(8) $x = (1,1,0,0)$ $y = (1,1,0,0)$

i) cosine:

$$\cos(x,y) = \frac{x \cdot y}{\|x\| \cdot \|y\|}$$

$$x \cdot y = (1,1,0,0) \cdot (1,1,0,0)$$
$$= 1(1) + 1(1) + 0(0) + 0(0) = 1+1 = 2$$

$$\|x\| = \|(1,1,0,0)\| = \sqrt{1^2+1^2+0^2+0^2} = \sqrt{2}$$

$$\|y\| = \sqrt{1^2+1^2+0^2+0^2} = \sqrt{2}$$

$$\cos(x,y) = \frac{2}{\sqrt{2}\cdot\sqrt{2}} = \frac{2}{2} = 1$$

ii) correlation:

$$corr(x,y) = \frac{cov(x,y)}{\sigma(x)\cdot\sigma(y)}$$

$$mean(x) = (1+1+0+0)/4 = 1/2$$

$$mean(y) = (1+1+0+0)/4 = 1/2$$

$$cov(x,y) = \frac{1}{N-1}\sum(x - mean(x))(y - mean(y))$$
$$= \frac{1}{4-1}\left[\left(1-\frac{1}{2}\right)\left(1-\frac{1}{2}\right) + \left(1-\frac{1}{2}\right)\left(1-\frac{1}{2}\right) + \left(0-\frac{1}{2}\right)\left(0-\frac{1}{2}\right) + \left(0-\frac{1}{2}\right)\left(0-\frac{1}{2}\right)\right]$$
$$= \frac{1}{3}\left[\frac{1}{4}+\frac{1}{4}+\frac{1}{4}+\frac{1}{4}\right] = \frac{1}{3}\left(\frac{4}{4}\right) = \frac{1}{3}$$

$$\sigma(x) = \sqrt{\frac{\sum(x_i - \bar{x})^2}{N}}$$
$$= \sqrt{\frac{\left(1-\frac{1}{2}\right)^2 + \left(1-\frac{1}{2}\right)^2 + \left(0-\frac{1}{2}\right)^2 + \left(0-\frac{1}{2}\right)^2}{4}}$$
$$= \sqrt{\frac{\frac{1}{4}+\frac{1}{4}+\frac{1}{4}+\frac{1}{4}}{4}} = \sqrt{\frac{4/4}{4}} = \frac{1}{2}$$

$$\sigma(y) = \sqrt{\frac{\Sigma(y_i - \bar{y})^2}{N}} = \frac{1}{2}$$

$$\text{covr}(x,y) = \frac{\frac{1}{3}}{\frac{1}{2} \cdot \frac{1}{2}} = \frac{1/3}{1/4} = \frac{4}{3}$$

iii) Euclidean

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

$$= \sqrt{(1-1)^2 + (1-1)^2 + (0-0)^2 + (0-0)^2} = 0$$

iv) Jaccard

$$\text{Jaccard} = \frac{\text{number of matching presences}}{\text{number of attributes not involved in 00 matches}}$$

$$= \frac{f_{11}}{f_{01} + f_{10} + f_{11}} = \frac{2}{0+0+2} = \frac{2}{2} = 1$$

9

(a) One of the problem is if there are enough duplicates then the nearest ~~neig~~ neighbor list might be just the duplicates. Another problem is that the order of duplicate objects will depend on the details of algorithm & order of objects in the data set for the nearest neighbor.

(b) One of the ways is to have just one object for each group of duplicate objects so each neighbour can represent either a single object or a group of duplicate objects.

10

(a) Binary, qualitative, ordinal
(b) Continuous, quantitative, ratio
(c) Discrete, qualitative, ordinal
(d) Continuous, quantitative, ratio
(e) Discrete, qualitative, ordinary
(f) Continuous, quantitative, interval/ratio (need to consider if sea level is arbitrary origin)
g) Discrete, quantitative, ratio
h) Discrete, qualitative, nominal
i) Discrete, qualitative, ordinal
j) Discrete, qualitative, ordinary.
k) Continuous, quantitative, interval/ratio.
l) Discrete, quantitative, ratio
m) Discrete, qualitative, nominal.