THU-70250043-0, Pattern Recognition (Spring 2021) **Homework: 6**

k-NN and Metric

Lecturer: Changshui Zhang zcs@mail.tsinghua.edu.cn

Hong Zhao vzhao@tsinghua.edu.cn

Student: Jingxuan Yang yangjx20@mails.tsinghua.edu.cn

Property of Euclidean Distance

1. When the metric space is a finite-dimensional Euclidean space, please prove that the Voronoi cells induced by the single-nearest neighbor algorithm must always be convex. Does this property hold when the metric becomes Manhattan distance?

In mathematics, a Voronoi diagram is a partition of a plane into regions close to each of a given set of objects. In the simplest case, these objects are just finitely many points in the plane (called seeds, sites, or generators). For each seed there is a corresponding region consisting of all points of the plane closer to that seed than to any other. These regions are called Voronoi cells, as shown in Figure 1. Similarly, Voronoi cells of a discrete set in higher-order Euclidean space are known as generalized polyhedra.

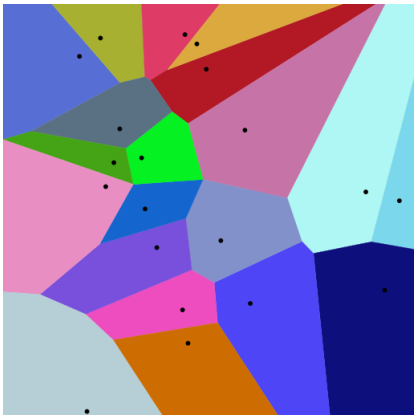

图 1: Voronoi cells with Euclidean distance [1].

Hint: Convex means for any two points x_1 and x_2 in a cell, all points on the segment linking x_1 and x_2 must also lie in the cell.

证明: 令 $\{x_1, x_2, \ldots, x_n\}$ 为 n 个样本点, V_i, $i = 1, 2, \ldots, n$ 表示 x_i 所在的由 ℓ_2 范数 (Euclidean Distance) 最近邻法生成的 Voronoi 区域. 设 ξ_1, $\xi_2 \in V_i$, 即有

$$\begin{aligned} \|\xi_1 - x_i\| &\leqslant \|\xi_1 - x_j\|, \ \forall\, j \neq i \\ \|\xi_2 - x_i\| &\leqslant \|\xi_2 - x_j\|, \ \forall\, j \neq i \end{aligned} \tag{1}$$

对 $\xi = \lambda\xi_1 + (1-\lambda)\xi_2$, $0 \leqslant \lambda \leqslant 1$, $\forall\, j \neq i$, 由上式及余弦定理

$$\langle a, b\rangle = a^\top b = \frac{\|a\|^2 + \|b\|^2 - \|a-b\|^2}{2} \tag{2}$$

可得

$$\begin{aligned} \|\xi - x_i\|^2 &= \|\lambda\xi_1 + (1-\lambda)\xi_2 - x_i\|^2 \\ &= \|\lambda(\xi_1 - x_i) + (1-\lambda)(\xi_2 - x_i)\|^2 \\ &= \lambda^2\|\xi_1 - x_i\|^2 + 2\lambda(1-\lambda)(\xi_1 - x_i)^\top(\xi_2 - x_i) + (1-\lambda)^2\|(\xi_2 - x_i)\|^2 \\ &= \lambda^2\|\xi_1 - x_i\|^2 + \lambda(1-\lambda)(\|\xi_1 - x_i\|^2 + \|\xi_2 - x_i\|^2 - \|\xi_1 - \xi_2\|^2) + (1-\lambda)^2\|(\xi_2 - x_i)\|^2 \\ &= \lambda\|\xi_1 - x_i\|^2 - \lambda(1-\lambda)\|\xi_1 - \xi_2\|^2 + (1-\lambda)\|(\xi_2 - x_i)\|^2 \\ &\leqslant \lambda\|\xi_1 - x_j\|^2 - \lambda(1-\lambda)\|\xi_1 - \xi_2\|^2 + (1-\lambda)\|\xi_2 - x_j\|^2 \\ &= \|\xi - x_j\|^2 \end{aligned} \tag{3}$$

即

$$\|\xi - x_i\| \leqslant \|\xi - x_j\|, \ \forall\, j \neq i \tag{4}$$

因此, $\xi \in V_i$, 即由 ℓ_2 范数最近邻法生成的 Voronoi 区域都是凸的.

由 ℓ_1 范数 (Manhattan Distance) 最近邻法生成的 Voronoi 区域不是凸的, 如图 2 所示.

图 2: Voronoi cells with Manhattan distance [2].

Properties of Metric

2. Please prove that the Minkowski metric indeed possesses the three properties required of all metrics.

Hint: A metric $D(\cdot,\cdot)$ must have three properties: for all vectors a, b and c,

(1) Identity of indiscernibles: $D(a,b)=0$ if and only if $a=b$.

(2) Symmetry: $D(a,b)=D(b,a)$.

(3) Triangle inequality: $D(a,b)+D(b,c)\geqslant D(a,c)$.

证明: s ($s\geqslant 1$) 阶 Minkowski 距离为

$$D(x,y)=\left(\sum_{j=1}^{d}|x_j-y_j|^s\right)^{1/s} \tag{5}$$

首先证明 identity of indiscernibles, $\forall\ x,y$, 当 $x=y$ 时, 有

$$D(x,y)=D(x,x)=\left(\sum_{j=1}^{d}|x_j-x_j|^s\right)^{1/s}=0 \tag{6}$$

当 $D(x,y)=0$ 时, 有

$$\sum_{j=1}^{d}|x_j-y_j|^s=0 \tag{7}$$

由于每一项绝对值均非负, 则

$$x_j=y_j,\quad \forall\ j=1,2,\ldots,d \tag{8}$$

即 $x=y$.

其次证明对称性, $\forall\ x,y$, 有

$$D(x,y)=\left(\sum_{j=1}^{d}|x_j-y_j|^s\right)^{1/s}=\left(\sum_{j=1}^{d}|y_j-x_j|^s\right)^{1/s}=D(y,x) \tag{9}$$

下面证明三角不等式, $\forall\ x,y,z$, 由绝对值性质有

$$\begin{aligned} D(x,y)&=\left(\sum_{j=1}^{d}|x_j-y_j|^s\right)^{1/s}\\ &=\left(\sum_{j=1}^{d}|x_j-z_j+z_j-y_j|^s\right)^{1/s}\\ &\leqslant\left(\sum_{j=1}^{d}(|x_j-z_j|+|z_j-y_j|)^s\right)^{1/s} \end{aligned} \tag{10}$$

其中, 由 Hölder 不等式可得

$$\begin{aligned}\sum_{j=1}^{d}(|x_j-z_j|+|z_j-y_j|)^s &= \sum_{j=1}^{d}(|x_j-z_j|+|z_j-y_j|)\cdot(|x_j-z_j|+|z_j-y_j|)^{s-1}\\ &= \sum_{j=1}^{d}(|x_j-z_j|+|z_j-y_j|)\cdot(|x_j-z_j|+|z_j-y_j|)^{s/t},\quad \frac{1}{s}+\frac{1}{t}=1\\ &= \sum_{j=1}^{d}|x_j-z_j|\cdot(|x_j-z_j|+|z_j-y_j|)^{s/t}+\sum_{j=1}^{d}|z_j-y_j|\cdot(|x_j-z_j|+|z_j-y_j|)^{s/t}\\ &\leqslant \left[\left(\sum_{j=1}^{d}|x_j-z_j|^s\right)^{1/s}+\left(\sum_{j=1}^{d}|z_j-y_j|^s\right)^{1/s}\right]\left(\sum_{j=1}^{d}(|x_j-z_j|+|z_j-y_j|)^s\right)^{1/t}\end{aligned} \tag{11}$$

整理即得

$$\left(\sum_{j=1}^{d}(|x_j-z_j|+|z_j-y_j|)^s\right)^{1-1/t}\leqslant\left(\sum_{j=1}^{d}|x_j-z_j|^s\right)^{1/s}+\left(\sum_{j=1}^{d}|z_j-y_j|^s\right)^{1/s} \tag{12}$$

又 $1-\frac{1}{t}=\frac{1}{s}$, 则

$$\left(\sum_{j=1}^{d}(|x_j-z_j|+|z_j-y_j|)^s\right)^{1/s}\leqslant\left(\sum_{j=1}^{d}|x_j-z_j|^s\right)^{1/s}+\left(\sum_{j=1}^{d}|z_j-y_j|^s\right)^{1/s} \tag{13}$$

此即 Minkowski 不等式, 所以

$$\begin{aligned}D(x,y) &\leqslant \left(\sum_{j=1}^{d}(|x_j-z_j|+|z_j-y_j|)^s\right)^{1/s}\\ &\leqslant \left(\sum_{j=1}^{d}|x_j-z_j|^s\right)^{1/s}+\left(\sum_{j=1}^{d}|z_j-y_j|^s\right)^{1/s}\\ &= D(x,z)+D(z,y)\end{aligned} \tag{14}$$

因此, 三角不等式成立. □

本题用到的 Hölder 不等式可由 Jensen 不等式证明如下 [3]. 简便起见, 对向量 $\boldsymbol{a}=(a_1,a_2,\ldots,a_n)\in\mathbb{R}^n$ 和 $p\geqslant 1$, 定义 ℓ_p 范数为

$$\|\boldsymbol{a}\|_p=\left(\sum_{k=1}^{n}|a_k|^p\right)^{1/p}, \tag{15}$$

则 Hölder 不等式可表述为对 $p\geqslant 1$ 和 $\forall\ \boldsymbol{a},\boldsymbol{b}\in\mathbb{R}_+^n$, 有

$$\sum_{k=1}^{n}a_kb_k\leqslant\|\boldsymbol{a}\|_p\,\|\boldsymbol{b}\|_q,\quad \frac{1}{p}+\frac{1}{q}=1. \tag{16}$$

证明: 对凸函数 $f(x)=x^p$ 应用 Jensen 不等式有

$$\left(\sum_{k=1}^{n}\lambda_k x_k\right)^p \leqslant \sum_{k=1}^{n}\lambda_k x_k^p, \quad \sum_{k=1}^{n}\lambda_k=1,\ \lambda_k\geqslant 0,\ k=1,2,\ldots,n \tag{17}$$

取 $a_k=\lambda_k^{1/p}x_k$, $c_k=\lambda_k^{1/q}$, 则上式变为

$$\sum_{k=1}^{n}a_k c_k \leqslant \left(\sum_{k=1}^{n}a_k^p\right)^{1/p}, \quad \sum_{k=1}^{n}c_k^q=1,\ c_k\geqslant 0,\ k=1,2,\ldots,n \tag{18}$$

取

$$c_k=\frac{b_k}{\|\boldsymbol{b}\|_q} \tag{19}$$

满足

$$\sum_{k=1}^{n}c_k^q=1,\ c_k\geqslant 0,\ k=1,2,\ldots,n \tag{20}$$

则有

$$\sum_{k=1}^{n}a_k\frac{b_k}{\|\boldsymbol{b}\|_q}\leqslant\left(\sum_{k=1}^{n}a_k^p\right)^{1/p}=\|\boldsymbol{a}\|_p \tag{21}$$

即

$$\sum_{k=1}^{n}a_k b_k\leqslant\|\boldsymbol{a}\|_p\ \|\boldsymbol{b}\|_q \tag{22}$$

因此, Hölder 不等式成立. □

k-NN Classifier

3. Let $\mathcal{D}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_n\}$ be a set of n independent labelled samples and let $\mathcal{D}_k(\boldsymbol{x})=\{\boldsymbol{x}'_1,\boldsymbol{x}'_2,\ldots,\boldsymbol{x}'_k\}$ be the k nearest neighbors of $\boldsymbol{x}$. Recall that the k-nearest-neighbor rule for classifying $\boldsymbol{x}$ is to give $\boldsymbol{x}$ the label most frequently represented in $\mathcal{D}_k(\boldsymbol{x})$. Consider a two-category problem with $P(\omega_1)=P(\omega_2)=1/2$. Assume further that the conditional densities $p(x|\omega_i)$ are uniform within unit hyperspheres, and the two categories center on two points ten units apart. Figure 3 shows a diagram of this situation.

3.1. Show that if k is odd, the average probability of error is given by

$$P_n(e)=\frac{1}{2^n}\sum_{j=0}^{(k-1)/2}\binom{n}{j}.$$

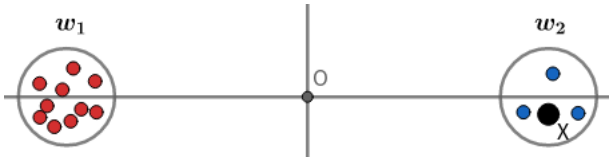

图 3: A diagram of assumed situation. When $k \geqslant 7$, X is misclassified as there are only 3 samples in ω_2.

解: 令 $P(x \in \omega_i, \omega_j)$, $i, j = 1, 2$ 表示 x 属于 ω_j 类, 但是 k-NN 算法将其判断为 ω_i 类. 当 ω_j 类的样本数量小于 $(k-1)/2$ 时, 就会出现分类错误的情况, 即此时 $i \neq j$.

因此, 错误率为

$$\begin{aligned} P_n(e) &= P(x \in \omega_1, \omega_2) + P(x \in \omega_2, \omega_1) \\ &= P(x \in \omega_1 | \omega_2) P(\omega_2) + P(x \in \omega_2 | \omega_1) P(\omega_1) \\ &= \frac{1}{2} \sum_{j=0}^{(k-1)/2} \binom{n}{j} \left(\frac{1}{2}\right)^j \left(\frac{1}{2}\right)^{n-j} + \frac{1}{2} \sum_{j=0}^{(k-1)/2} \binom{n}{j} \left(\frac{1}{2}\right)^j \left(\frac{1}{2}\right)^{n-j} \\ &= \frac{1}{2^n} \sum_{j=0}^{(k-1)/2} \binom{n}{j} \end{aligned} \tag{23}$$

3.2. Show that for this case the single-nearest neighbor rule has a lower error rate than the k-nearest-neighbor error rate for $k > 1$.

解: 当 $k = 1$ 时, 由上题可知最近邻法错误率为

$$\tilde{P}_n(e) = \frac{1}{2^n} \tag{24}$$

当 $k > 1$ 时

$$\begin{aligned} P_n(e) &= \frac{1}{2^n} \sum_{j=0}^{(k-1)/2} \binom{n}{j} \\ &= \frac{1}{2^n} \sum_{j=1}^{(k-1)/2} \binom{n}{j} + \frac{1}{2^n} \\ &> \frac{1}{2^n} = \tilde{P}_n(e) \end{aligned} \tag{25}$$

即最近邻法错误率 $\tilde{P}_n(e) < P_n(e)$.

3.3. If k is odd and is allowed to increase with n but is restricted by $k < a\sqrt{n}$, where a is a positive constant, show that $P_n(e) \to 0$ as $n \to \infty$.

解: 当 $n \to \infty$ 时, 有

$$\begin{aligned}
\lim_{n\to\infty} P_n(e) &= \lim_{n\to\infty} \frac{1}{2^n} \sum_{j=0}^{(k-1)/2} \binom{n}{j} \\
&\leqslant \lim_{n\to\infty} \frac{1}{2^n} \sum_{j=0}^{\lfloor (a\sqrt{n}-1)/2 \rfloor} \binom{n}{j} \\
&= \lim_{n\to\infty} \frac{1}{2^n} \sum_{j=0}^{\lfloor (a\sqrt{n}-1)/2 \rfloor} n^j, \quad \binom{n}{j} = \frac{n!}{j!(n-j)!} \sim n^j \\
&= \lim_{n\to\infty} \frac{n^{(a\sqrt{n}-1)/2}}{2^n} \\
&= \lim_{n\to\infty} \frac{n^{a\sqrt{n}/2}}{2^n} \\
&= \lim_{n\to\infty} \frac{\exp(\frac{a}{2}\sqrt{n}\log n)}{\exp(n\log 2)} \\
&= \lim_{n\to\infty} \exp\left(\frac{a}{2}\sqrt{n}\log n - n\log 2\right) \\
&= \lim_{n\to\infty} \exp\left[-n\log 2 \cdot \left(1 - \frac{a\log n}{2\sqrt{n}\log 2}\right)\right] \\
&= \lim_{n\to\infty} \exp\left[-n\log 2 \cdot \left(1 - \frac{an^{-1}}{n^{-1/2}\log 2}\right)\right] \\
&= \lim_{n\to\infty} \exp(-n\log 2) \\
&= 0
\end{aligned} \tag{26}$$

又 $P_n(e) \geqslant 0$, 因此

$$\lim_{n\to\infty} P_n(e) = 0 \tag{27}$$

Programming: k-NN Classifier on MNIST

4. Please implement k-NN classifier and run on MNIST [4]. You need to follow the official train/test split of MNIST. Compare the performance with the following settings:

- Using 100, 300, 1000, 3000, 10000 training samples.
- Using different values of k.
- Using at least three different distance metrics.

In this assignment, you are *NOT* allowed to use any existing libraries or code snippets that provides k-NN algorithm.

解: 首先分析训练样本数量对算法性能的影响, 分别使用 $n = 100,\ 300,\ 1000,\ 3000,\ 10000$ 个训练样本作为训练集, 比较最近邻分类器 ($k = 1$) 的性能变化, 计算采用 Minkowski 距离 $p = 2$, 即欧氏距离. k-NN 算法的准确率和运行时间分别如图 4 和图 5 所示.

由图可知, 算法的正确率随着样本数量增加而增加, 在样本数量增加到一定值后, 正确率的增长速度变慢; 算法的运行时间随着训练样本数量的增加而近乎线性增加.

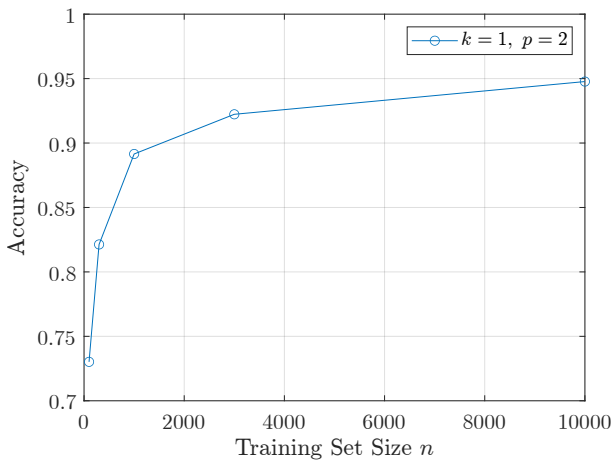

图 4: 正确率与训练样本数变化关系

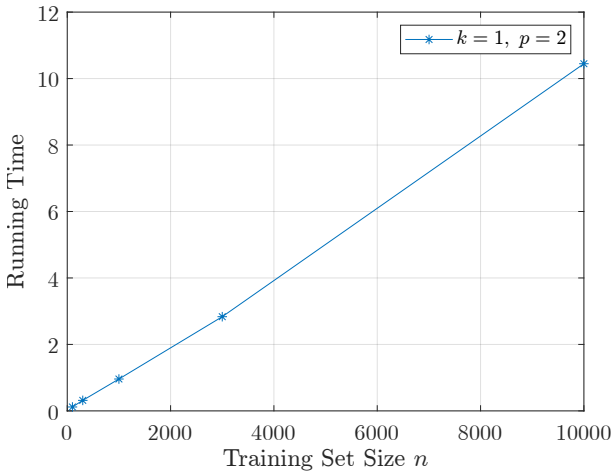

图 5: 运行时间与训练样本数变化关系

其次, 分析 k-NN 算法的参数 k 对算法性能的影响, 训练样本数取 $n = 3000$, Minkowski 距离参数取 $p = 2$, 即欧氏距离, 取 $k = [1, 100]$, 得到 k-NN 算法的准确率和运行时间分别如图 6 和图 7 所示.

由图可知, 算法的正确率随着 k 的增加先增加后减小, $k = 3$ 时正确率最高, 取过大的 k 反而会使得算法的正确率下降; 算法的运行时间基本稳定在 2.85 s, 与 k 的取值基本无关. 因为 k-NN 的时间复杂度主要体现在计算待预测样本点与训练集所有样本点之间的距离, 选择不同的 k 值对实际算法运行时间无明显影响.

最后分析不同的距离对算法性能的影响, 训练样本数取 $n = 3000$, 比较最近邻分类器 ($k = 1$) 的性能变化, Minkowski 距离参数取 $p = 1,\ 2,\ 4,\ \infty$, 得到 k-NN 算法的准确率和运行时间分别如图 8 和图 9 所示.

由图可知, 算法的正确率随着 p 的增大而有所增加，但 Chebyshev 距离 ($p = \infty$) 的正确率最低; 算法的运行时间与 ℓ_p 范数的计算复杂度有关, ℓ_1, ℓ_2 与 ℓ_∞ 范数的复杂度基本相同, 而 ℓ_4 范数的复杂度远远高于其他三种范数的复杂度.

Literature Reading

Please read a paper about metric learning [5].

Hint: You do not have to submit anything for this reading section.

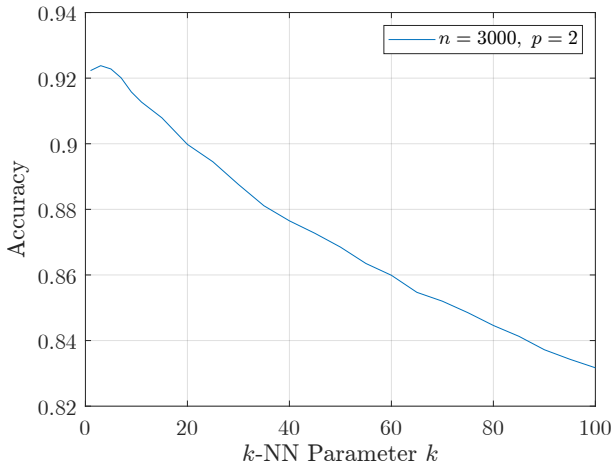

图 6: 正确率与 k-NN 参数 k 变化关系

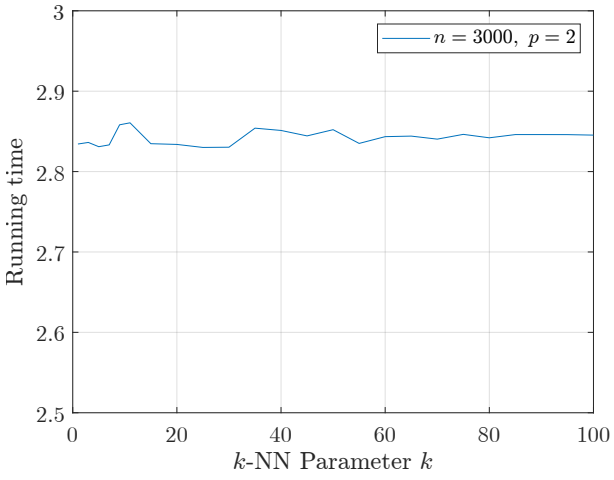

图 7: 运行时间与 k-NN 参数 k 变化关系

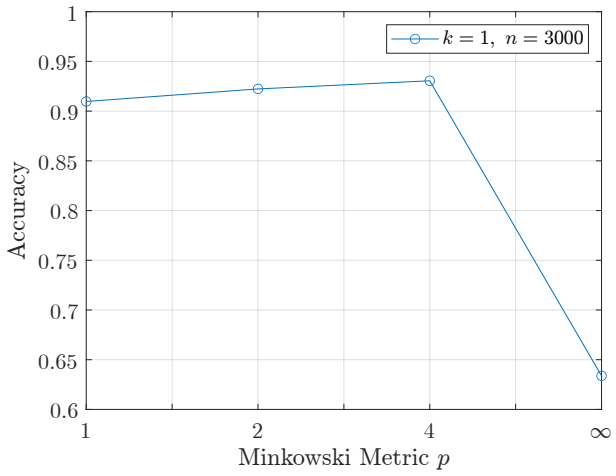

图 8: 正确率与 Minkowski 距离参数 p 变化关系

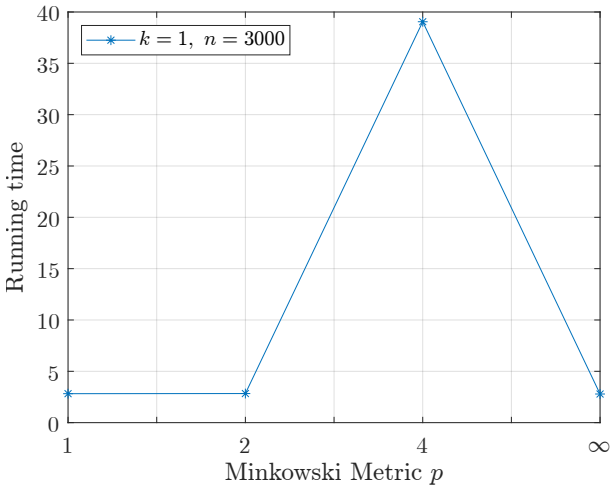

图 9: 运行时间与 Minkowski 距离参数 p 变化关系

参考文献

[1] File: Euclidean Voronoi diagram.svg. (2020, October 10). Wikimedia Commons, the free media repository. Retrieved 09:39, March 29, 2021 from https://commons.wikimedia.org/wiki/File:Euclidean_Voronoi_diagram.svg

[2] File: Manhattan Voronoi Diagram.svg. (2020, September 17). Wikimedia Commons, the free media repository. Retrieved 08:26, June 19, 2021 from https://commons.wikimedia.org/wiki/File:Manhattan_Voronoi_Diagram.svg

[3] J.M. Steele, The Cauchy-Schwarz Master Class, An Introduction to the Art of Mathematical Inequalities, MAA Problem Book Series, Cambridge University Press, 2008.

[4] LeCun Y. The MNIST database of handwritten digits[J]. http://yann.lecun.com/exdb/mnist/, 1998.

[5] Xing E P, Ng A Y, Jordan M I, et al. Distance metric learning with application to clustering with side-information[C] // NIPS. 2002, 15(505-512): 12.