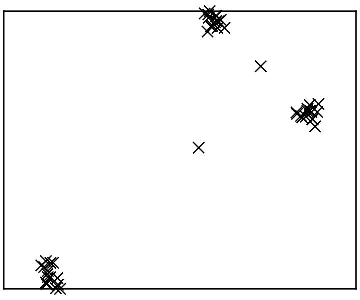
Data Mining & Knowledge Discovery Exam 28.02.23 Sample Solution

points	grade
27-	1.0
25-26.5	1.3
23-24.5	1.7
21-22.5	2.0
19-20.5	2.3
17-18.5	2.7
15-16.5	3.0
13-14.5	3.3
11-12.5	3.7
9-10.5	4.0

Problem 1: General Understanding (7 of 35 points)

Consider the data set shown in this scatter plot:



a) Is there noise in this data set?

Yes.

b) How many outliers are in this data set?

0

c) How many inliers are in this data set?

2

d) How many classes are in this data set?

None, the data set is unlabeled.

e) How many clusters are in this data set?

3

f) How many clusters are in the one–dimensional PCA mapping of this data set?

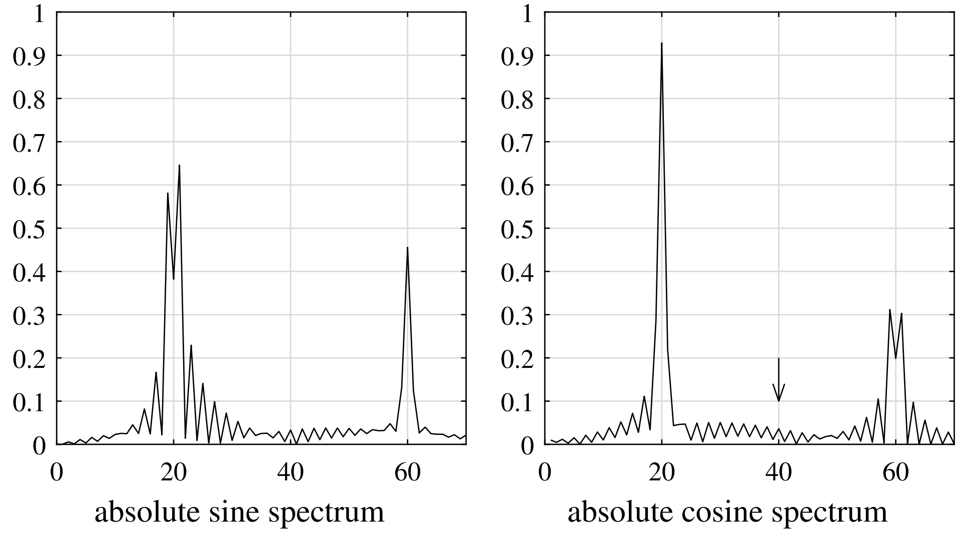
2

g) How many clusters are in the one–dimensional Sammon mapping of this data set?

3

Problem 2: Spectral analysis (8 of 35 points)

Consider these absolute sine and cosine spectra:



a) Find an equation for all possible time series that yield absolute sine and cosine spectra similar to these!

$$\hat{x}_{20} \approx \sqrt{0.38^2 + 0.93^2} \approx 1, \quad \varphi_{20} \approx \pm \operatorname{atan} \frac{\pm 0.38}{\pm 0.93} \approx \pm \frac{\pi}{8} \approx \pm 0.4,$$

$$\hat{x}_{60} \approx \sqrt{0.46^2 + 0.2^2} \approx \frac{1}{2}, \quad \varphi_{60} \approx \pm \operatorname{atan} \frac{\pm 0.46}{\pm 0.2} \approx \pm \frac{3}{8}\pi \approx \pm 1.2$$

$$x_t \approx \pm \cos\left(20\omega t \pm \frac{\pi}{8}\right) \pm \frac{1}{2} \cos\left(60\omega t \pm \frac{3}{8}\pi\right), \quad \omega > 0 \text{ (6)}$$

b) What is the reason for the peak marked with the arrow in the cosine spectrum?

Limited number of elements of the time series. (2)

Problem 3: Relations (10 of 35 points)

Consider a pair of two points $\{x, y\} \subset \mathbb{R}^+ \times \mathbb{R}^+$ with city block distance $\sqrt{3}/2$ and cosine similarity $\sqrt{3}/2$. Find two (different) pairs $\{x, y\}$ of such points.

For convenience we set $x_1 = y_1$ and $x_2 = 0$, so city block distance $\sqrt{3}/2$ implies $y_2 = \sqrt{3}/2$. Inserting these into the formula for cosine similarity yields

$$\frac{x_1 y_1 + x_2 y_2}{\sqrt{(x_1^2 + x_2^2)(y_1^2 + y_2^2)}} = \frac{x_1^2}{\sqrt{x_1^2 (x_1^2 + \frac{3}{4})}} = \frac{1}{\sqrt{(1 + \frac{3}{4x_1^2})}} = \frac{\sqrt{3}}{2}$$

$$\Rightarrow 1 + \frac{3}{4x_1^2} = \frac{4}{3} \Rightarrow 4x_1^2 = 9 \Rightarrow x_1 = \frac{3}{2}$$

so one pair of points is

$$\left\{ \left(\frac{3}{2}, 0 \right), \left(\frac{3}{2}, \frac{1}{2}\sqrt{3} \right) \right\}$$

The city block distance and the cosine similarity are symmetric, so swapping the two dimensions yields another pair of points:

$$\left\{ \left(0, \frac{3}{2} \right), \left(\frac{1}{2}\sqrt{3}, \frac{3}{2} \right) \right\}$$

Problem 4: Classification (10 of 35 points)

Consider a data set with more than 5 elements and randomly distributed class labels 1 and 2. Let p be the probability of class 1.

a) What is the probability of class 1 for the nearest-neighbor classifier?

$$p \quad (1)$$

b) What is the probability of class 1 for the 3-nearest-neighbor classifier?

$$3p^2(1-p) + p^3 = -2p^3 + 3p^2 \quad (2)$$

c) What is the probability of class 1 for the 5-nearest-neighbor classifier?

$$\binom{5}{3} p^3(1-p)^2 + 5p^4(1-p) + p^5 = 6p^5 - 15p^4 + 10p^3 \quad (3)$$

d) For $p = \frac{1}{4}$, $p = \frac{1}{2}$, and $p = \frac{3}{4}$, which of these three classifiers yields the highest probability of class 1? Explain!

nearest-neighbor classifier:

$$\begin{aligned}
p = \frac{1}{4}: &\quad \frac{128}{512} \\
p = \frac{1}{2}: &\quad \frac{1}{2} \\
p = \frac{3}{4}: &\quad \frac{384}{512}
\end{aligned}$$

3-nearest-neighbor classifier:

$$\begin{aligned}
p = \frac{1}{4}: &\quad -2\frac{1}{4^3} + 3\frac{1}{4^2} = \frac{5}{32} = \frac{80}{512} \\
p = \frac{1}{2}: &\quad -2\frac{1}{2^3} + 3\frac{1}{2^2} = \frac{1}{2} \\
p = \frac{3}{4}: &\quad -2\frac{3^3}{4^3} + 3\frac{3^2}{4^2} = \frac{27}{32} = \frac{432}{512}
\end{aligned}$$

5-nearest-neighbor classifier:

$$\begin{aligned}
p = \frac{1}{4}: &\quad 6\frac{1}{4^5} - 15\frac{1}{4^4} + 10\frac{1}{4^3} = \frac{53}{512} \\
p = \frac{1}{2}: &\quad 6\frac{1}{2^5} - 15\frac{1}{2^4} + 10\frac{1}{2^3} = \frac{1}{2} \\
p = \frac{3}{4}: &\quad 6\frac{3^5}{4^5} - 15\frac{3^4}{4^4} + 10\frac{3^3}{4^3} = \frac{459}{512}
\end{aligned}$$

For $p = \frac{1}{4}$, the nearest-neighbor classifier yields the highest probability. For $p = \frac{1}{2}$ all three yields the same probability. For $p = \frac{3}{4}$, the 5-nearest-neighbor classifier yields the highest probability. Explanation: For equal class probabilities $p = 1/2$, for symmetry reasons the probability of a majority across neighbors does not change with the number of neighbors. But the more neighbors are considered, the less likely is a majority for a class with probability $< 1/2$, and the more likely is a majority for a class with probability $> 1/2$. (4)