Contents

CHAPTER 1 Data Mining and Analysis

1.7 EXERCISES

Q1. Show that the mean of the centered data matrix $\mathbf{Z}$ in Eq. (1.5) is $\mathbf{0}$.

Answer: Each centered point is given as: $\mathbf{z}_i = \mathbf{x}_i - \boldsymbol{\mu}$. Their mean is therefore:

$$\begin{aligned}\frac{1}{n}\sum_{i=0}^{n}\mathbf{z}_i &= \frac{1}{n}\sum_{i=0}^{n}(\mathbf{x}_i - \boldsymbol{\mu})\\ &= \frac{1}{n}\sum_{i=0}^{n}\mathbf{x}_i - \frac{1}{n}\cdot n\cdot\boldsymbol{\mu}\\ &= \boldsymbol{\mu} - \boldsymbol{\mu} = \mathbf{0}\end{aligned}$$

Q2. Prove that for the L_p-distance in Eq. (1.2), we have

$$\delta_\infty(\mathbf{x},\mathbf{y}) = \lim_{p\to\infty}\delta_p(\mathbf{x},\mathbf{y}) = \max_{i=1}^{d}\left\{|x_i - y_i|\right\}$$

for $\mathbf{x},\mathbf{y}\in\mathbb{R}^d$.

Answer: We have to show that

$$\lim_{p\to\infty}\left(\sum_{i=1}^{d}|x_i - y_i|^p\right)^{\frac{1}{p}} = \max_{i=1}^{d}\left\{|x_i - y_i|\right\}$$

Assume that dimension a is the max, and let $m = |x_a - y_a|$. For simplicity, we assume that $|x_i - y_i| < m$ for all $i \neq a$.

If we divide and multiply the left hand side with m^p we get:

$$\left(m^p\sum_{i=1}^{d}\left(\frac{|x_i - y_i|}{m}\right)^p\right)^{\frac{1}{p}} = m\left(1 + \sum_{i\neq a}\left(\frac{|x_i - y_i|}{m}\right)^p\right)^{\frac{1}{p}}$$

As $p \to \infty$, each term $\left(\frac{|x_i - y_i|}{m}\right)^p \to 0$, since $m > |x_i - y_i|$ for all $i \neq a$. The finite summation $\sum_{i \neq a} \left(\frac{|x_i - y_i|}{m}\right)^p$ converges to 0 as $p \to \infty$, as does $1/p$.

Thus $\delta_\infty(\mathbf{x}, \mathbf{y}) = m \times 1^0 = m = |x_a - y_a| = \max_{i=1}^{d}\{|x_i - y_i|\}$

Note that the same result is obtained even if we assume that dimensions other than a achieve the maximum value m. In the worst case, we have $m = |x_i - y_i|$ for all d dimensions. In this case, the expression on LHS becomes

$$\lim_{p\to\infty} m \left(\sum_{i=1}^{d} 1^p\right)^{1/p} = \lim_{p\to\infty} m d^{1/p} = \lim_{p\to\infty} m d^0 = m$$