B656 DATA MINING 3 February 2018

Homework 1 *Chuck Jia*

Problem 1

(a) The original document-term matrix $(m_{ij})_{n\times m}$ has raw counts as elements, i.e., m_{ij} is the count of a term j in document i. It can be inconvenient to directly use the raw counts to compare term frequencies from documents with different length. For example, a thick book of ten million words may contain 100 occurrences of the word “Python". While a short essay of 300 words has only 50 occurrences of “Python", it is clear that the word occurs more often in the essay than in the book, though the counts in the essay are smaller than in the book.

In the new document-term matrix $(x_{ij})_{n\times m}$, where $x_{ij} = \frac{m_{ij}}{m_i}\log\frac{n}{n_j}$, the $\frac{m_{ij}}{m_i}$ part transforms raw count m_{ij} into the frequency of the term i in the document j. This enables us to make direct comparisons across documents, even when documents have different word counts. The $\log\frac{n}{n_j}$ part penalizes words that appear in many documents, therefore fade out the effects of more common terms. This can be useful in many tasks. For example, if we want to determine the auther of an article by looking at word frequencies, terms "the" and "a" may not provide very useful information, as they appear too often in the English language.

(b) The $log\frac{n}{n_j}$ part gives lower weights on more common terms, and therefore “filters out" the common terms. But in some tasks, some of the common words can also provide much information. Compared to $x_{ij} = \frac{m_{ij}}{m_i}$ and $x_{ij} = m_{ij}$, we might lose some information from common terms.

Also, if a term j_0 appears in every document, then for every document i, $x_{ij_0} = 0$. Although the frequency of the term may vary greatly in different documents, which can potentially contain useful information, we loses all of the information on the term using this coding.

(c) If a term j_0 appears in only one document, then x_{ij_0} is 0 for all but one $i = i_0$, where i_0^{th} document is where the term appears. The value of $x_{i_0 j_0}$ is $\frac{\log n}{m_{i_0}}$.

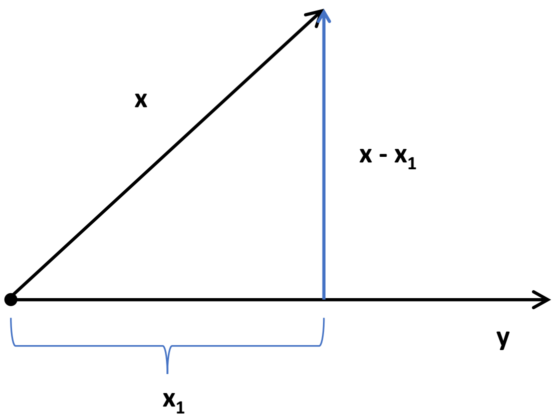
If a term j_0 appears in every document, then for every document i, $x_{ij_0} = 0$.

Problem 2

Assume that none of $\mathbf{x}$ and $\mathbf{y}$ are 0.

By the definition of cosine function, the length of the projection of $\mathbf{x}$ onto $\mathbf{y}$ is $\|x\|\cos(\mathbf{x},\mathbf{y})$, and its direction is along $\mathbf{y}$. Therefore, the projection of $\mathbf{x}$ onto $\mathbf{y}$ is

$$\mathbf{x}_1 = \text{proj}_{\mathbf{y}}\mathbf{x} = \|\mathbf{x}\|\cos(\mathbf{x},\mathbf{y})\cdot\frac{\mathbf{y}}{\|\mathbf{y}\|} = \frac{\|\mathbf{x}\|\cos(\mathbf{x},\mathbf{y})}{\|\mathbf{y}\|}\mathbf{y} \tag{E1}$$



The vector $\mathbf{x}$ and the projection $\mathbf{x}_1$, as two sides, form a right triangle, by definition of projections. The other side that is perpendicular to the projection can be written as the difference between $\mathbf{x}_1$ and $\mathbf{x}$. Therefore,

$$(\mathbf{x}-\mathbf{x}_1) \perp \mathbf{y}$$

which implies

$$(\mathbf{x}-\mathbf{x}_1)^T\mathbf{y} = 0$$

Using (E1),

$$\left(\mathbf{x} - \frac{||\mathbf{x}||\cos(\mathbf{x},\mathbf{y})}{||\mathbf{y}||}\mathbf{y}\right)^T \mathbf{y} = 0 \tag{E2}$$

Note that the left hand side

$$\begin{aligned}
\text{LHS of (E2)} =& \mathbf{x}^T\mathbf{y} - \left(\frac{||\mathbf{x}||\cos(\mathbf{x},\mathbf{y})}{||\mathbf{y}||}\mathbf{y}\right)^T \mathbf{y} \\
=& \mathbf{x}^T\mathbf{y} - \left(\frac{||\mathbf{x}||\cos(\mathbf{x},\mathbf{y})}{||\mathbf{y}||}\mathbf{y}^T\right) \mathbf{y} \\
=& \mathbf{x}^T\mathbf{y} - \frac{||\mathbf{x}||\cos(\mathbf{x},\mathbf{y})}{||\mathbf{y}||}\mathbf{y}^T \mathbf{y} \\
=& \mathbf{x}^T\mathbf{y} - \frac{||\mathbf{x}||\cos(\mathbf{x},\mathbf{y})}{||\mathbf{y}||}||\mathbf{y}||^2 \\
=& \mathbf{x}^T\mathbf{y} - ||\mathbf{x}||\,||\mathbf{y}||\cos(\mathbf{x},\mathbf{y})
\end{aligned}$$

Using this back in (E2), we have

$$\mathbf{x}^T\mathbf{y} - ||\mathbf{x}||\,||\mathbf{y}||\cos(\mathbf{x},\mathbf{y}) = 0$$

That is,

$$\mathbf{x}^T\mathbf{y} = ||\mathbf{x}||\,||\mathbf{y}||\cos(\mathbf{x},\mathbf{y})$$

□

Problem 3

Assume all sets are finite and non-empty.

Definition 1. A *metric* on a set X is a function $d : X \times X \to \mathbb{R}$ having the following properties:

(1) $d(x,y) \geq 0$ for all $x, y \in X$;

(2) $d(x,y) = 0$ if and only if $x = y$, for all $x, y \in X$;

(3) $d(x,y) = d(y,x)$, for all $x, y \in X$;

(4) (Triangle Inequality) $d(x,y) + d(y,z) \geq d(x,z)$, for all $x, y, z \in X$

(a) d_1 is a metric.

Proof. Let A, B, and C be any 3 sets from space $\mathcal{S}$. We verify the four conditions in Definition 1 above.

(1) Since the cardinality of any set is always non-negative, $|A-B| \geq 0$ and $|B-A| \geq 0$. Thus $d_1(A,B) = |A-B| + |B-A| \geq 0$.

(2) We prove $d_1(A,B) = 0 \Leftrightarrow A = B$ from two directions.

- ($\Rightarrow$) If $d_1(A,B) = 0$, then $|A-B| + |B-A| = 0$. Since $|A-B| \geq 0$ and $|B-A| \geq 0$, $|A-B| = |B-A| = 0$.[1] Thus, $A-B = B-A = \varnothing$. Note that $A \cup B = (A \cap B) \cup (A-B) \cup (B-A) = A \cap B$. Therefore, $A = B$.
- ($\Leftarrow$) If $A = B$, then $A-B = B-A = \varnothing$. Therefore, $d_1(A,B) = |A-B| + |B-A| = 0$.

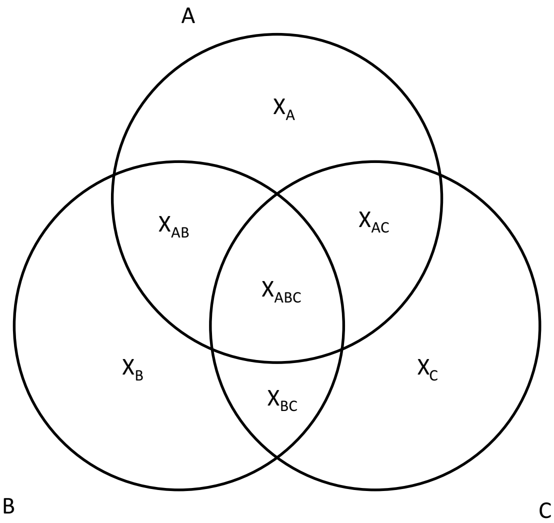
(3) Now we prove the symmetry condition.

$$d_1(A,B) = |A-B| + |B-A| = |B-A| + |A-B| = d_1(B,A)$$

(4) Now we prove the triangle inequality. We partition $A \cup B \cup C$ into X_A, X_B, $\cdots$, X_{ABC} 7 disjoint subsets, as shown below in the Venn diagram.

$$\begin{aligned}
& d_1(A,B) + d_1(B,C) \\
=& |A-B| + |B-A| + |B-C| + |C-B| \\
=& |X_A \cup X_{AC}| + |X_B \cup X_{BC}| + |X_B \cup X_{AB}| + |X_C \cup X_{AC}| \\
=& |X_A| + |X_{AC}| + |X_B| + |X_{BC}| + |X_B| + |X_{AB}| + |X_C| + |X_{AC}| \\
& \qquad \rightsquigarrow \text{Since all set here are mutually disjoint} \\
=& (|X_A| + |X_{AB}|) + (|X_C| + |X_{BC}|) + 2(|X_{AC}| + |X_B|) \\
=& |X_A \cup X_{AB}| + |X_C \cup X_{BC}| + 2(|X_{AC}| + |X_B|) \quad \rightsquigarrow \text{By disjointness} \\
=& |A-C| + |C-A| + 2(|X_{AC}| + |X_B| \\
\geq& |A-C| + |C-A| \\
=& d_1(A,C)
\end{aligned}$$

[1] If either $|A-B|$ or $|B-A|$ or both are strictly positive, then their sum would be greater than 0, which contradicts our assumption that $d_1(A,B) = |A-B| + |B-A| = 0$. Therefore, $|A-B| = |B-A| = 0$.



The 3 circles represent the sets A, B and C. The $X_{subscript}$ sets represent disjoint sections.

(b) d_2 is a metric.

Proof. We verify the four conditions in Definition 1 above.

(1) Since cardinality of a set is always non-negative, $|A-B| \geq 0$, $|B-A| \geq 0$, and $|A \cup B| > 0$ (by our assumption that A and B are not empty). Therefore, $d_2(A,B) = \frac{|A-B|+|B-A|}{|A \cup B|} \geq 0$.

(2) We prove $d_2(A,B) = 0 \Leftrightarrow A = B$ from two directions.

- ($\Rightarrow$) If $d_2(A,B) = \frac{|A-B|+|B-A|}{|A \cup B|} = 0$, then $|A-B|+|B-A| = 0$. Using exactly the same argument as in part (2) of the proof in part (a), we have $A = B$.
- ($\Leftarrow$) If $A = B$, then $A - B = B - A = \varnothing$, and hence $|A-B| = |B-A| = 0$. Therefore, $d_2(A,B) = \frac{|A-B|+|B-A|}{|A \cup B|} = 0$.

(3) We verify the symmetry condition. By commutativity of addition between real numbers and the union operation,

$$d_2(A,B) = \frac{|A-B|+|B-A|}{|A \cup B|} = \frac{|B-A|+|A-B|}{|B \cup A|} = d_2(B,A)$$

(4) Now we prove the triangle inequality.

First, we prove the following 2 lemmas.

Lemma 1. Let X and Y be any two sets from space $\mathcal{S}$. Then,

$$d_2(X,Y) = 2 \cdot \frac{d_1(X,Y)}{|X|+|Y|+d_1(X,Y)} \tag{E3}$$

where $d_1(X,Y) = |X-Y|+|Y-X|$ is the metric we used in part (a).

Proof of Lemma 1. Note that $X \cup Y = (X-Y) \cup (Y-X) \cup (X \cap Y)$, in which $X-Y$, $Y-X$, and $X \cap Y$

are mutually disjoint sets. Thus,

$$
\begin{aligned}
|X\cup Y| =& |X-Y|+|Y-X|+|X\cap Y| \\
=& \frac{1}{2}\Big(|X-Y|+|X-Y|+|Y-X|+|Y-X|+|X\cap Y|+|X\cap Y|\Big) \\
& \quad \rightsquigarrow \text{Duplicate all terms and divide by 2} \\
=& \frac{1}{2}\Big[(|X-Y|+|X\cap Y|)+(|Y-X|+|X\cap Y|)+|X-Y|+|Y-X|\Big] \\
=& \frac{1}{2}\Big(|X|+|Y|+|X-Y|+|Y-X|\Big)
\end{aligned}
\tag{E4}
$$

Therefore,

$$
\begin{aligned}
d_2(X,Y) =& \frac{|X-Y|+|Y-X|}{|X\cup Y|} \\
=& \frac{|X-Y|+|Y-X|}{\frac{1}{2}\big(|X|+|Y|+|X-Y|+|Y-X|\big)} \quad \rightsquigarrow \text{By (E4)} \\
=& 2\cdot\frac{d_1(X,Y)}{|X|+|Y|+d_1(X,Y)}
\end{aligned}
$$

□ *(End of proof for Lemma 1)*

Lemma 2. Let x, y, and c be positive real numbers. If $y \geq x$, then

$$\frac{x}{c+x} \leq \frac{y}{c+y}$$

Proof of Lemma 2. Computing directly,

$$\frac{y}{c+y}-\frac{x}{c+x}=\frac{cy+xy-(cx+xy)}{(c+x)(y+c)}=\frac{c(y-x)}{(c+x)(c+y)}\geq 0$$

□ *(End of proof for Lemma 2)*

Now we prove the main result of triangle inequality.

As a direct result of Lemma 1,

$$d_2(A,C) = 2\cdot\frac{d_1(A,C)}{|A|+|C|+d_1(A,C)} \tag{E5}$$

Since d_1 is a metric (proved in (a)), by the triangle inequality,

$$d_1(A,C) \leq d_1(A,B)+d_1(B,C) \tag{E6}$$

Now using this inequality and Lemma 2 in (E5),

$$\begin{aligned} d_2(A,C) =& 2\cdot\frac{d_1(A,C)}{|A|+|C|+d_1(A,C)} \\ \leq& 2\cdot\frac{d_1(A,B)+d_1(B,C)}{|A|+|C|+d_1(A,B)+d_1(B,C)} \\ & \quad \rightsquigarrow \text{By Lemma 2 with } x=d_1(A,C),\ y=d_1(A,B)+d_1(B,C) \\ =& 2\cdot\left(\frac{d_1(A,B)}{|A|+|C|+d_1(A,B)+d_1(B,C)}+\frac{d_1(B,C)}{|A|+|C|+d_1(A,B)+d_1(B,C)}\right) \end{aligned} \tag{E7}$$

Note that

$$\begin{aligned} |A|+|C|+d_1(B,C) =& |A|+|C|+|B-C|+|C-B| \\ =& |A|+(|C|+|B-C|)+|C-B| \\ =& |A|+|B\cup C|+|C-B| \\ \geq& |A|+|B\cup C| \\ \geq& |A|+|B| \end{aligned} \tag{E8}$$

Similarly,

$$|A|+|C|+d_1(A,B) \geq |B|+|C| \tag{E9}$$

Back to (E7),

$$\begin{aligned} d_2(A,C) =& 2\cdot\left(\frac{d_1(A,B)}{|A|+|C|+d_1(A,B)+d_1(B,C)}+\frac{d_1(B,C)}{|A|+|C|+d_1(A,B)+d_1(B,C)}\right) \\ =& 2\cdot\left(\frac{d_1(A,B)}{\big[|A|+|C|+d_1(B,C)\big]+d_1(A,B)}+\frac{d_1(B,C)}{\big[|A|+|C|+d_1(A,B)\big]+d_1(B,C)}\right) \\ \leq& 2\cdot\left(\frac{d_1(A,B)}{|A|+|B|+d_1(A,B)}+\frac{d_1(B,C)}{|B|+|C|+d_1(B,C)}\right) \quad \rightsquigarrow \text{By (E8) and (E9)} \\ =& d_2(A,B)+d_2(B,C) \quad \rightsquigarrow \text{By Lemma 1} \end{aligned} \tag{E10}$$

□ *(End of proof for triangle inequality)*

(c) d_3 is not a metric.

Counterexample:

Let $A=\{1\}$, $C=\{2\}$, and $B=\{1,2\}=A\cup C$. Then

$$d_3(A,B)=1-\left(\frac{1}{2}\cdot\frac{|A\cap B|}{|A|}+\frac{1}{2}\cdot\frac{|A\cap B|}{|B|}\right)=1-\left(\frac{1}{2}\cdot\frac{1}{1}+\frac{1}{2}\cdot\frac{1}{2}\right)=\frac{1}{4}$$

$$d_3(B,C)=1-\left(\frac{1}{2}\cdot\frac{|B\cap C|}{|B|}+\frac{1}{2}\cdot\frac{|B\cap C|}{|C|}\right)=1-\left(\frac{1}{2}\cdot\frac{1}{2}+\frac{1}{2}\cdot\frac{1}{1}\right)=\frac{1}{4}$$

$$d_3(A,C)=1-\left(\frac{1}{2}\cdot\frac{|A\cap C|}{|A|}+\frac{1}{2}\cdot\frac{|A\cap C|}{|C|}\right)=1-\left(\frac{1}{2}\cdot\frac{0}{1}+\frac{1}{2}\cdot\frac{0}{1}\right)=1$$

Therefore,

$$d_3(A,B)+d_3(B,C)=\frac{1}{4}+\frac{1}{4}=\frac{1}{2}<1=d_3(A,C)$$

which implies that d_3 does not satisfy the triangle inequality and hence it is not a metric.

(d) d_4 is not a metric.
Counterexample: (Same as in (c))
Let $A = \{1\}$, $C = \{2\}$, and $B = \{1,2\} = A \cup C$. Then

$$d_4(A,B) = 1 - \left(\frac{1}{2}\cdot\frac{|A|}{|A\cap B|} + \frac{1}{2}\cdot\frac{|B|}{|A\cap B|}\right)^{-1} = 1 - \left(\frac{1}{2}\cdot\frac{1}{1} + \frac{1}{2}\cdot\frac{2}{1}\right)^{-1} = \frac{1}{3}$$

$$d_4(B,C) = 1 - \left(\frac{1}{2}\cdot\frac{|B|}{|B\cap C|} + \frac{1}{2}\cdot\frac{|C|}{|B\cap C|}\right)^{-1} = 1 - \left(\frac{1}{2}\cdot\frac{2}{1} + \frac{1}{2}\cdot\frac{1}{1}\right)^{-1} = \frac{1}{3}$$

$$\begin{aligned} d_4(A,C) =& 1 - \left(\frac{1}{2}\cdot\frac{|A|}{|A\cap C|} + \frac{1}{2}\cdot\frac{|C|}{|A\cap C|}\right)^{-1} \\ =& 1 - \left(\frac{1}{2}\frac{|A|+|C|}{|A\cap C|}\right)^{-1} \\ =& 1 - \frac{2|A\cap C|}{|A|+|C|} \\ =& 1-0 \\ =& 1 \end{aligned}$$

Therefore,

$$d_4(A,B) + d_4(B,C) = \frac{1}{3} + \frac{1}{3} = \frac{2}{3} < 1 = d_4(A,C)$$

which implies that d_4 does not satisfy the triangle inequality and hence it is not a metric.

Problem 4

(a) d is a metric for $p \geq 1$.
Proof. First, we define the function $\|\cdot\| : \mathbb{R}^k \to \mathbb{R}$ as

$$\|x\| = \left[\left(\sum_{i:x_i>0} x_i\right)^p + \left(\sum_{i:x_i<0} (-x_i)\right)^p\right]^{1/p}, \quad \forall x = (x_1,\cdots,x_k) \in \mathbb{R}^k,\ p \geq 1 \tag{E11}$$

Note that $d(x,y)$ can be written as

$$d(x,y) = \|x-y\| \tag{E12}$$

for all $x = (x_1,\cdots,x_k) \in \mathbb{R}^k$ and $y = (y_1,\cdots,y_k) \in \mathbb{R}^k$.

◎ OUTLINE OF THE PROOF.
The proof we give below will consist of 2 parts:

- Part 1. We prove that $\|\cdot\|$ is a norm.
- Part 2. We show that d is indeed a metric.

A few words on the design choices:

- Part 2 of the proof is in fact trivial, given the result in part 1. It is a general result that any norm $\|\cdot\|$ naturally induces a metric d in the same space by defining $d(x,y)=\|x-y\|$. (E12) implies that our d is the induced metric from the norm $\|\cdot\|$. In part 2, we will prove that d is a metric, but notice that the proof applies to any norm, regardless of how it is defined.
- The reason why we choose to work on the norm instead of directly on the metric is because it greatly simplifies our notations, especially in the proof of the triangle inequality. As shown below in the proof, by working with the norm, we do not need to deal with differences $x_i - y_i$ in the definition of our metric, which makes our argument more clear and concise.[2]

◎ PROOF PART 1.

Definition 2. Let Ω be a vector space over field $F=\mathbb{R}$ or $\mathbb{C}$. A function $\|\cdot\|:\Omega\to\mathbb{R}$ is a *norm* if

(1) $\|x\|\geq 0$ for all $x\in\Omega$.

(2) If $\|x\|=0$, then $x=0$.

(3) $\|\lambda x\|=|\lambda|\cdot\|x\|$, for all $\lambda\in F$ and $x\in\Omega$.

(4) (Triangle inequality) $\|x+y\|\leq\|x\|+\|y\|$ for all $x,y\in\Omega$.

We show that $\|\cdot\|$ defined in (E11) is a norm by verifying all four conditions in Definition 2 with $\Omega=\mathbb{R}^k$ and $F=\mathbb{R}$.

(1) For any $x\in\mathbb{R}^k$, since $\sum_{i:x_i>0}x_i\geq 0$ and $\sum_{i:x_i<0}(-x_i)\geq 0$, we have

$$\|x\| = \left[\left(\sum_{i:x_i>0} x_i\right)^p + \left(\sum_{i:x_i<0}(-x_i)\right)^p\right]^{1/p} \geq 0$$

(2) If $\|x\|=0$, then

$$\left[\left(\sum_{i:x_i>0} x_i\right)^p + \left(\sum_{i:x_i<0}(-x_i)\right)^p\right]^{1/p} = 0$$

which implies

$$\left(\sum_{i:x_i>0} x_i\right)^p + \left(\sum_{i:x_i<0}(-x_i)\right)^p = 0 \tag{E13}$$

Since $\sum_{i:x_i>0}x_i\geq 0$ and $\sum_{i:x_i<0}(-x_i)\geq 0$,

$$\sum_{i:x_i>0} x_i = \sum_{i:x_i<0}(-x_i) = 0 \tag{E14}$$

[2]Mathematically speaking, all the argument we make below can be made equivalently on d, by using $\|x\|=d(x,0)$, barring some small changes in technicality. However, the norm is a convenient notation that makes the argument more readable and focused on the key ideas.

(Similar argument as before: if either or both of them is not 0, then at least one of $\left(\sum_{i:x_i>0} x_i\right)^p$ and $\left(\sum_{i:x_i<0}(-x_i)\right)^p$ would be greater than 0, which implies $\left(\sum_{i:x_i>0} x_i\right)^p + \left(\sum_{i:x_i<0}(-x_i)\right)^p > 0$, contradicting (E13).)

Since all the summands $\{x_i : x_i > 0\}$ and $\{-x_i : x_i < 0\}$ in (E14) are positive, there cannot be any x_i such that $x_i > 0$ or $x_i < 0$. Therefore, $x_i = 0$ for all $i = 1, \cdots, k$, i.e. $x_i = 0$.

(3) For any $x \in \mathbb{R}^k$,

- If $\lambda > 0$, then λx_i have the same sign with x_i all i. Thus,

$$\begin{aligned}
\|\lambda x\| =& \|(\lambda x_1, \cdots, \lambda x_k)\| = \left[\left(\sum_{i:\lambda x_i>0}(\lambda x_i)\right)^p + \left(\sum_{i:\lambda x_i<0}(-\lambda x_i)\right)^p\right]^{1/p} \\
=& \left[\left(\sum_{i:x_i>0}(\lambda x_i)\right)^p + \left(\sum_{i:x_i<0}(-\lambda x_i)\right)^p\right]^{1/p} \\
=& \left[\lambda^p \left(\sum_{i:x_i>0} x_i\right)^p + \lambda^p \left(\sum_{i:x_i<0}(-x_i)\right)^p\right]^{1/p} \\
=& \lambda \left[\left(\sum_{i:x_i>0} x_i\right)^p + \left(\sum_{i:x_i<0}(-x_i)\right)^p\right]^{1/p} \\
=& \lambda \|x\| = |\lambda| \cdot \|x\|
\end{aligned}$$

- If $\lambda < 0$, then λx_i have the opposite sign with x_i all i.

$$\begin{aligned}
\|\lambda x\| =& \|(\lambda x_1, \cdots, \lambda x_k)\| = \left[\left(\sum_{i:\lambda x_i>0}(\lambda x_i)\right)^p + \left(\sum_{i:\lambda x_i<0}(-\lambda x_i)\right)^p\right]^{1/p} \\
=& \left[\left(\sum_{i:x_i<0}(\lambda x_i)\right)^p + \left(\sum_{i:x_i>0}(-\lambda x_i)\right)^p\right]^{1/p} \\
=& \left[(-\lambda)^p \left(\sum_{i:x_i<0}(-x_i)\right)^p + (-\lambda)^p \left(\sum_{i:x_i>0} x_i\right)^p\right]^{1/p} \\
=& (-\lambda) \left[\left(\sum_{i:x_i<0}(-x_i)\right)^p + \left(\sum_{i:x_i>0} x_i\right)^p\right]^{1/p} \\
=& |\lambda| \cdot \|x\|
\end{aligned}$$

- If $\lambda = 0$, then $\|\lambda x\| = \|0\| = 0 = |\lambda| \cdot \|x\|$.

(4) Now we prove the triangle inequality. Let $x, y \in \mathbb{R}^k$.

$$
\begin{aligned}
&\|x+y\| \\
&= \left[\left(\sum_{i:x_i+y_i>0}(x_i+y_i)\right)^p + \left(\sum_{i:x_i+y_i<0}(-x_i-y_i)\right)^p\right]^{1/p} \\
&= \left[\left(\sum_{i:x_i+y_i>0}x_i + \sum_{i:x_i+y_i>0}y_i\right)^p + \left(\sum_{i:x_i+y_i<0}(-x_i) + \sum_{i:x_i+y_i<0}(-y_i)\right)^p\right]^{1/p} \\
&= \Bigg[\left(\sum_{\substack{i:x_i+y_i>0\\x_i>0}}x_i + \sum_{\substack{i:x_i+y_i>0\\x_i<0}}x_i + \sum_{\substack{i:x_i+y_i>0\\y_i>0}}y_i + \sum_{\substack{i:x_i+y_i>0\\y_i<0}}y_i\right)^p \\
&\qquad + \left(\sum_{\substack{i:x_i+y_i<0\\x_i>0}}(-x_i) + \sum_{\substack{i:x_i+y_i<0\\x_i<0}}(-x_i) + \sum_{\substack{i:x_i+y_i<0\\y_i>0}}(-y_i) + \sum_{\substack{i:x_i+y_i<0\\y_i<0}}(-y_i)\right)^p\Bigg]^{1/p} \\
&\le \left[\left(\sum_{\substack{i:x_i+y_i>0\\x_i>0}}x_i + \sum_{\substack{i:x_i+y_i>0\\y_i>0}}y_i\right)^p + \left(\sum_{\substack{i:x_i+y_i<0\\x_i<0}}(-x_i) + \sum_{\substack{i:x_i+y_i<0\\y_i<0}}(-y_i)\right)^p\right]^{1/p}
\end{aligned}
\tag{E15}
$$

$\rightsquigarrow$ In the step above, we threw out all the negative terms:

$$\sum_{\substack{i:x_i+y_i>0\\x_i<0}}x_i,\quad \sum_{\substack{i:x_i+y_i>0\\y_i<0}}y_i,\quad \sum_{\substack{i:x_i+y_i<0\\x_i>0}}(-x_i), \text{ and } \sum_{\substack{i:x_i+y_i<0\\y_i>0}}(-y_i)$$

Now for convenience and clarity, we use the following notation

$$a_1 = \sum_{\substack{i:x_i+y_i>0\\x_i>0}}x_i, \quad a_2 = \sum_{\substack{i:x_i+y_i<0\\x_i<0}}(-x_i), \quad b_1 = \sum_{\substack{i:x_i+y_i>0\\y_i>0}}y_i, \quad b_2 = \sum_{\substack{i:x_i+y_i<0\\y_i<0}}(-y_i)$$

Using these notations, (E15) can be written as

$$\|x+y\| = \Big((a_1+b_1)^p + (a_2+b_2)^p\Big)^{1/p} \tag{E16}$$

Now by the sum version of the Minkowski inequality (i.e. the integral Minkowski inequality with counting measure),

$$
\begin{aligned}
\|x+y\| &= \Big((a_1+b_1)^p + (a_2+b_2)^p\Big)^{1/p} \\
&= \Big(|a_1+b_1|^p + |a_2+b_2|^p\Big)^{1/p} && \rightsquigarrow \text{Since } a_1, a_2, b_1, b_2 \text{ are all positive} \\
&\le \big(|a_1|^p + |a_2|^p\big)^{1/p} + \big(|b_1|^p + |b_2|^p\big)^{1/p} && \rightsquigarrow \text{By Minkowski's inequality} \\
&= \big(a_1^p + a_2^p\big)^{1/p} + \big(b_1^p + b_2^p\big)^{1/p} && \rightsquigarrow \text{Since } a_1, a_2, b_1, b_2 \text{ are all positive}
\end{aligned}
\tag{E17}
$$

Note that a_1, a_2, b_1, b_2 are all positive and

$$\begin{aligned} a_1 &= \sum_{\substack{i:x_i+y_i>0\\ x_i>0}} x_i \leq \sum_{i:x_i>0} x_i, & a_2 &= \sum_{\substack{i:x_i+y_i<0\\ x_i<0}} (-x_i) \leq \sum_{i:x_i<0} (-x_i) \\ b_1 &= \sum_{\substack{i:x_i+y_i>0\\ y_i>0}} y_i \leq \sum_{i:y_i>0} y_i, & b_2 &= \sum_{\substack{i:x_i+y_i<0\\ y_i<0}} (-y_i) \leq \sum_{i:y_i<0} (-y_i) \end{aligned} \tag{E18}$$

Using (E18) in (E17), we have

$$\begin{aligned} &\|x+y\| \\ \leq &\left(a_1^p + a_2^p\right)^{1/p} + \left(b_1^p + b_2^p\right)^{1/p} \\ \leq &\left[\left(\sum_{i:x_i>0} x_i\right)^p + \left(\sum_{i:x_i<0} (-x_i)\right)^p\right]^{1/p} + \left[\left(\sum_{i:y_i>0} y_i\right)^p + \left(\sum_{i:y_i<0} (-y_i)\right)^p\right]^{1/p} \\ &\qquad \rightsquigarrow \text{We used (E18) in the step above} \\ = &\|x\| + \|y\| \end{aligned} \tag{E19}$$

Therefore, the triangle inequality for $\|\cdot\|$ is verified.

Now we have proved that $\|\cdot\|$ is indeed a norm.

◎ PROOF PART 2.

Now we show that d is a metric. We verify the four conditions in Definition 1.

(1) By (E12), $d(x,y) = \|x-y\| \geq 0$, since $\|\cdot\|$ is a norm.

(2) We show $d(x,y) = 0 \Leftrightarrow x = y$ from two directions.

- $(\Rightarrow)$ If $d(x,y) = 0$, then by (E12), $\|x-y\| = 0$. By condition (2) in the norm definition, $x - y = 0$.
- $(\Leftarrow)$ If $x = y$, then $x - y = 0 = (0, \cdots, 0) \in \mathbb{R}^k$, $\|x-y\| = 0$. [3]

(3) Using condition (3) in Definition 2,

$$d(x,y) = \|x-y\| = \|(-1)(y-x)\| = |-1| \cdot \|y-x\| = \|y-x\| = d(y,x)$$

(4) Again using the form in (E12),

$$\begin{aligned} d(x,z) = &\|x-z\| \\ = &\|(x-y)+(y-z)\| \\ \leq &\|x-y\| + \|y-z\| \quad \rightsquigarrow \text{By the triangle inequality of the norm} \\ = &d(x,y) + d(y,z) \quad \rightsquigarrow \text{By (E12)} \end{aligned}$$

Therefore, we have completed our proof that $d(x,y)$ is a metric for $p \geq 1$. □

[3] We are using "For any norm $\|\cdot\|$, if $x = 0 \in \Omega$, then $\|x\| = 0$". This is implied by condition (3) in Definition 2: If $\|0\| \neq 0$, then $\|0\| = \|2 \cdot 0\| = 2\|0\|$. This implies $(2-1)\|0\| = 0$, which lead to contradiction with the assumption $\|0\| \neq 0$.

(b) When $0 < p < 1$, there are two cases depending on the dimension.

◎ CASE 1. When dimension $k = 1$.

In this case, d is a metric on $\mathbb{R}^1$.

By definition,

$$d(x,y) = \begin{cases} \Big((x-y)^p\Big)^{1/p} = x-y = |x-y|, & \text{if } x > y, \\ \Big((y-x)^p\Big)^{1/p} = y-x = |x-y|, & \text{if } x < y, \\ 0 = |x-y|, & \text{if } x = y, \end{cases}$$
$$= |x-y|$$

for all $x, y \in \mathbb{R}$.

Clearly, $d(x,y) = |x-y|$ is a metric, since by the properties of the absolute value operation on $\mathbb{R}$, for all $x,y,z \in \mathbb{R}$,

(1) $d(x,y) = |x-y| \geq 0$,

(2) $d(x,y) = |x-y| = 0$ if and only if $x = y$,

(3) $d(x,y) = |x-y| = |y-x| = d(y,x)$,

(4) $d(x,z) = |x-z| = |(x-y)+(y-z)| \leq |x-y| + |y-z| = d(x,y) + d(y,z)$, by the triangle inequality on the real line.

◎ CASE 2. When dimension $k \geq 2$.

In this case, d is not a metric on $\mathbb{R}^k$ for any $k \geq 2$.

Counterexample:

For any $k \geq 2$ and $0 < p < 1$, let

$$\begin{aligned} x =& (1,0,0,\cdots,0) \in \mathbb{R}^k, \\ y =& \mathbf{0} = (0,0,\cdots,0) \in \mathbb{R}^k, \\ z =& (0,1,0,0\cdots,0) \in \mathbb{R}^k \end{aligned}$$

Then

$$\begin{aligned} d(x,y) &= (1^p)^{1/p} = 1 \\ d(y,z) &= (1^p)^{1/p} = 1 \\ d(x,z) &= (1^p + 1^p)^{1/p} = 2^{1/p} \end{aligned}$$

Note that since $0 < p < 1$, $\frac{1}{p} > 1$. Therefore $2^{1/p} > 2$, and

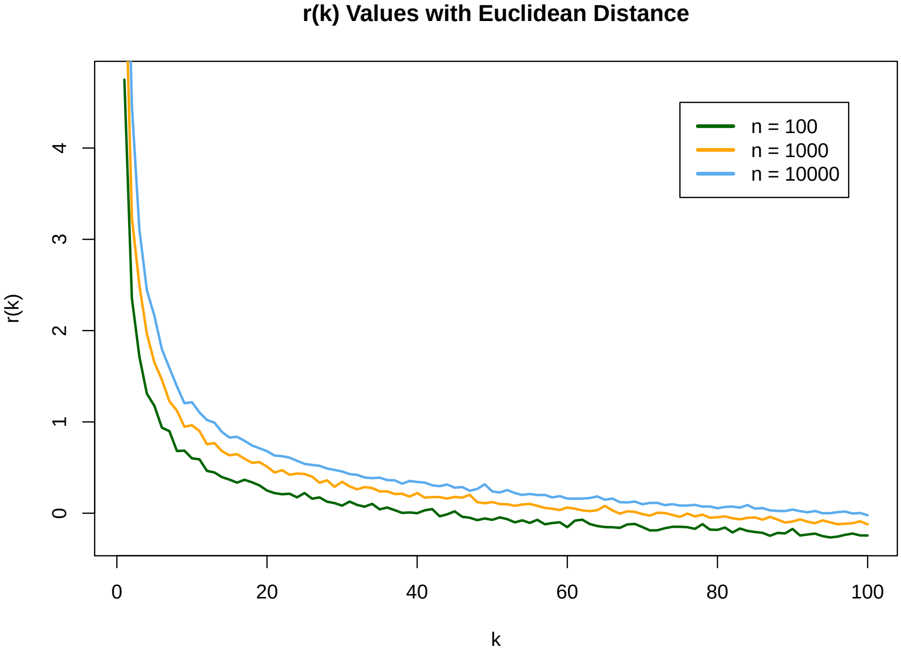
$$d(x,z) = 2^{1/p} > 2 = d(x,y) + d(y,z)$$

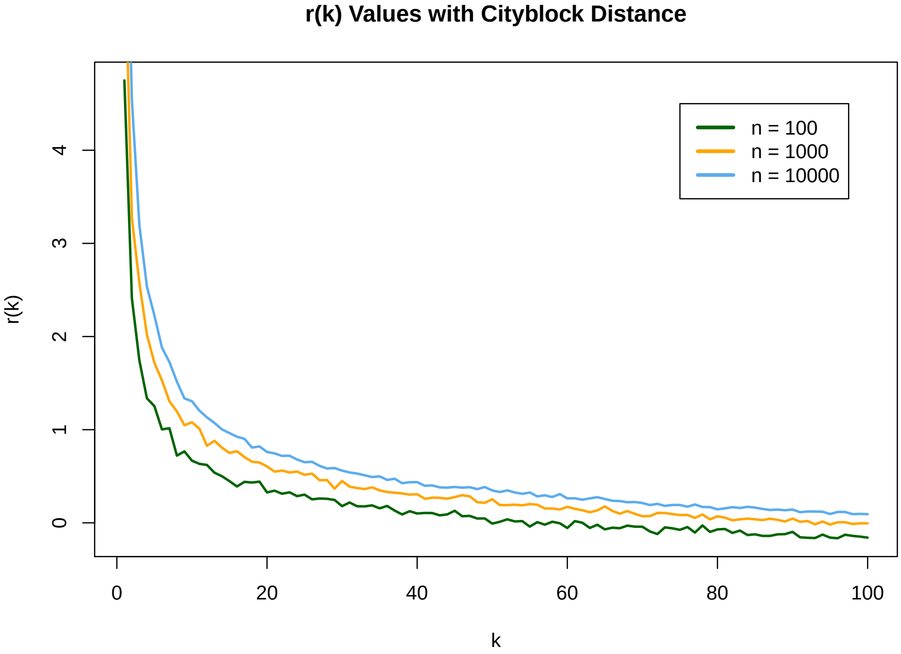
which implies that d does not satisfy the triangle inequality and hence is not a metric for any $k \geq 2$ and $0 < p < 1$.

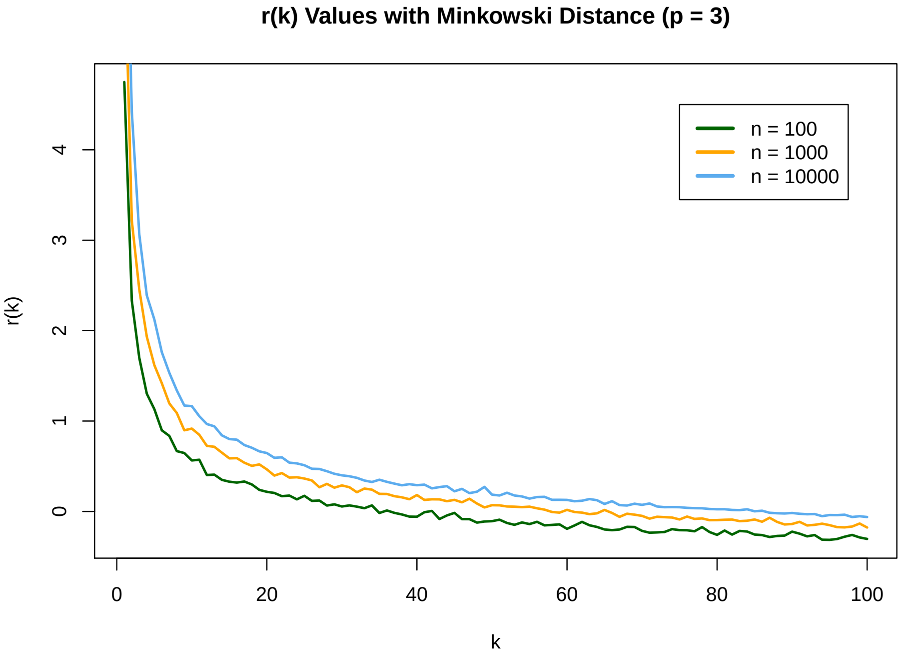
Problem 5

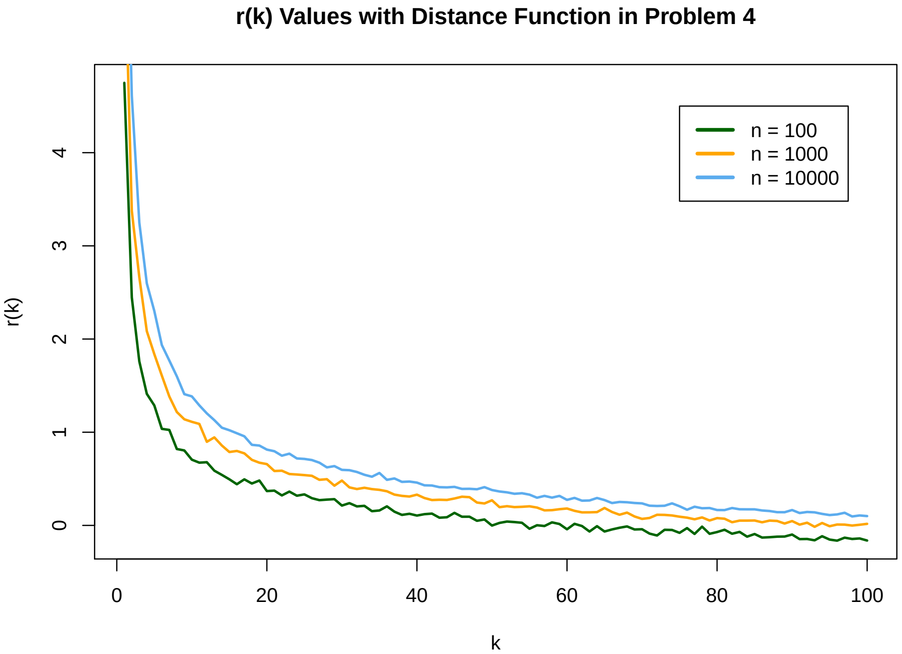
(a) We calculated all $r(k)$ values for $k = 1, \cdots, 100$, $n = \{100, 1000, 10000\}$, and with 5 different distance functions. In the following plots, all the r values are computed with a unit uniform random generator.

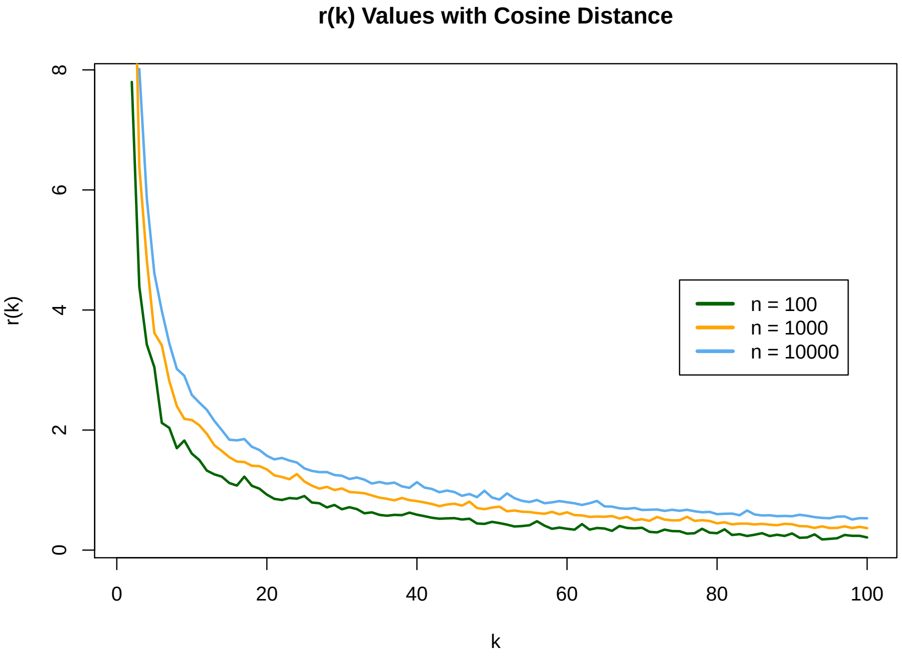
Plots of $r(k)$ with Uniform Random Generator





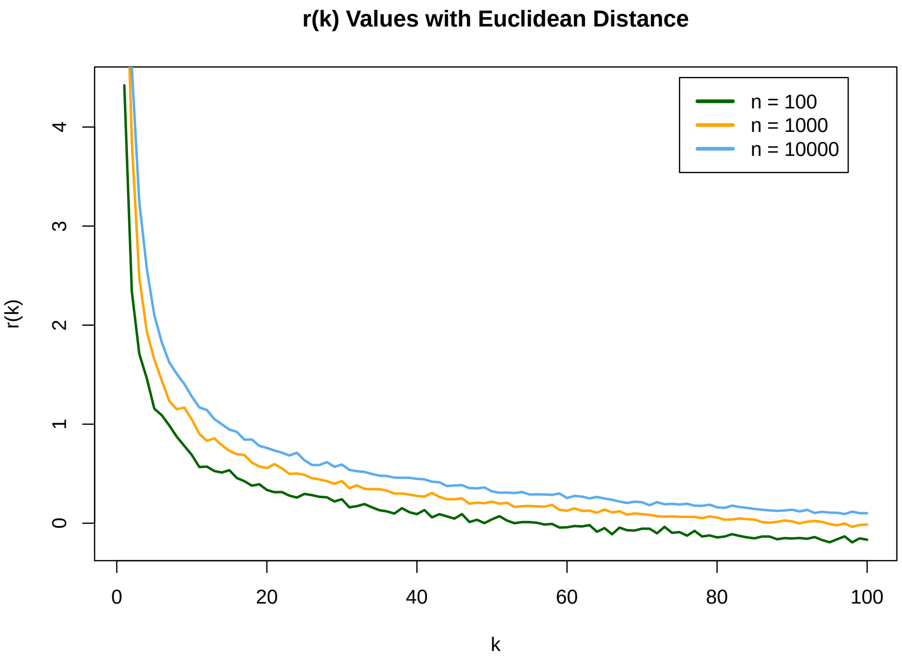

r(k) Values with Minkowski Distance (p = 3)
n = 100
n = 1000
n = 10000
r(k)
k
0
1
2
3
4
0
20
40
60
80
100

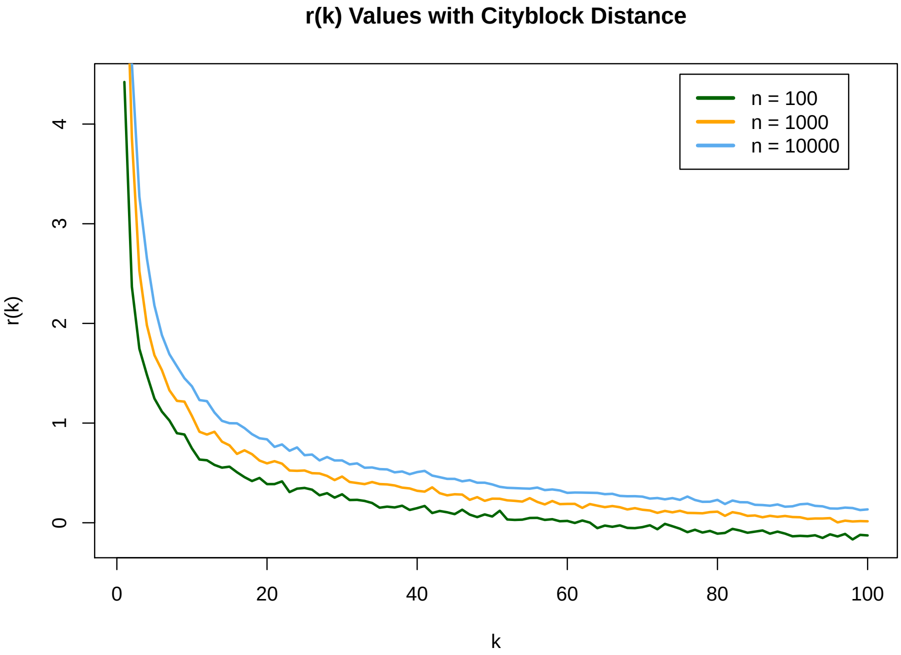

r(k) Values with Distance Function in Problem 4
n = 100
n = 1000
n = 10000
r(k)
k
0
1
2
3
4
0
20
40
60
80
100

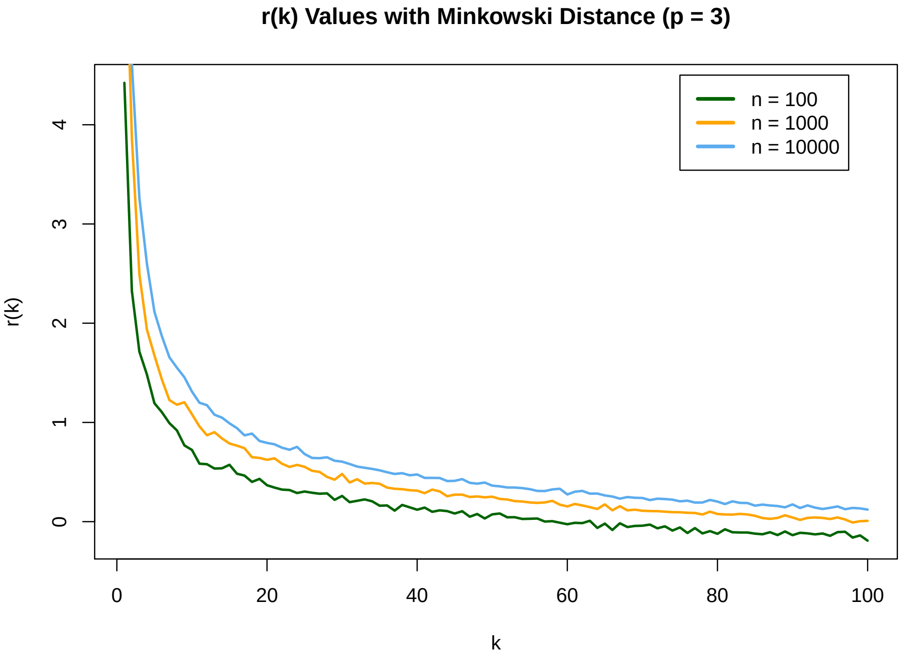


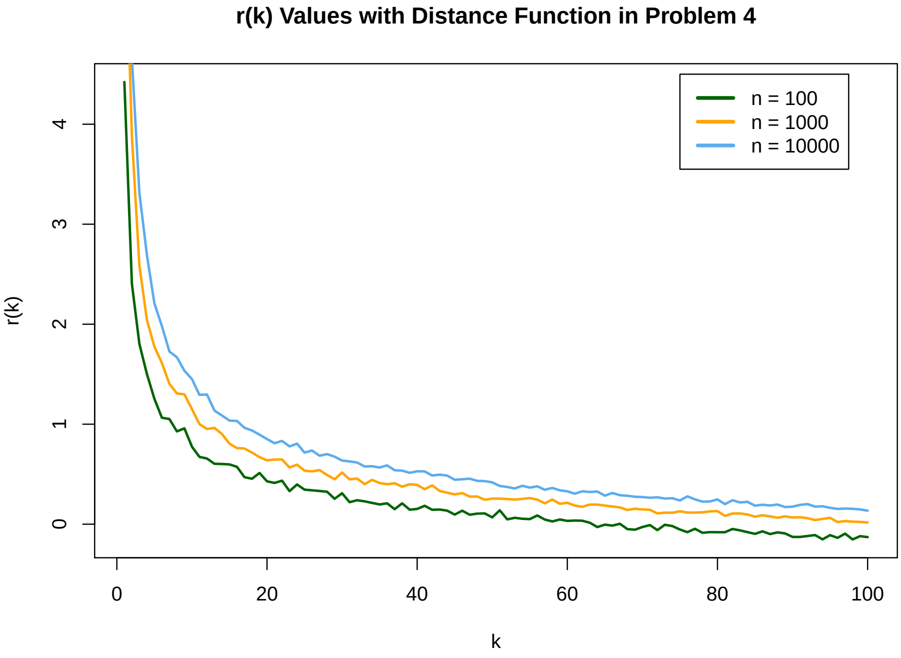
(b) We calculated all $r(k)$ values for $k = 1, \cdots, 100$, $n = \{100, 1000, 10000\}$, and with 5 different distance functions. In the following plots, all the r values are computed with a normal random generator with 0 mean and identity covariance matrix.

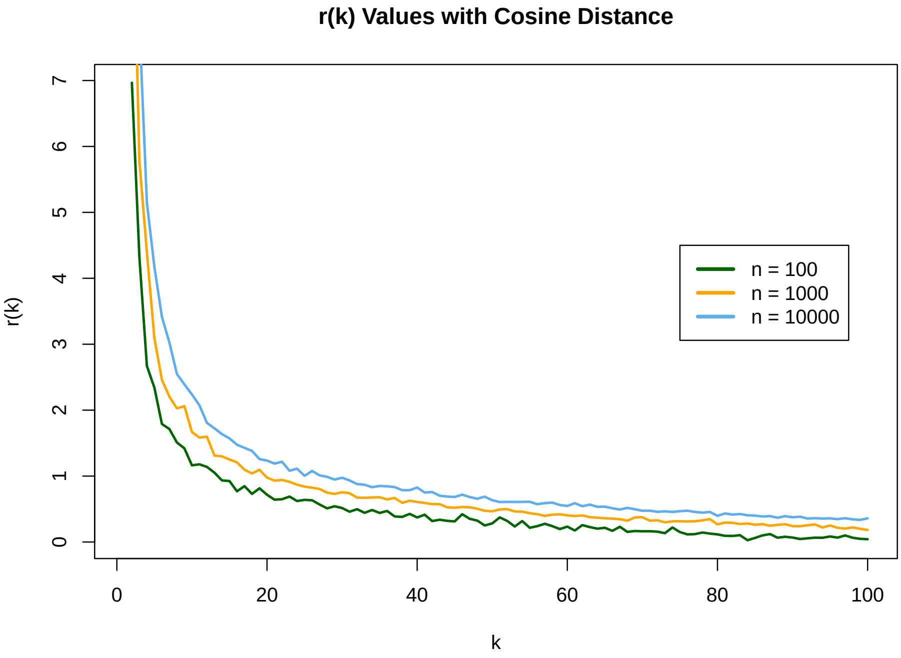
Plots of $r(k)$ with Normal Random Generator




r(k) Values with Cityblock Distance
n = 100
n = 1000
n = 10000
r(k)
k
0
1
2
3
4
0
20
40
60
80
100


r(k) Values with Minkowski Distance (p = 3)
n = 100
n = 1000
n = 10000
r(k)
k
0
1
2
3
4
0
20
40
60
80
100





(c) Some observations from the experiment:

- As n increases, $r(k)$ in general increases, as we can see from the relative position of the 3 curves. This is consistent with my expectation before the experiment. When the number of data points n goes large, the chance of getting a pair of very close points increases. That is, d_{min} is likely to be smaller. Similarly, when n is larger, the chance of getting a pair of very far away points also increases. Thus, d_{max} increases. Therefore, $r(k) = \log_2 \frac{d_{max}(k) - d_{min}(k)}{d_{min}(k)}$ increases as n goes larger.
- As k increases, the dimension of the space increases, which gives more space for variation. As a result,

points tend to be farther away from each other. Therefore, d_{min} increases and $r(k) = \log_2 \frac{d_{max}(k) - d_{min}(k)}{d_{min}(k)}$ decreases. The plots are consistent with the expectations as well.

- Although all the plots are different, they are very similar in shape and scale. Therefore, for this problem, the five distance functions are not very different in their effects.

- As n increases and k increases, we can observe dramatic increase in computation time. When n and k are small, the computation is almost instantly completed, using less than 0.001 seconds. However, when $n = 10000$ and $k = 100$, it takes around 4 to 8 seconds, depending on different norms, which is significantly longer than small n's and k's[4]. This is consistent with the phenomenon of "curse of dimensionality".

Problem 6

(b) We ran the decision tree program on 3 different data sets from UCI machine learning repository: “Glass", “Ionosphere", and “Transfusion". Following are the results.

Fraction of Correctly Classified Data Point on Test Set

Impurity Measure: Gini Index

Experiment	**Glass**	**Ionosphere**	**Transfusion**
1	66.9811%	90.2857%	70.7774%
2	66.0377%	87.4285%	68.3646%
3	70.7547%	86.2857%	72.1179%
4	62.2641%	89.7142%	69.9731%
5	64.1509%	92.5714%	69.7050%
Average	66.0377%	89.2571%	70.1876%

[4]Run time tested on local machine with MacOS

Fraction of Correctly Classified Data Point on Test Set

Impurity Measure: Information Gain

Experiment	Glass	Ionosphere	Transfusion
1	67.9245%	89.1428%	71.8498%
2	65.0943%	88.5714%	69.1689%
3	57.5471%	89.7142%	70.5093%
4	65.0943%	88.5714%	68.9008%
5	75.4716%	92.5714%	71.0455%
Average	66.22636%	89.71424%	70.29486%

Discussion

- As we can see, the results from all 3 data sets showed that we learned information by using the decision tree. Since the Glass set has 7 classes, and Ionosphere and Transfusion sets both have 2 classes, random guessing would give correct rates of around 15%, 50%, and 50% respectively. Therefore our program performed significantly better results than random guess.

(c)

Comparison of Gini Index and Information Gain.

Comparing the two measures of impurity, we can see that they perform very similarly in quality on these 3 sets, with information gain being slightly better on the average correct rates on all the sets.

Overall Quality of the Results.

As we can see that the overall quality is very good. Note that the Glass set has 7 classes, and Ionosphere and Transfusion sets both have 2 classes. By random guess, the correct rate would be around 15%, 50%, and 50% respectively[5]. Our results on all 3 sets provided significantly better results than random guess. In particular, the program performed especially well on the Ionosphere set.

[5] Random experiments also included in the code.