CS 5525
Solutions to Homework Assignment 1
Meghendra Singh

September 13, 2016

[10] 1.

The following table classifies the attributes and presents the reasoning behind the classification:

Attribute	**Type**	**Class**	**Reasoning**
(a) Angles as measured in degrees between 0 and 360	Continuous or Discrete	Ratio	If we consider decimal degrees (For E.g.: 30.1275°) there can be infinitely many degree measures between any two degree measure, making the attribute continuous. If we do not consider the subdivisions or consider the *Sexagesimal* units ($1^\circ = 60$ minutes and 1 minute = 60 seconds) then the attribute can be considered discrete as there are only a finite number of measures between any two Sexagesimal angle measures. The Class is ratio because any two degree measures are distinct, have an order, can be added, subtracted, multiplied and divided meaningfully (two angles can be multiplied resulting in a meaningful *solid angle*).
(b) Bronze, Silver, and Gold medals as awarded at the Olympics	Discrete	Ordinal	There are only three values possible for these medals (gold, silver and bronze), hence they are discrete. Also two medals are distinct and have an order, but the properties of addition and multiplication don't hold for them. Hence they belong to the Ordinal class.

(c) Cell Phone Numbers	Discrete	Ordinal	There can only be a finite number of cell phone numbers, hence these are discrete. Also, any two cell phone numbers are distinct, can be considered to have an order (For E.g.: 54012345 ¡ 54012346). But adding or multiplying two cell phone numbers can result in an invalid cell phone number. (For E.g. adding the two valid 10 digit cell phone numbers: 9999999999 and 910000200 will result in 10910000199, which has 11 digits and hence is invalid). Therefore the attribute is ordinal.
(d) Different departments in the University	Discrete	Nominal	There is no meaningful scale on which different departments at a university can be ordered and compared. Hence they are discrete and nominal.

[10] 2.

The following table gives the relevant transformations for the vector A= [1204, -212, 30.21, 12.0, 56]:

Transformation method	**Result**
Z-score normalization	[1.75600, -0.76591, -0.33453, -0.36696, -0.28860] $\mu = 218.04$, $\sigma = 561.48$
Decimal scaling	[0.1204000, -0.0212000, 0.0030210, 0.0012000, 0.0056000] $j = 4$; Since, Max($\|A\|$) = 1204)
Min-Max normalization [0, 1]	[1.00000, 0.00000, 0.17105, 0.15819, 0.18927]; $\mathrm{min_A}$ = -212, $\mathrm{max_A}$ = 1204

[10] 3.

Upon concatenating Vector A from **Problem 2** to the vector C= [30 , 50, 10, 30, 40, -25, 95], we get the following vector: AC = [1204, -212, 30.21, 12.0, 56, 30 , 50, 10, 30, 40,

-25, 95]. The following table presents the results of discretizing the vector **AC** into 3 bins using Equal Width Discretization and Equal Frequency Discretization:

Dscretization method	**Result**
(1) Equal Width Discretization	Here we need to divide the data into k=3 intervals of equal size. Hence, the width of each interval can be given by: w = ((Max(AC) - Min(AC)) / 3); Therefore, w = (1204 - (-212))/3; w = 472; The two bin intervals boundaires are: (-212 + 472) = 260 and (-212 + 2*472) = 732; and the three bins are $-\infty$ to 260, 260 to 732 and 732 to ∞. This gives the following three vectors representing the values in the three bins: (1) [-212, -25, 10, 12, 30, 30, 30.21, 40, 50, 56, 95] (2) [ ] (3) [1204]
(2) Equal Frequency Discretization	In this case we need to choose bin intervals in such a way that there is almost an equal number of elements in each of the three bins. If we choose the bin intervals as: $-\infty$ to 13, 13 to 41 and 41 to ∞ we will get 4 elements in three bins. This results in the following three vectors representing the values in the three bins: (1) [-212, -25, 10, 12] (2) [30, 30, 30.21, 40] (3) [50, 56, 95, 1204]

---

[10] 4.

(a) The following table gives the Hamming distance and Jaccard similarity between the two vectors: x = 0101010001 and y = 0100011000

Measure	**Result**
(a) Hamming distance	Since Hamming distance is the number of bits that are different between two vectors, the Hamming distance between x and y = **3 or 0.3**
(b) Jaccard similarity	Jaccard similarity coefficient is given by number of 11 matches between the binary vectors (M_{11}) divided by the number of 11, 01 or 10 matches between the binary vectors ($M_{11} + M_{01} + M_{10}$). In the given vectors $M_{11} = 2$, $M_{01} = 1$ & $M_{10} = 2$. Therefore, the Jaccard similarity is $= 2/(2 + 1 + 2) = \mathbf{0.4}$

(b) The Jaccard Coefficient is similar to Simple Matching Coefficient (SMC), because in both the cases, similarity between the two vectors is computed relative to the number of matches and mis-matches between the two vectors. While, Hamming distance is similar to cosine similarity because Hamming distance considers the difference between each element of two vectors. Since the angle between two vectors is considered in Cosine similarity the more the difference between individual elements of the two vectors, the greater would be the angle between them. In summary, while Jaccard coefficient and SMC try to estimate the similarity between two vectors, the Hamming distance and Cosine similarity try to estimate the difference or distance between two vectors.

---

[10] 5.

The following table gives the relevant similarity and distance measure values between the vectors x = (0, 1, 0, 1) and y = (1, 0, 1, 0):

Similarity/Distance Measure	**Result**
(a) Cosine	$cos(x, y) = (x \cdot y)/\|x\|\|y\|$ $= (0/1.4142 * 1.4142)$ $= 0$
(b) Correlation	$corr(x, y) = covariance(x, y)/(\sigma_x * \sigma_y)$ $= -0.333\overline{3}/(0.57735 * 0.57735)$ ≈ -1

(c) Euclidean	$dist(x,y) = \sqrt{\sum_{k=1}^{n}(x_k - y_k)^2}$ $= 2$
(d) Jaccard	$J(x,y) = M_{11}/(M_{11} + M_{01} + M_{10})$ $= 0/4$ $= 0$

---

[10] 6.

The following table shows the minimum and maximum possible values for the different distance measures between Mike and John as Mike is jogging:

Distance Measure	**Minimum value**	**Maximum value**
(a) Manhattan	1 mile	$\sqrt{2}$ miles
(b) Euclidean	1 mile	1 mile
(c) Chebyshev	1 mile	1 mile

---

[10] 7.

The following table gives the range (Lower bound and upper bound) for the different distance and similarity measures for a n-dimensional space:

Distance/Similarity Measure	**Range**
(i) Euclidean Distance	The *straight-line* distance between any two points or *vectors* can range from: 0 to ∞
(ii) Cosine Similarity	Cosine similarity measures the $\cos\theta$ of the angle (θ) between two vectors, its range is bound by the maximum and minimum possible values of $\cos\theta$. Therefore, the range for cosine similarity is: -1 to 1

(iii) Simple Matching Coefficient (SMC)	SMC is the ratio of number of matching attribute values to the total number of attribute values. Therefore, this ranges from: 0 (no matching values) to 1 (when all values match)
(iv) Hamming Distance	Hamming distance is the number of attribute values that are different betweeen two vectors. Hence, this can range from: 0 (all values match between the vectors) to ∞ (the two vector are of infinite length and none of the values match)

---

[10] 8.

(a) Two vectors x and y have zero mean. What is the relationship of the cosine measure and correlation between them?

$$Given, \overline{x} = 0, \overline{y} = 0 \tag{1}$$

$$covariance(x, y) = \frac{1}{n-1}\sum_{k=1}^{n}(x_k - \overline{x}) * (y_k - \overline{y}) \tag{2}$$

From (1) and (2);

$$covariance(x, y) = \frac{1}{n-1}\sum_{k=1}^{n}(x_k * y_k) \tag{3}$$

Also,

$$\sigma(x) = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(x_k - \overline{x})^2} \tag{4}$$

From (1) and (4);

$$\sigma(x) = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(x_k)^2} \tag{5}$$

Similarly,

$$\sigma(y) = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(y_k)^2} \tag{6}$$

As,

$$corr(x,y) = \frac{covariance(x,y)}{\sigma_x * \sigma_y} \tag{7}$$

From (3),(5),(6) and (7) we have,

$$corr(x,y) = \frac{\frac{1}{n-1}\sum_{k=1}^{n}(x_k * y_k)}{\sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(x_k)^2} * \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(y_k)^2}} \tag{8}$$

$$= \frac{\sum_{k=1}^{n}(x_k * y_k)}{\sqrt{\sum_{k=1}^{n}(x_k)^2} * \sqrt{\sum_{k=1}^{n}(y_k)^2}} \tag{9}$$

$$= \frac{(x \cdot y)}{|x||y|} \tag{10}$$

Since,

$$cos(x,y) = \frac{x \cdot y}{|x||y|} \tag{11}$$

Therefore,

$$corr(x,y) = cos(x,y) \tag{12}$$

Hence, if the mean of two vectors is zero, their correlation coefficient equals the cosine similarity between them.

(b) Derive the mathematical relationship between cosine similarity and Euclidean distance when each data object vector has an L2 length (magnitude) of 1. (NOTE: your final answer should be independent of the original vectors).

Given,

$$|x| = 1, |y| = 1 \tag{1}$$

Cosine similarity between x and y is given by,

$$cos(x,y) = \frac{x \cdot y}{|x||y|} \tag{2}$$

By (1) and (2),

$$cos(x,y) = x \cdot y \tag{3}$$

$$= \sum_{k=1}^{n}(x_k * y_k) \tag{4}$$

Now Euclidean distance between x and y is given by,

$$dist(x, y) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2} \quad (5)$$

$$= \sqrt{\sum_{k=1}^{n} ({x_k}^2 - {y_k}^2 + 2 * x_k * y_k)} \quad (6)$$

$$= \sqrt{\sum_{k=1}^{n} ({x_k}^2) + \sum_{k=1}^{n} ({y_k}^2) - 2 * \sum_{k=1}^{n} (x_k * y_k)} \quad (7)$$

$$= \sqrt{|x| + |y| - 2 * \sum_{k=1}^{n} (x_k * y_k)} \quad (8)$$

By (1), (4) and (8), we have,

$$dist(x, y) = \sqrt{1 + 1 - 2 * cos(x, y)} \quad (9)$$

$$= \sqrt{2 * (1 - cos(x, y))} \quad (10)$$

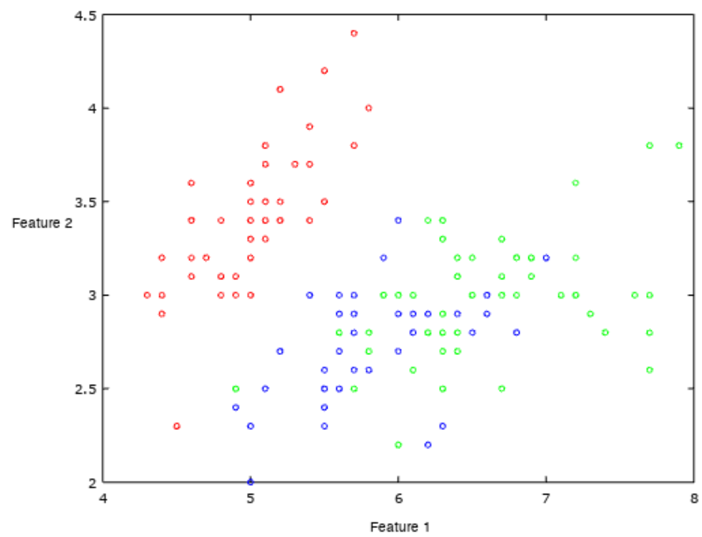
---

[20] 9.

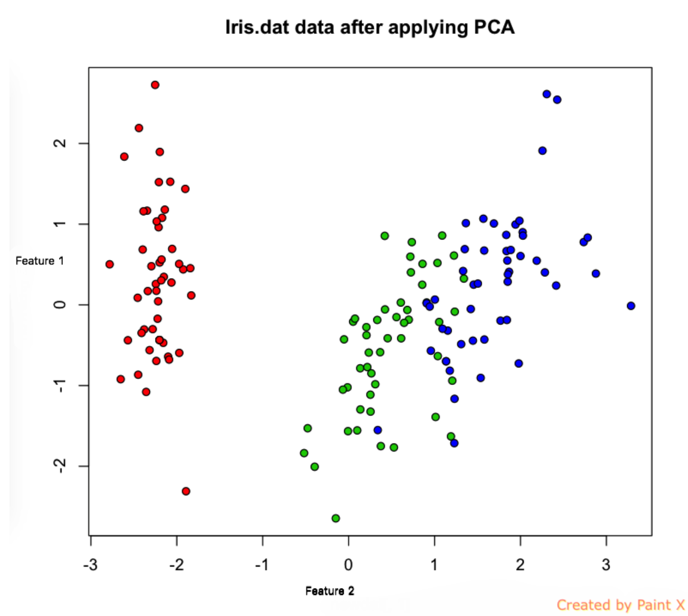
(a) The data matrix has **150** data points, **4** features and **3** classes. The following table gives the basic statistics for the iris data set:

Statistic	**feature 1**	**feature 2**	**feature 3**	**feature 4**
Mean	5.8433	3.0540	3.7587	1.1987
Median	5.8000	3.0000	4.3500	1.3000
Std. Dev.	0.82807	0.43359	1.76442	0.76316
Min Value	4.30000	2.00000	1.00000	0.10000
Max Value	7.9000	4.4000	6.9000	2.5000

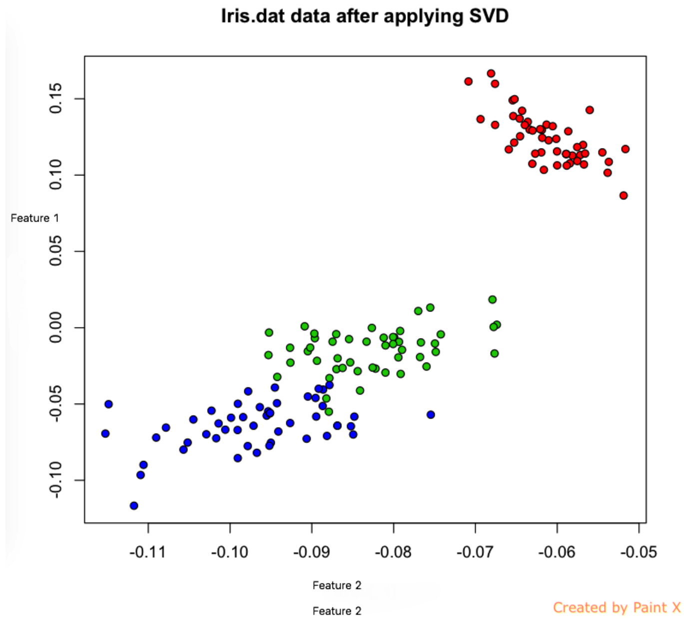
(b) The following figure shows the plot of the first two features of the data. The three classes are shown by red (class = 1), blue (class = 2) and green (class = 3) colors:



(c) The following figure shows the plot of the two features of the data after applying PCA for dimensionality reduction. The three classes are shown by red (class = 1), blue (class = 2) and green (class = 3) colors:



(d) The following figure shows the plot of the two features of the data after applying SVD for dimensionality reduction. The three classes are shown by red (class = 1), blue (class = 2) and green (class = 3) colors:



The 2 figures (c and d) are very different because we get different feature values after dimensionality reduction in both the cases. In both the cases, the red colored class (label=1) is clearly separated from the other two classes.
