

# Contents

<i>Contents</i>	1
<b>1 Data Mining and Analysis</b> . . . . .	3
1.7 Exercises	3
<b>PART I DATA ANALYSIS FOUNDATIONS</b>	5
<b>2 Numeric Attributes</b> . . . . .	7
2.7 Exercises	7
<b>3 Categorical Attributes</b> . . . . .	16
3.7 Exercises	16
<b>4 Graph Data</b> . . . . .	20
4.6 Exercises	20
<b>5 Kernel Methods</b> . . . . .	26
5.6 Exercises	26
<b>6 High-dimensional Data</b> . . . . .	29
6.9 Exercises	29
<b>7 Dimensionality Reduction</b> . . . . .	39
7.6 Exercises	39
<b>PART II FREQUENT PATTERN MINING</b>	45
<b>8 Itemset Mining</b> . . . . .	47
8.5 Exercises	47
<b>9 Summarizing Itemsets</b> . . . . .	56
9.6 Exercises	56
	<b>1</b>

<b>10</b>	<b>Sequence Mining</b> . . . . .	<b>63</b>
	10.5 Exercises	63
<b>11</b>	<b>Graph Pattern Mining</b> . . . . .	<b>75</b>
	11.5 Exercises	75
<b>12</b>	<b>Pattern and Rule Assessment</b> . . . . .	<b>84</b>
	12.4 Exercises	84
<b>PART III</b>	<b>CLUSTERING</b>	<b>89</b>
<b>13</b>	<b>Representative-based Clustering</b> . . . . .	<b>91</b>
	13.5 Exercises	91
<b>14</b>	<b>Hierarchical Clustering</b> . . . . .	<b>99</b>
	14.4 Exercises	99
<b>15</b>	<b>Density-based Clustering</b> . . . . .	<b>106</b>
	15.5 Exercises	106
<b>16</b>	<b>Spectral and Graph Clustering</b> . . . . .	<b>111</b>
	16.5 Exercises	111
<b>17</b>	<b>Clustering Validation</b> . . . . .	<b>118</b>
	17.5 Exercises	118
<b>PART IV</b>	<b>CLASSIFICATION</b>	<b>123</b>
<b>18</b>	<b>Probabilistic Classification</b> . . . . .	<b>125</b>
	18.5 Exercises	125
<b>19</b>	<b>Decision Tree Classifier</b> . . . . .	<b>129</b>
	19.4 Exercises	129
<b>20</b>	<b>Linear Discriminant Analysis</b> . . . . .	<b>137</b>
	20.4 Exercises	137
<b>21</b>	<b>Support Vector Machines</b> . . . . .	<b>141</b>
	21.7 Exercises	141
<b>22</b>	<b>Classification Assessment</b> . . . . .	<b>145</b>
	22.5 Exercises	145

## 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

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

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

**Answer:** We have to show that

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

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  $p \rightarrow \infty$ , each term  $\left(\frac{|x_i - y_i|}{m}\right)^p \rightarrow 0$ , 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  $p \rightarrow \infty$ , 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

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