} = \boldsymbol{y}$.

If $x^{(k)} \to x$ and $x^{(k)} \to y$,

If $x^{(k)} \to x$ and $x^{(k)} \to y$,

then $\lim_{k\to\infty} x^{(k)} = x$ and $\lim_{k\to\infty} x^{(k)} = y$

If $x \neq y$,

$$\lim_{k\to\infty} x^{(k)} = x \neq y \neq \lim_{k\to\infty} x^{(k)}$$

which is impossible,

so $x = y$.

3.

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 $b \in \mathbb{R}$.

mean of $a_1, a_2, \ldots a_m = \dfrac{\sum_{i=1}^{m} a_i}{m}$

Let $f(a) = (a_1-b)^2 + (a_2-b)^2 + \ldots (a_m-b)^2$

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

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

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

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

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

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

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

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

$$= 0$$

... we

$= 0$ (m)

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

$f''(a) = 2m > 0$ 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 $a_1, a_2, \ldots a_m$.

b].

(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 $b \in \mathbb{R}$.

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

From the definition of median, we know that median divides $a_1, a_2, \ldots, a_m$ into two groups, namely $S = \{a_s\}$ and $L = \{a_e\}$ where

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

For S,

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

When $b = \tilde{x}$,

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

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

For L,

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

When $b = \tilde{x}$,

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

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

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

Since the minimum of absolute function occurs when $f'(x) = 0$ or undefined,

So when $b = \tilde{x}$, it minimizes $\sum_{i=1}^{m} |a_i - b|$.

4

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

(a) 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.

(b) 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.

(c) 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 $x_i \geqslant 0$, $1 \leq i \leq k$

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

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

≥ 0

b). If each vector sums to one,

$1^T x_k = 1$ for all k.

then for z_j output by k-mean algorithm,

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

For k-median,

$$1^T z_j = \underset{i \in G_j}{\text{median}} (1^T x_i)$$

which may not $= 1$.

∴ k-mean represents proportion but k-median does not.

C).The i-th entry of group representative z_j generated by k-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 ith entry of group representative z_j generated by k-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.