1. Three students, Linda, Tuan, and Javier, are given five laboratory rats each for a nutritional experiment. Each rat's weight is recorded in grams. Linda feeds her rats **Formula A**, Tuan feeds his rats **Formula B**, and Javier feeds his rats **Formula C**. At the end of a specified time period, each rat is weighed again, and the net gain in grams is recorded. Using a significance level of 5%, test the hypothesis that the three formulas produce the same mean weight gain.

Linda’s rats	Tuan’s rats	Javier’s rats
43.5	47.0	51.2
39.4	40.5	40.9
41.3	38.9	37.9
46.0	46.3	45.0
38.2	44.2	48.6

There are three samples, we use ANOVA F-test. Assume the weight of Linda’s rats Y_1, the weight of Tuan’s rats Y_2, the weight of Javier’s rats Y_3.

$\overline{Y}$ =43.26, $\overline{Y_1}$ =41.68, $\overline{Y_2}$ =43.38, $\overline{Y_3}$ =44.72

SST=$n_1(\bar{Y}_1-\bar{Y})^2+n_2(\bar{Y}_2-\bar{Y})^2+n_3(\bar{Y}_3-\bar{Y})^2$

= $5*(41.68-43.26)^2+5*(43.38-43.26)^2+5*(44.72-43.26)^2$

= 5*(2.4964+0.0144+2.1316) = 23.212

SSE=$(Y_{11}-\bar{Y}_1)^2+(Y_{21}-\bar{Y}_1)^2+\cdots+(Y_{51}-\bar{Y}_1)^2$

$+(Y_{12}-\bar{Y}_2)^2+(Y_{22}-\bar{Y}_2)^2+\cdots+(Y_{52}-\bar{Y}_2)^2$

$+(Y_{13}-\bar{Y}_3)^2+(Y_{23}-\bar{Y}_3)^2+\cdots+(Y_{53}-\bar{Y}_3)^2$

= $(43.5-41.68)^2+(39.4-41.68)^2+(41.3-41.68)^2+(46.0-41.68)^2+(38.2-41.68)^2$

$+(47.0-43.38)^2+(40.5-43.38)^2+(38.9-43.38)^2+(46.3-43.38)^2+(44.2-43.38)^2$

$+(51.2-44.72)^2+(40.9-44.72)^2+(37.9-44.72)^2+(45.0-44.72)^2+(48.6-44.72)^2$

= 39.428+50.668+118.228 = 208.324

p-1 = 2

n-p = 12

$$F=\frac{SST/(p-1)}{SSE/(n-p)}=\frac{23.212/2}{208.324/12}=0.67$$

F_{crit} = 3.89

df_N = p-1 = 2

$df_D = n-p = 12$

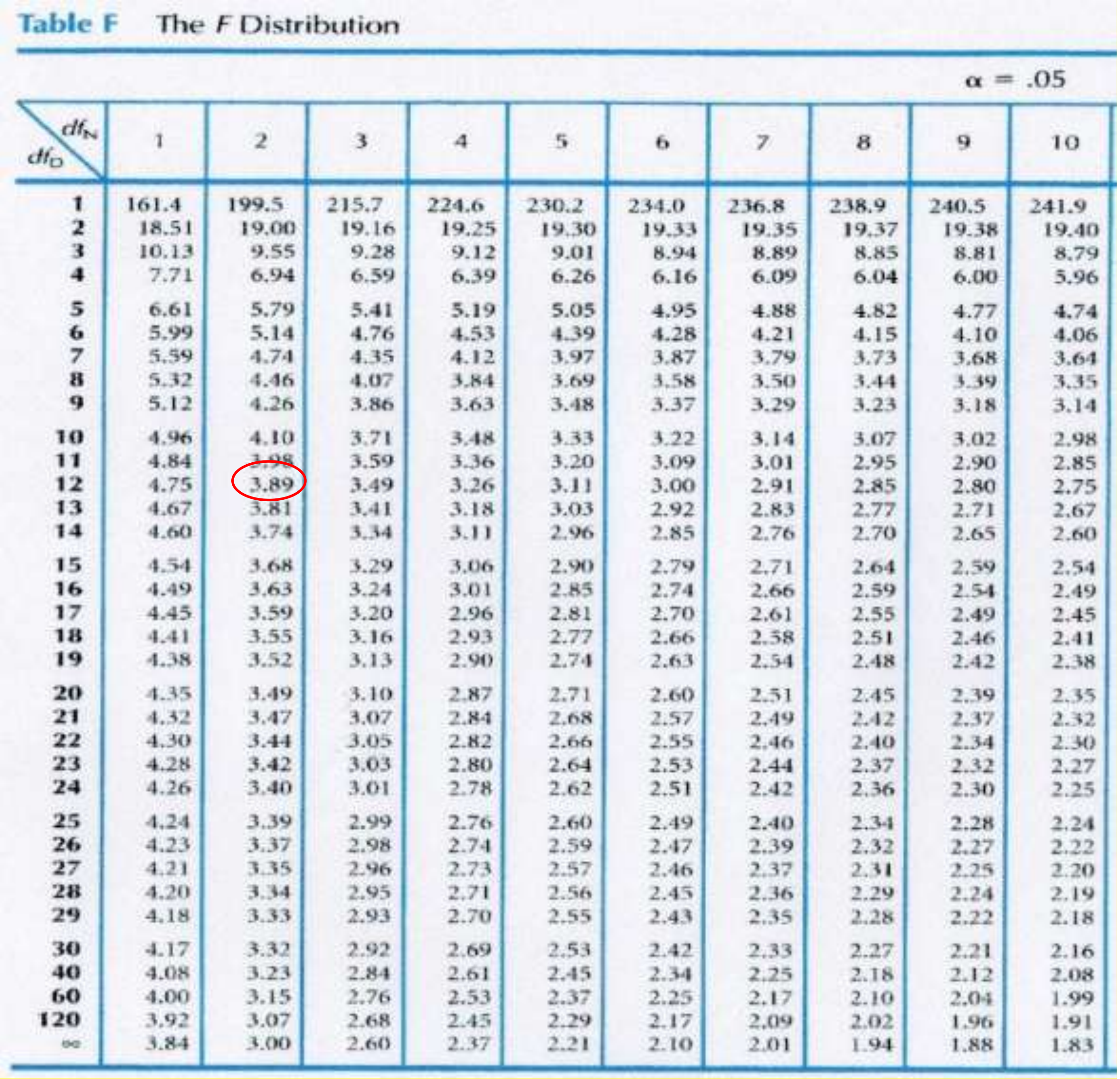


Table F The *F* Distribution

$\alpha = .05$

df_D \ df_N	1	2	3	4	5	6	7	8	9	10
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99
120	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96	1.91
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83

$F = 11.606/17.36 = 0.67 < 3.89$

Fail to reject the null hypothesis, that the three formulas produce the same mean weight gain.

2. UK Biobank database has recruited more than 5 million participants from different cities of UK. It collected the participant's lab test results, lifestyle and disease history from questionnaire and clinical records. In Table 1, we showed the selected information (including disease, weight, age and smoking status) of 30 white British participants.

Participant ID	Disease	Weight (kg)	Age	Smoking
1	Type 2 Diabetes	70	69	Non-Smoker

2	Lung cancer	54	74	Smoker
3	Type 2 Diabetes	68	31	Non-Smoker
4	Lung cancer	51	26	Smoker
5	Lung cancer	49	89	Non-Smoker
6	Lung cancer	46	54	Smoker
7	Hypertension	52	86	Smoker
8	Lung cancer	40	46	Smoker
9	Hypertension	59	54	Smoker
10	Type 2 Diabetes	65	31	Non-Smoker
11	Lung cancer	40	64	Smoker
12	Hypertension	79	64	Smoker
13	Type 2 Diabetes	84	22	Smoker
14	Hypertension	63	23	Non-Smoker
15	Type 2 Diabetes	74	41	Smoker
16	Type 2 Diabetes	52	89	Smoker
17	Hypertension	50	75	Non-Smoker
18	Lung cancer	56	25	Smoker
19	Type 2 Diabetes	88	60	Non-Smoker
20	Lung cancer	44	26	Smoker
21	Hypertension	85	73	Non-Smoker
22	Lung cancer	58	73	Non-Smoker
23	Type 2 Diabetes	81	45	Non-Smoker
24	Hypertension	75	28	Smoker
25	Hypertension	60	23	Smoker
26	Lung cancer	56	46	Smoker
27	Type 2 Diabetes	81	53	Smoker
28	Hypertension	61	23	Non-Smoker
29	Type 2 Diabetes	64	70	Non-Smoker
30	Hypertension	83	31	Smoker

Table 1. The information of thirty participants of UK Biobank

2.1. According to the information shown in **Table 1**, calculate the mean weight and its 95% confidence interval of the participants with Type 2 Diabetes? Is it sample mean or population mean?

Answer:

Mean weight: $\frac{70+68+65+84+74+52+88+81+81+64}{10} = 72.7\text{kg}$

Sample standard deviation is

$S =$

$$= \sqrt{\frac{(70-72.7)^2+(68-72.7)^2+(65-72.7)^2+(84-72.7)^2+(74-72.7)^2+(52-72.7)^2+(88-72.7)^2+(81-72.7)^2+(81-72.7)^2+(64-72.7)^2}{10-1}}$$

$$= \sqrt{\frac{7.29+22.09+59.29+127.69+1.69+428.49+234.09+68.89+68.89+75.69}{9}}$$

$$= \sqrt{\frac{1094.1}{9}} = 11.03$$

$$S_{\bar{x}} = \frac{11.03}{\sqrt{10}} = 3.49$$

95% confidence interval: [72.7-2*3.49, 72.7+2*3.49]=[65.72,79.68]

This is a sample mean

2.3. Calculate Pearson correlation, Spearman correlation between age and weight for the smokers with lung cancer.

Answer:

Participant ID	Disease	Weight (kg)	Age	Smoking
2	Lung cancer	54	74	Smoker
4	Lung cancer	51	26	Smoker
6	Lung cancer	46	54	Smoker
8	Lung cancer	40	46	Smoker
11	Lung cancer	40	64	Smoker
18	Lung cancer	56	25	Smoker
20	Lung cancer	44	26	Smoker
26	Lung cancer	56	46	Smoker

Pearson correlation:

$$r_{XY} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}}$$

(age) $\overline{X}$ = (74+26+54+46+64+25+26+46)/8 = 45.125

(weight) $\overline{Y}$ = (54+51+46+40+40+56+44+56)/8 = 48.375

$$r_{XY} = \frac{(74-45.125)\times(54-48.375)+\cdots+(46-45.125)\times(56-48.375)}{\sqrt{(54-48.375)^2+\cdots+(56-48.375)^2}\times\sqrt{(74-45.125)^2+\cdots+(46-45.125)^2}}$$

$$= \frac{-137.375}{17.885 \times 49.060}$$

= -0.157

Pearson correlation: -0.157

2.4 Are the weights of smokers and non-smokers with hypertension significantly different each other (α=0.05, t_{crit}=±2.306)? Please state (A) what is the null hypothesis and alternative hypothesis? (B) What kind of t-test would you like to use? (C) One tailed or two tailed tests?

Answer:

(A) H_0 (null hypothesis): The weights between smokers and non-smokers with hypertension are the same

H_a (Alternative hypothesis): The weights between smokers and non-smokers with hypertension are different from each other

(B) We will select the independent sample T Test

(C) We will select two-tailed T test

Full procedure for independent t-test

1. Identify

Pop1 (variable x): The weights of smokers with hypertension

Pop2 (variable y): The weights of non-smokers with hypertension

2. State the null and research hypotheses

H_0 (null hypothesis): The average weights between smokers and non-smokers with hypertension are the same

H_1 (Alternative hypothesis): The average weights between smokers and non-smokers with hypertension are different from each other

3. Determine characteristics of comparison distribution (distribution of differences between means)

Population: $\mu_x = \mu_y$ (i.e., no difference between means)

$$\bar{x} = 68, \bar{y} = 64.75$$

$$S_x^2 = \frac{\sum_i (x_i - \bar{x})^2}{n-1} = \frac{(52-68)^2 + \cdots (83-68)^2}{6-1} = 159.2$$

$$S_y^2 = \frac{\sum_i (y_i - \bar{y})^2}{n-1} = \frac{(63-64.75)^2 + \cdots (61-64.75)^2}{4-1} = 214.92$$

$$df_{Total} = df_x + df_y = 5 + 3 = 8$$

$$S_{pooled}^2 = \frac{df_x}{df_{Total}} S_x^2 + \frac{df_y}{df_{Total}} S_y^2$$

$$=\frac{5}{8} * 159.2 + \frac{3}{8} * 214.91$$

= 99.5+80.59=180.09

$$S_{\bar{x}}^2 = \frac{180.09}{6} = 30.02$$

$$S_{\bar{y}}^2 = \frac{180.09}{4} = 45.02$$

$$S_{Difference} = \sqrt{30.02 + 45.02} = 8.66$$

4. Determine critical value (cutoffs)

In Behavioral Sciences, we use α = .05 (5%)

Our hypothesis is *nondirectional*, so our hypothesis test is *two-tailed*.

t_{crit}= ±2.306

5. Calculate a test statistic

$$t = \frac{\bar{x} - \bar{y}}{S_{Difference}} = \frac{68 - 64.75}{8.66} = \frac{3.25}{8.66} = 0.375$$

6. Make a decision

Because $t = 0.375 < 2.306$, we fail to reject the null hypothesis and make the conclusion that the average weights between smokers and non-smokers with hypertension are the same.

3. Which of the following measures are resistant to outliers? (AC)

A. Median
B. Mean
C. Interquartile Range (IQR)
D. Range

4. Which is the most likely method to avoid sampling bias? (C)

A. Pick more number items from the population
B. Pick the odd-numbered data items from the population
C. Randomly pick data from the population
D. Pick the first 10 data items from the population

5. When the correlation coefficient, r, is close to one: (B)

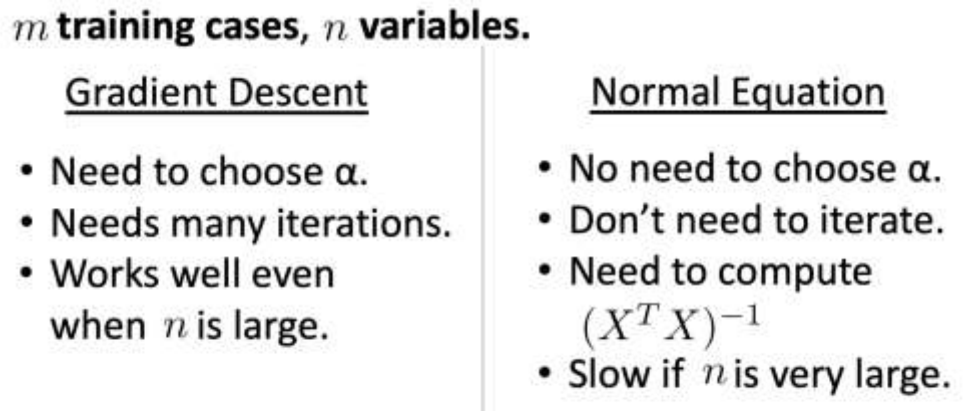
A. there is no relationship between the two variables
B. there is a strong linear relationship between the two variables
C. it is impossible to tell if there is a relationship between the two variables
D. the slope of the regression line will be close to one

6. Given a dataset that follows normal distribution, if you want to compare two independent samples, which of the following statistical tests should be chosen? (C)

A. Single sample t-test

B. Paired sample t-test
C. Two-group t test
D. ANOVA

8. What are the differences between gradient descent and normal equation to estimate the parameters in linear regression?



m **training cases,** n **variables.**

Gradient Descent

- Need to choose α.
- Needs many iterations.
- Works well even when n is large.

Normal Equation

- No need to choose α.
- Don't need to iterate.
- Need to compute $(X^TX)^{-1}$
- Slow if n is very large.

9**.** A food factory mainly produces the bagged vegetable, which could produce 5,000 bags every day. In expectation, the weight of each bagged vegetable is 200g. In order to guarantee high quality of the products, the quality assurance department will examine 8 bagged vegetables every day. Now we have their weights for three days (Assume all the weights listed below follow normal distribution)

Day1: 210, 220, 183, 198, 205, 210, 180, 195

Day2: 198,196,155,138,122,100,220,210

Day3: 183,175,120,182,110,135,188,159

9.1 What are the mean and standard deviation of Day1's products? Is the mean of the products in Day1 the "sample mean" or "population mean"?

Mean: $\bar{X} = \frac{210+220+183+198+205+210+180+195}{8} = 200.125$

Standard deviation: $s = \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{(210-200.125)^2 + \cdots (195-200.125)^2}{8-1}} = 13.85$

It is the "sample mean"

9.2 What is the median, 25 percentile and 75 percentile in Day2's products?

First, we sort the product weights in Day 2: 100,122,138,155,196,198,210,220

100,122,138, 155, 155,196, 196, 198,210,220

$$m = \frac{155+196}{2} = 175.5$$

$$Q1 = \frac{122 + 138}{2} = 130$$

$$Q3 = \frac{198 + 210}{2} = 204$$

9.3 Are the mean weights of the products from Day 1, Day 2 and Day 3 significantly different from each other (F_{crit}= ±3.467 at confidence level 0.05)? Please list all the essential steps.

9.3.1 What are the null hypothesis and alternative hypothesis?

H_0: The average weight of the products from the three days are the same

H_a: At least the average weight of one day is different from the others

9.3.2 What is the average weight for each day and what is the overall average weight for the three days?

$$\bar{Y}_1 = 200.125$$

$$\bar{Y}_2 = 167.375$$

$$\bar{Y}_3 = 156.5$$

$$\bar{Y} = 174.667$$

9.3.3 What are the between group variation and in group variation?

SST=$n_1(\bar{Y}_1 - \bar{Y})^2 + n_2(\bar{Y}_2 - \bar{Y})^2 + n_3(\bar{Y}_3 - \bar{Y})^2$

=$8 * (200.125 - 174.667)^2 + 8 * (167.375 - 174.667)^2 + 8 * (156.5 - 174.667)^2$

$= 8 * 648.11 + 8 * 53.17 + 8 * 330.04$=8250.56

SSE=$(Y_{11} - \bar{Y}_1)^2 + (Y_{21} - \bar{Y}_1)^2 + \cdots + (Y_{81} - \bar{Y}_1)^2$

+$(Y_{12} - \bar{Y}_2)^2 + (Y_{22} - \bar{Y}_2)^2 + \cdots + (Y_{82} - \bar{Y}_2)^2$

+$(Y_{13} - \bar{Y}_3)^2 + (Y_{23} - \bar{Y}_3)^2 + \cdots + (Y_{83} - \bar{Y}_3)^2$=1342.875+13957.875+6650=21950.75

9.3.4 What is the degrees of freedoms?

p=3, n=24, v1=p-1=3-1=2, v2=n-p=24-3=21

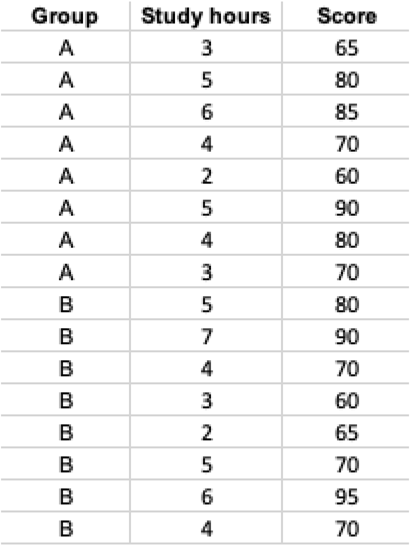
9.3.5 What is the F value?

$$F = \frac{8250.56/2}{21950.75/21} = \frac{4125.28}{1045.27} = 3.95 > 3.467$$

9.3.6 What is the conclusion?

Conclusion: we reject the null hypothesis, at $\alpha = 0.05$. At least the average weight of one day is different from the others.

10. A visualization design question



Group	Study hours	Score
A	3	65
A	5	80
A	6	85
A	4	70
A	2	60
A	5	90
A	4	80
A	3	70
B	5	80
B	7	90
B	4	70
B	3	60
B	2	65
B	5	70
B	6	95
B	4	70

Calculate Pearson correlation between study hours and score for the students in Group A.

(study hours) $\overline{X}$ = (3+5+6+4+2+5+4+3)/8 = 4

(score) $\overline{Y}$ = (65+80+85+70+60+90+80+70)/8 = 75

$$r_{XY} = \frac{(3-4)\times(65-75)+(5-4)\times(80-75)+\cdots+(3-4)\times(70-75)}{\sqrt{(3-4)^2+(5-4)^2\ldots+(3-4)^2}\times\sqrt{(65-75)^2+(80-75)^2\ldots+(70-75)^2}}$$

$$= \frac{85}{\sqrt{12}\times\sqrt{750}}$$

= 0.896

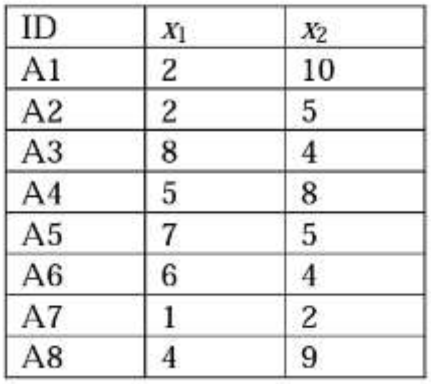
11. Explain the reason why the nonlinear activation function is required in a neural network model.

Answer: Without non linear activation function, the neural network will be reduced to a linear model.

12. Discuss the differences between perceptron algorithm and support vector machines.

Answer: SVM aims to find the separating hyperplane with maximum margin whereas perceptron does not incorporate this maximum margin principle. Perceptron algorithm will not converge on non-separable data where SVM can. SVM can solve nonlinear classification problem whereas the perceptron algorithm can not.

13. Suppose that we want to cluster eight data points as shown in the following table (x1 and x2 are the two features) into three clusters.

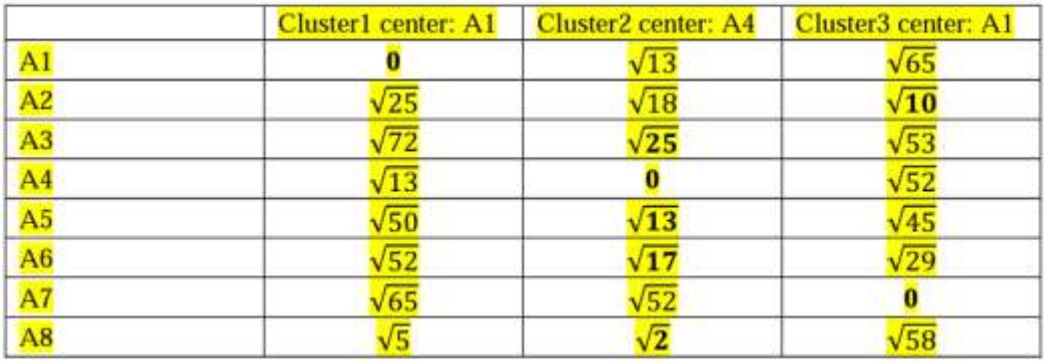


ID	x_1	x_2
A1	2	10
A2	2	5
A3	8	4
A4	5	8
A5	7	5
A6	6	4
A7	1	2
A8	4	9

The distance function is Euclidean distance. Suppose initially we assign A1, A4, A7 as the center of each cluster. Run the k-means algorithm for 1 iteration, at the end of the iteration show

(a) The cluster assignments (i.e., which samples belong to which clusters)

Answer:



	Cluster1 center: A1	Cluster2 center: A4	Cluster3 center: A1
A1	**0**	$\sqrt{13}$	$\sqrt{65}$
A2	$\sqrt{25}$	$\sqrt{18}$	**$\sqrt{10}$**
A3	$\sqrt{72}$	**$\sqrt{25}$**	$\sqrt{53}$
A4	$\sqrt{13}$	**0**	$\sqrt{52}$
A5	$\sqrt{50}$	**$\sqrt{13}$**	$\sqrt{45}$
A6	$\sqrt{52}$	**$\sqrt{17}$**	$\sqrt{29}$
A7	$\sqrt{65}$	$\sqrt{52}$	**0**
A8	$\sqrt{5}$	**$\sqrt{2}$**	$\sqrt{58}$

cluster 1: {A1}; cluster 2:{A3, A4, A5, A6, A8}; cluster 3: {A2, A7}

(b) The center of the new clusters

Answer: new cluster centers {2, 10}, {6, 6}, {1.5, 3.5}

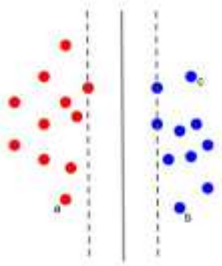
14. Describe at least two limitations of k-means algorithm. And how to address these two limitations?

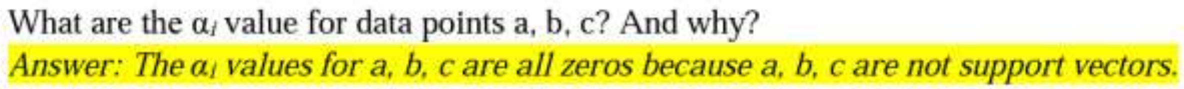
Answer: (1) K-means is extremely sensitive to cluster center initialization. We can try multiple initialization and choose the best result to address this limitation; (2) Works only if the cluster are round shaped and of equal size/density cluster. Probabilistic cluster methods can address this limitation; (3) Can not capture non-linear structure. Kernel k-means can address this limitation; (4) Makes hard assignments of points to clusters. Gaussian mixture model can address this limitation.

15. In K-means algorithm and K-nearest neighbor algorithm, which one can be used for classification problem? And why?

Answer: K-nearest neighbor algorithm can be used for classification problem. K-means is a clustering algorithm.

16. Consider the following figure that shows the decision boundary of SVM on a dataset.

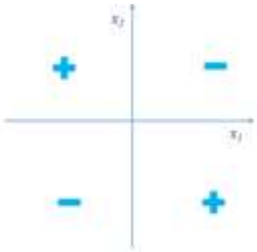


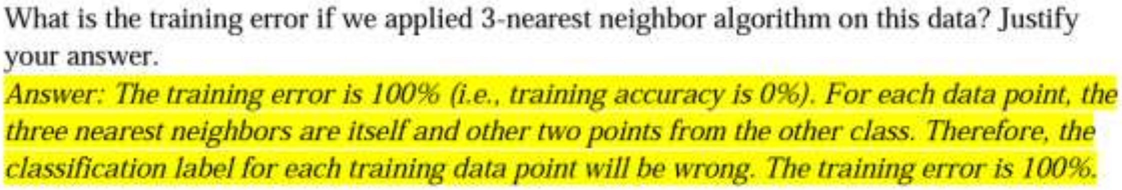


What are the α_i value for data points a, b, c? And why?

Answer: The α_i values for a, b, c are all zeros because a, b, c are not support vectors.

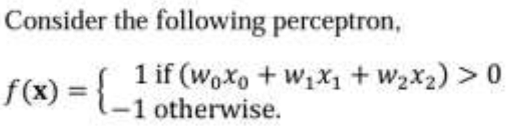
17. Consider the following data set





What is the training error if we applied 3-nearest neighbor algorithm on this data? Justify your answer.

Answer: The training error is 100% (i.e., training accuracy is 0%). For each data point, the three nearest neighbors are itself and other two points from the other class. Therefore, the classification label for each training data point will be wrong. The training error is 100%.

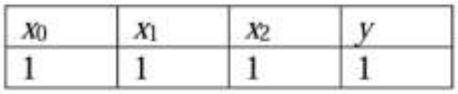


Consider the following perceptron,

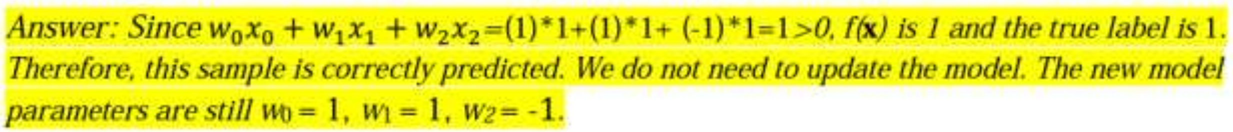
$$f(\mathbf{x}) = \begin{cases} 1 \text{ if } (w_0x_0 + w_1x_1 + w_2x_2) > 0 \\ -1 \text{ otherwise.} \end{cases}$$

18.

Suppose the current model parameters are $w_0 = 1$, $w_1 = 1$, $w_2 = -1$. Is the following data sample correctly predicted? What are the new model parameters after updating model based on this data sample?

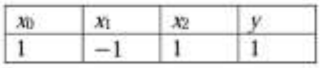


x_0	x_1	x_2	y
1	1	1	1

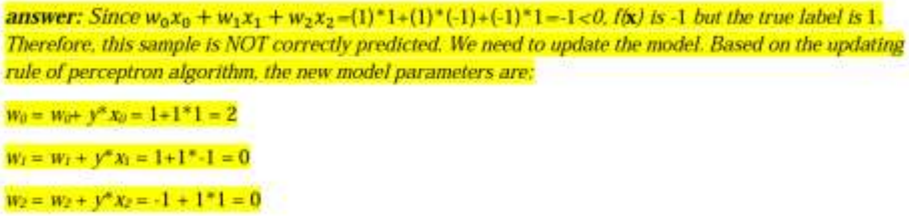


*Answer: Since $w_0x_0 + w_1x_1 + w_2x_2$=(1)*1+(1)*1+ (-1)*1=1>0, f(**x**) is 1 and the true label is 1. Therefore, this sample is correctly predicted. We do not need to update the model. The new model parameters are still w_0 = 1, w_1 = 1, w_2= -1.*

19. Suppose the current model parameters are $w_0 = 1$, $w_1 = 1$, $w_2 = -1$. Is the following data sample correctly predicted? What are the new model parameters after updating model based on this data sample?



x_0	x_1	x_2	y
1	−1	1	1



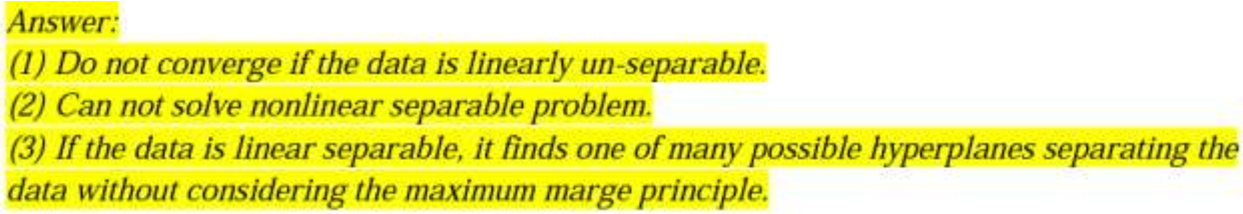
answer: *Since $w_0x_0 + w_1x_1 + w_2x_2$=(1)*1+(1)*(-1)+(-1)*1=-1<0, f(**x**) is -1 but the true label is 1. Therefore, this sample is NOT correctly predicted. We need to update the model. Based on the updating rule of perceptron algorithm, the new model parameters are:*

$w_0 = w_0 + y^* x_0$ = 1+1*1 = 2

$w_1 = w_1 + y^* x_1$ = 1+1*-1 = 0

$w_2 = w_2 + y^* x_2$ = -1 + 1*1 = 0

20. Describe two limitations of perceptron algorithm. And how to address to these two limitations.



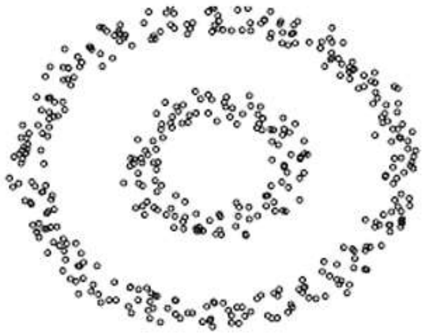
Answer:

(1) Do not converge if the data is linearly un-separable.

(2) Can not solve nonlinear separable problem.

(3) If the data is linear separable, it finds one of many possible hyperplanes separating the data without considering the maximum marge principle.

21. Given the following dataset, will k-means work well on it? Explain your answer in detail.



Answer: k-means will not work well on this dataset. K-means works only if the clusters are round shaped. This data contains nonlinear structure that k means algorithm cannot capture.