CSC 4740/6740 **Data mining** **D**ue by: Sept 26.

Name:

Assignment 2

Q1: **Explain** the following *normalization methods by using formula* , the value ranges of output etc. (10 points)

(a) min-max normalization

(b) z-score normalization

(c) z-score normalization using the mean absolute deviation instead of standard deviation

(d) normalization by decimal scaling

Q2: Use the methods below to *normalize* the following group of data: (10 points)

200, 300, 400, 600, 1000

(a) min-max normalization by setting $min = 0$ and $max = 1$

(b) z-score normalization

(c) z-score normalization using the mean absolute deviation instead of standard deviation

(d) normalization by decimal scaling

Q3: Use a flowchart to explain the following procedures for *attribute subset selection*: (10 points)

(a) stepwise forward selection

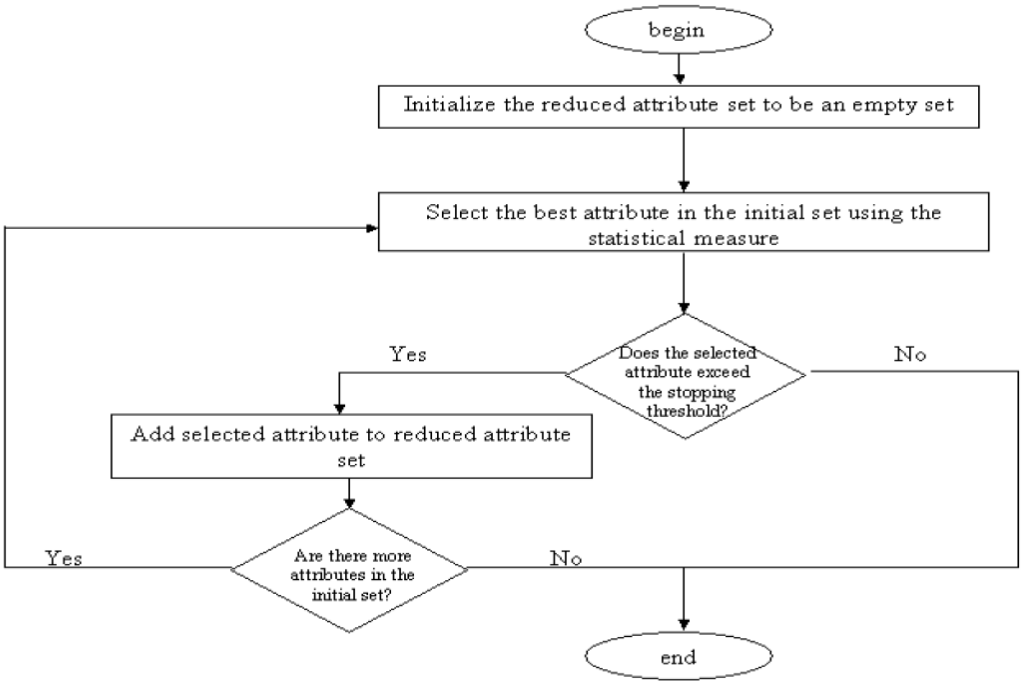
begin
Initialize the reduced attribute set to be an empty set
Select the best attribute in the initial set using the statistical measure
Yes
Does the selected attribute exceed the stopping threshold?
No
Add selected attribute to reduced attribute set
Yes
Are there more attributes in the initial set?
No
end

Figure 3.1: Stepwise forward selection.

CSC 4740/6740 **Data mining** **D**ue by: Sept 26.

Name:

Q4: The following table contains the attributes *name, trait-1, trait-2, trait-3*, and *trait-4*,etc where *name* is an object identifier, attributes are describing personal traits of individuals who desire a penpal. Suppose that a service exists that attempts to find pairs of compatible penpals by computing the similarity of each pair. Compute the similarity of all pairs and which pair is the most compatible pair under 1) symmetric attributes 2) asymmetric attributes assumption ? (10 points)

name	Trait 1	Trait 2	Trait 3	Trait 4	Trait 5	Trait6	Trait 7
Kevin	P	P	P	N	N	N	P
Eric	P	N	P	P	N	N	N
Caroline	N	P	N	P	N	N	N
…	.	…	…		.		.

Q5: The following is data collected from 30 participants from 10 attributes, (55 points)

Step 1: Please check data quality, do necessary cleaning and/or transformation steps, and explain what you did, and why (10 points) Hint: consider missing, outliers, scale/normalization

Step 2: a) Compute covariance matrix of data among all variables after step1. (this should be a 10-by-10 matrix) Note: Python : be aware that **Each row of data array represents a variable** (5 points),

b) compute the total variance of data = sum of diagonal elements of covariance matrix (5 points),

c) compute correlation (Pearson's correlation) between variable 1 and variable 2 (5 points),

Step 3: perform **Principal component analysis (PCA)** to generate a number of Principal Components (PCs) capturing >85% of total data variance.

a) Plot percentage of variances of each Principal Components(PC) in a decreasing order (5 points)
b) How many components do you need to capture > 85% total data variance? (5 points)
c) Plot the PC (or projection direction) of N components you selected (5 points, N lines in one plot or N separate plots)

d) Plot the generated (**NEW**) top P PC variables you selected (5 points, N lines in one plot or N separate plots)

e) Compute the covarion matrix of the NEW P PC variables (this should be P-by-P matrix), **and** compute the total variance of PCs (sum of diagonal elements of covariance matrix), compare this value with the total variance of data in Step2_b, what % variance kept in PCs (5 points),

f) compute correlation (Pearson's correlation) between new variable PC1 and new variable PC2 (5 points),

CSC 4740/6740 **Data mining** **D**ue by: Sept 26.

Name:

Q6: Graduate students Please apply UMap or tSNE for visualization of clean data in 2D. Since we only have 30 data points, please select parameters with smaller neighbors. (5 points)

CSC 4740/6740 **Data mining** **D**ue by: Sept 26.

Name:

ID	**Fluid IQ**	**Crystallized IQ**	**Vocabulary**	**Inhibitory control**	**Memory**	**Mental flexibility**	**Processing Speed**	**Attention Problem**	**Anxiety problem**	**Social problems**
1	122	77	131	81	86	86	80	8.6	7	8
2	103	77	98	69	97	84	57	10	8.5	9
3	148	91	153	89	109	87	67	7.8	8	7.2
4	137	107	142	106	105	102	94	7.6	6.6	5.6
5	0	0	0	0	0	0	0	7.8	6.8	7.5
6	101	87	102	75	94	90	88	8.8	7.7	6.7
7	102	89	122	86	97	90	88	7.6	7.6	7
8	72	63	82	57	67	88	78	8.2	8.5	7
9	148	91	120	97	105	74	86	9.8	9	8.2
10	116	99	109	95	105	101	101	7.2	7.3	8.2
11	107	101	102	101	105	97	97	7.8	7.6	7
12	131	90	153	93	101	88	84	8.4	8.8	7
13	110	75	98	71	97	89	67	7.4	8.8	7.5
14	84	82	98	93	82	91	86	7.4	8.2	6.9
15	125	99	112	98	105	105	90	8.2	9.7	8
16	110	95	109	91	101	90	94	6.5	7.7	6.4
17	113	100	112	99	105	96	97	9.2	7	6.2
18	95	93	92	97	94	94	90	8.2	8.5	7.5
19	66	80	72	84	86	95	71	8	8.3	7
20	103	91	120	96	97	72	92	8	7.9	8.8
21	142	96	122	92	109	98	67	7	8.3	6.7
22	84	91	92	102	82	91	90	7.2	7.6	6.7
23	116	84	131	90	70	85	101	8.6	7.6	7.3
24	110	81	98	86	101	74	90	7.6	7.6	7.3
25	116	82	120	91	101	81	74	8	7.3	6.7
26	97	74	98	67	86	79	78	7.2	9.4	8.3
27	84	72	77	76	74	73	88	8	7.9	7
28	142	106	120	104	109	100	90	7.6	5.3	5.9
29	84	85	92	97	86	93	82	8	5.6	6.4
30	97	81	98	91	97	90	59	8.4	6.6	6.5