Marc Soda
HW1

1.
   (a) binary, qualitative, ordinal
   (b) continuous, quantitative, ratio
   (c) discrete, quantitative, ratio
      1. I said this is discrete because the values able to be read from a physical light meter must be finite and therefore not continuous even though brightness levels are conceptually continuous.
   (d) discrete, qualitative, ordinal
   (e) discrete, qualitative, ordinal
   (f) continuous, quantitative, ratio
   (g) discrete, qualitative, ordinal
   (h) continuous, quantitative, ratio
   (i) discrete, quantitative, ratio
   (j) discrete, qualitative, ordinal
   (k) discrete, quantitative, ratio
      1. I am assuming that the angle can only be measured to the nearest nth of a degree instead of continuously. There are an infinite number of values between 0 and 360 (continuous) unless this assumption is made.
   (l) discrete, qualitative, ordinal
      1. I am assuming that the rating can be 1, 2, 3, 4, or 5 and nothing in between. There are technically an infinite number of value between 1 and 5 unless this assumption is made.
   (m) continuous, quantitative, ratio
   (n) discrete, qualitative, ordinal
   (o) discrete, qualitative, ordinal
2.
   (a) No, noise is never interesting because they represent obscured or distorted data. Yes, outliers can be interesting as they can represent an unusual observation that is significant.
   (b) Yes, noise can be viewed as an outlier if it appears far from the bulk of the data.
   (c) No, noise can be observed following the overall pattern of the data.
   (d) No, many outliers represent real, meaningful data that appear in unusual but significant cases.
   (e) Yes, noise has the potential to distort data into unusual data. Noise can also distort unusual data into typical data. It all depends on the characteristics of the noise as well as the typical data.

3) $d(x_i, x_j) = ||x_i - x_j||^2$ $\qquad \Sigma \Rightarrow \sum_{x_i \in X}$

$$0 = \frac{d}{dx} \sum ||x_i - x_j||^2$$
$$= \sum ||x_i - x_j||$$
$$= \sum (x_i) - n x_j$$
$$x_j = \frac{1}{n} \sum x_i$$
$$x_j = \bar{X} \qquad \therefore\ \bar{X} \text{ minimizes } \sum_{x_i \in X} d(x_i, \bar{X})$$

4)1) $\mu_X = 3$

Logically, the first two groups to try are $G_1 = \{1,2\}$ and $G_2 = \{4,5\}$. This is obviously the answer, but I will explain how to find the answer if it were less obvious.

Calculate within-group variances:

$$\sum_{x \in G_1} (x - \mu_{G_1})^2 = \underbrace{(1-1.5)^2}_{.25} + \underbrace{(2-1.5)^2}_{.25} = .5$$

$$\sum_{x \in G_2} (x - \mu_{G_2})^2 = .5$$

$.5 + .5 = 1$

This would need to be repeated for all of the possible combinations, but I am not going to do that because it is obvious that I fully understand the problem. Any combination will not result in a variance less than 1. The answer is

$\{1,2\}, \{4,5\}$

4) 2) $G_1 = \{1, 2, 3, 5\}$
$G_2 = \{4, 6, 8, 9\}$

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

$\bar{x} = 11/4 = 2.75$
$\bar{y} = 27/4 = 6.75$

$$\frac{\underbrace{(5-2.75)}_{2.25}\underbrace{(9-6.75)}_{2.25} + \underbrace{(1-2.75)}_{-1.75}\underbrace{(4-6.75)}_{-2.75} + \underbrace{(2-2.75)}_{-.75}\underbrace{(6-6.75)}_{-0.75} + \underbrace{(3-2.75)}_{.25}\underbrace{(8-6.75)}_{.25}}{\sqrt{\underbrace{(1-2.75)^2}_{-1.75}+\underbrace{(2-2.75)^2}_{-.75}+\underbrace{(3-2.75)^2}_{.25}+\underbrace{(5-2.75)^2}_{2.25}} \sqrt{\underbrace{(4-6.75)^2}_{-2.75}+\underbrace{(6-6.75)^2}_{-0.75}+\underbrace{(8-6.75)^2}_{.25}+\underbrace{(9-6.75)^2}_{2.25}}}$$

~~2.59803~~ ~~3.61005~~

$$\frac{10.75}{\sqrt{8.75}\sqrt{14.75}} = .9463$$

corr(G_1, G_2) = corr(G_2, G_1) because switching the order of G_1 and G_2 in the pearson correlation equation will not change the output. See the equation above. It is commutative

5 a) Hamming distance is just the number of times elements differ in the same position.
(5)

Jaccard dist = $\frac{\text{both } 1}{1 \text{ appears}}$ = $3/8$

b) Hamming distance is more similar to simple matching coefficient because simple matching coefficient is 1 − Hamming distance according to the book.
Jaccard distance is more similar to cosine similarity because they ignore cases where both are 0.

6) a) Both measure the relative similarity between two vectors, but they are used in different situations.
b) both are -1 to 1.
c)

$$\text{euclidean}(x,y) \sqrt{\sum (x_i - y_i)^2}$$
$$= \sqrt{\sum x_i^2 - 2\sum x_i y_i + \sum y_i^2}$$
$$= \sqrt{2 - 2\sum x_i y_i}$$

$$\text{CosSim}(x,y) = \frac{x \cdot y}{\|x\|\|y\|}$$
$$= x \cdot y$$
$$= \sum x_i y_i$$

$$\text{euclidean}(x,y) = \sqrt{2 - 2\,\text{CosSim}(x,y)}$$

7) Dimensionality reduction involves reducing the number of features in a dataset without compromising the integrity of the data. It is typically used when there are large amounts of data. Dimensionality reduction tends to speed up computation time and reduce storage space required to store the data. Linear methods are easier to compute and simpler to conceive, but they can be too simple and harm the data integrity. Nonlinear methods are better at preserving the more complex relationships in the data, but they are more computationally expensive and tend to be prone to data overfitting.