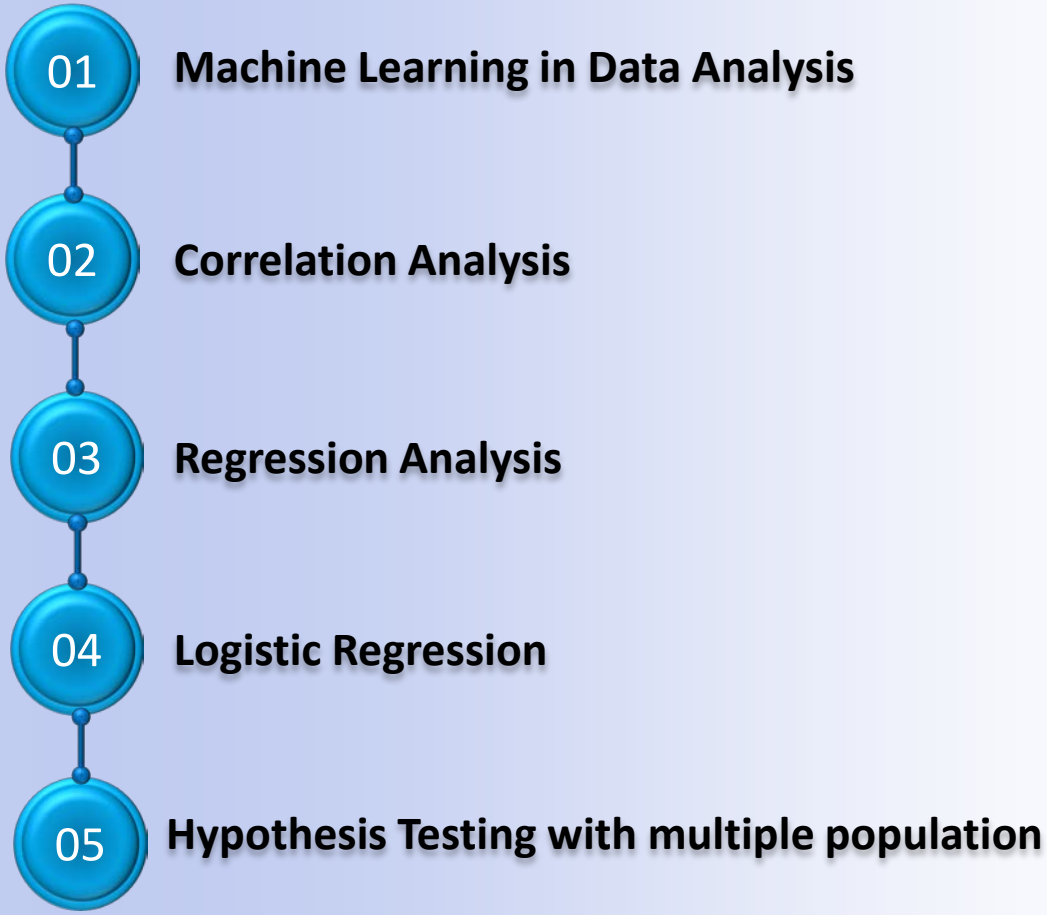
Data Analytics

Tutorial - III

Indian Institute of Technology Kharagpur

04th November 2023

Tutorial Overview



Machine Learning
Data Analytics

True False Questions

1. **Is there a difference between Statistics and Machine Learning?** Yes
2. **Is Logistic regression a supervised machine learning algorithm?** Yes
3. **Is Logistic regression mainly used for Regression?** No
4. **Is it possible to apply a logistic regression algorithm on a 3-class Classification problem?** Yes
5. **Maximum Likelihood cannot be used in Logistic Regression.** No
6. **We must calculate accuracy for evaluating linear regression model.** No
7. **The line in linear regression is called decision boundary.** No
8. **$(-\infty, \infty)$ is the range of logit function in the domain x=[0,1].** Yes
9. **As a data analyst the problem statements will be provided to you most of the time.** No
10. **Should we use machine learning if the data distribution in known in prior?** No

MCQs

Which of the following statement(s) is(are) NOT true?

a) Pearson's correlation analysis is applicable to only numeric data.
b) Spearman's correlation analysis is applicable to only ordinal data.
c) $\chi 2$ correlation analysis is applicable to only categorical data.
d) Any non-parametric statistical learning approach is applicable when the entire population is available.

Correct Answer: (b)

Explanation:

Spearman's correlation analysis is applicable to both ordinal and numerical data because in both the cases, the rank of data can be calculated.

Which of the following statement(s) is(are) correct?

a) The statistical t-test is to be followed to infer mean of a normally distribute population when the population variance is known.
b) The statistical z-test is to be followed to infer the mean of a normally distributed population when the population variance is unknown.
c) To infer the variance of normally distributed population the statistical $\chi 2$- test can be followed.
d) To infer the mean of normally distributed population the statistical $\chi 2$- test can be followed.

Correct Answer: (c)

Explanation:

$\chi 2$- test is used to compare a sample variance to a theoretical population variance.

Which of the following is(are) NOT under parametric test?

a) Z-test of hypothesis testing
b) t-test of hypothesis testing
c) $\chi 2$-test of hypothesis testing
d) Chi-square test of correlation analysis

Correct Answer: (d)

Explanation:

Chi-square correlation analysis is used to test if two attributes are independent to each other. Like hypothesis testing, here we are to test the null hypothesis that two attributes are independent. The other are to test the populations parameters and hence they are parametric test.

A value of $r^2 \approx 0$ means ?

a) SST is greater than SSE
b) SSE is greater than SST
c) SSE is almost equal to SST
d) There is no correlation

Correct Answer: (c) and (d)

Explanation:

In the context of regression analysis, it is R^2 and then SSE is slightly smaller than SST as R^2=1-SSE/SST. On the other hand, in the context of correlation analysis, r^2 is the square of degree of correlation coefficient, and then there is no correlation.

How many model parameters are to be learned when a simple non-linear regression model is to be built with a training set of size *n*?

a) 1
b) 2
c) 3
d) n

Correct Answer: (c)

Explanation:

It depends on non-linearity, that is, degree of the regression model. For a simple linear model, the parameters are b_0 and b_1, that is, two parameters. For a simple non-linear model, it should be at least 3, for example, $y = a + bx + cx^2$.

The value of coefficient of determination lies between?

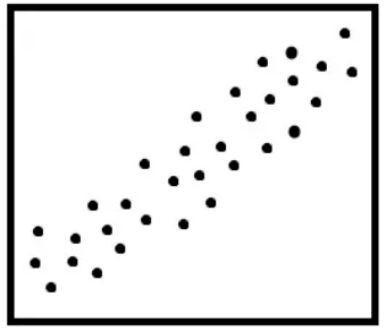
a) 0 to 1
b) -1 to 1
c) $-\infty$ to $+\infty$
d) 1 to ∞

Correct Answer: (a)

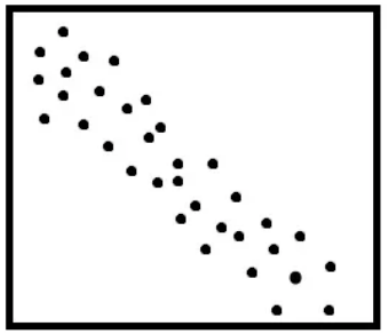
Explanation:

The coefficient of determination is r^2 where r is the degree of correlation and the values of r lie between -1 to 1.

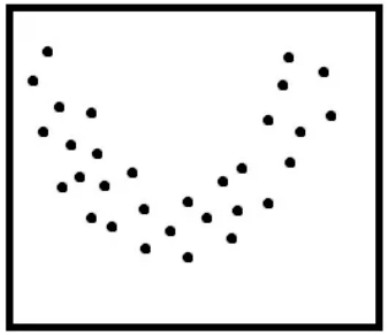
Which of the following data has non-linear association?



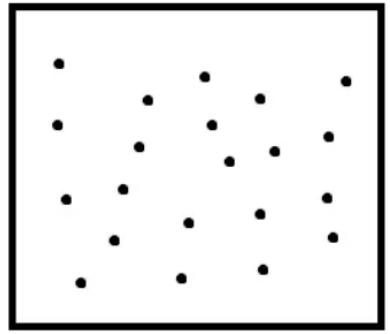
(a)



(b)



(c)



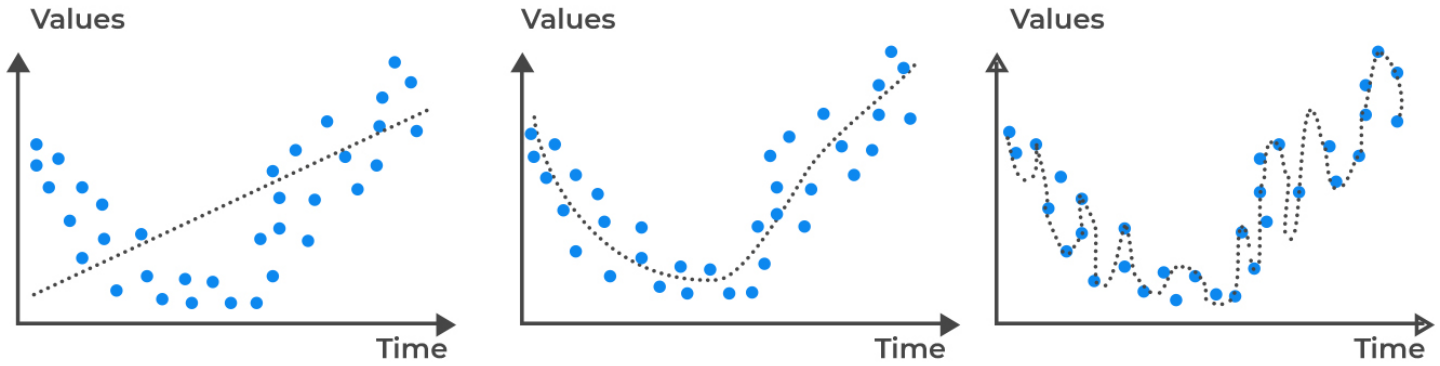
(d)

Correct Answer: **(c)**

Explanation:

(a) is positive linear association; (b) is negative linear association; (d) has no association.

Which of the following model is the most optimal model (*dotted line*)?



(a) **(b)** **(c)**

Correct Answer: (b)

Explanation:

(a) is underfitting – high bias; (b) is optimal; (c) is overfitting – high variance.

Which of the following metric(s) you will use for a regression problem?

a) RMSE
b) Accuracy
c) F1-Score
d) Recall

Correct Answer: (a)

Explanation:

(a) is distance based error metric which is the only metric that can be used in a regression task.

Which of the following test is suitable when the population variances are assumed equal?

a) Pooled t-test
b) Z-test
c) $\chi 2$-test
d) All are suitable

Correct Answer: (a)

Explanation:

The test that assumes equal population variances is referred to as the pooled t-test. Pooling refers to finding a weighted average of the two independent sample variances. The pooled test statistic uses a weighted average of the two sample variances.

Which of the following is incorrect about paired t-test?

a) It is used to compare two population means
b) One sample can be paired with observations in the other sample
c) Works well in case of extreme outliers
d) Can be used for comparison of two different methods of measurement

Correct Answer: (c)

Explanation:

For this test to be valid the differences only need to be approximately normally distributed. Therefore, it would not be advisable to use a paired *t-test* where there were any extreme outliers.

Example - A

Make a list of 10 food/household items purchased regularly by your family. Obtain the current prices of the items in three different shops.

Example - B

Draw a straight line between 20cm and 25cm in a paper. Collect 6 to 10 volunteers from school years 7, 9 and 11 to independently estimate the length.

Choose the correct option from below:

a) There is one factor (item) at three levels (3) in Example-A
b) There are two factors (item and shop) at two levels (10 and 3) in Example-A
c) There is one factor (school year) at three levels (7, 9 and 11) in Example-B
d) There is two factor (volunteers and school year) at one level (length) in Example-B

Correct Answer: (b) and (c)

Explanation:

Factor: A characteristic under consideration, thought to influence the measure observations.
Level: A value of the factor.

Practice Questions

Practice Question 1

A survey was conducted among 1500 people (300 literate and 1200 illiterate). In this survey, a person is either a literate or an illiterate and their participation in a poll (is either "cast" or "no cast") was recorded. It is observed that 250 literate people and 200 illiterate people casted their votes, respectively.

It is required to test (using χ^2 –correlation analysis) if there is any association between literacy and habit to casting vote of the electorate.

Find the following.

a) Give the structure of the contingency table suitable for χ^2 –test in this case.
b) Enter the observed frequencies and expected frequencies in the table.
c) Calculate the χ^2 value from the contingency table.
d) Mention the null hypothesis usually considered for χ^2 –test.
e) Test the hypothesis, if literacy is correlated with voting trend. Assume 0.01 is the confidence level.

a) The contingency table will take the following form:

		Literacy		
		Literate	Illiterate	Total
Casting habit	Yes			
	No			
	Total			1500

b) Observed frequencies and expected frequencies are shown in the table

		Literacy		
		Literate	Illiterate	Total
Casting habit	Yes	250 (90)	200 (360)	450
	No	50 (210)	1000 (840)	1050
	Total	300	1200	1500

Practice Question 1

A survey was conducted among 1500 people (300 literate and 1200 illiterate). In this survey, a person is either a literate or an illiterate and their participation in a poll (is either "cast" or "no cast") was recorded. It is observed that 250 literate people and 200 illiterate people casted their votes, respectively.

It is required to test (using χ^2 –correlation analysis) if there is any association between literacy and habit to casting vote of the electorate.

Find the following.

a) Give the structure of the contingency table suitable for χ^2 –test in this case.
b) Enter the observed frequencies and expected frequencies in the table.
c) Calculate the χ^2 value from the contingency table.
d) Mention the null hypothesis usually considered for χ^2 –test.
e) Test the hypothesis, if literacy is correlated with voting trend. Assume 0.01 is the confidence level.

c) Calculation of $\chi 2$ – value

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(200-360)^2}{360} + \frac{(50-210)^2}{210} + \frac{(1000-840)^2}{840}$$
$$= 507.93$$

d) The hypothesis for testing is

H_0: Literacy and voting habit is independent to each other.
H_1: Otherwise.

e) Testing the hypothesis

The degree of freedom for the data is (2-1) * (2-1) = 1

With 1 degree of freedom and $\alpha = 0.01$, the χ^2 value needed to reject the hypothesis is 10.828.
Since, the test statistic value of $\chi^2 = 507.93$ greater than critical χ^2–value, the null hypothesis is rejected.

That is, Literacy and casting habit is highly correlated.

Practice Question 2

Two attributes X and Y are related with probability mass function f(X,Y), which is as follows.

		Y		
	f(X,Y)	10	15	20
X	5	0.20	0.15	0.10
	10	0.05	0.15	0.15
	15	0.10	0.05	0.05

Calculate the following.

(a) μ_x, the mean of the sample X.
(b) μ_y, the mean of the sample Y.
(c) σ_x, the standard deviation in the sample X.
(d) σ_y, the standard deviation in the sample Y.
(e) S_{xy}, the variance between the sample X, Y
(f) The value of COV(X, Y).

a)

		Y			
	f(X,Y)	10	15	20	h(x)
X	5	0.20	0.15	0.10	0.45
	10	0.05	0.15	0.15	0.35
	15	0.10	0.05	0.05	0.20
	g(y)	0.35	0.35	0.30	1

b) $\mu_x = \sum_{x=5}^{15} xh(x) = 5*0.45 + 10*0.35 + 15*0.2 = 8.75$

c) $\sigma_x{}^2 = E(x^2) - \mu_x{}^2$

$E(x^2) = x^2 h(x)$

$= 5^2 * 0.45 + 10^2 * 0.35 + 15^2 * 0.2 = 91.25$

$\sigma_x{}^2 = 91.25 - 8.75^2 = 14.69$

$\sigma_x = \sqrt{14.69} = 3.83$

Practice Question 2

Two attributes X and Y are related with probability mass function f(X,Y), which is as follows.

		Y		
	f(X,Y)	**10**	**15**	**20**
X	**5**	**0.20**	**0.15**	**0.10**
	10	**0.05**	**0.15**	**0.15**
	15	**0.10**	**0.05**	**0.05**

Calculate the following.

(a) μ_x, the mean of the sample X.
(b) μ_y, the mean of the sample Y.
(c) σ_x, the standard deviation in the sample X.
(d) σ_y, the standard deviation in the sample Y.
(e) S_{xy}, the variance between the sample X, Y
(f) The value of COV(X, Y).

d) $\sigma_y^2 = E(y^2) - \mu_y^2$

$E(y^2) = y^2 g(y)$

$= 10^2 * 0.35 + 15^2 * 0.35 + 20^2 * 0.3 = 233.75$

$\sigma_y^2 = 233.75 - 14.75^2 = 16.19$

$\sigma_y = \sqrt{16.19} = 4.02$

e) $S_{xy} = \sum(x - \mu_x).(y - \mu_y).f(x,y)$

$\approx E(XY) - \mu_x.\mu_y$

$E(XY) = \sum_{x=5}^{15} \sum_{y=10}^{20} x.y.f(x,y)$

= 5*10*0.20 + 5*15*0.15 + 5*20*0.1 + 10*10*0.05 + 10*15*0.15 + 10*20*0.15 + 15*10*0.10 + 15*15*0.05 + 15*20*0.05

= 130

S_{xy} = 130 – 8.75*14.75 = 0.94

f) $COV(x,y) = \frac{S_{xy}}{\sigma_x.\sigma_y} = \frac{0.94}{3.83*4.02} = 0.06$

Practice Question 3

Suppose a sample of n students were given a diagnostic test before studying a particular module and then again after completing the module. We want to find out if, in general, our teaching leads to improvements in students' knowledge/skills (i.e. test scores). We can use the results from our sample of students to draw conclusions about the impact of this module in general.

Student	Pre-module score	Post-module score	Difference
1	18	22	+4
2	21	25	+4
3	16	17	+1
4	22	24	+2
5	19	16	-3
6	24	29	+5
7	17	20	+3
8	21	23	+2
9	23	19	-4
10	18	20	+2
11	14	15	+1
12	16	15	-1
13	16	18	+2
14	19	26	+7
15	18	18	0
16	20	24	+4
17	12	18	+6
18	22	25	+3
19	15	19	+4
20	17	16	-1

Let x = test score before the module, y = test score after the module

To test the null hypothesis that the true mean difference is zero, the procedure is as follows:

1. Calculate the difference ($d_i = y_i - x_i$) between the two observations on each pair, making sure you distinguish between positive and negative differences.
2. Calculate the mean difference, $\bar{d}$.
3. Calculate the standard deviation of the differences, s_d, and use this to calculate the standard error of the mean difference, $SE(\bar{d}) = \frac{s_d}{\sqrt{n}}$
4. Calculate the t-statistic, which is given by $T = \frac{\bar{d}}{SE(\bar{d})}$. Under the null hypothesis, this statistic follows a t-distribution with $n - 1$ degrees of freedom.
5. Use tables of the t-distribution to compare your value for T to the t_{n-1} distribution. This will give the p-value for the paired t-test.

Practice Question 3

Suppose a sample of n students were given a diagnostic test before studying a particular module and then again after completing the module. We want to find out if, in general, our teaching leads to improvements in students' knowledge/skills (i.e. test scores). We can use the results from our sample of students to draw conclusions about the impact of this module in general.

Student	Pre-module score	Post-module score	Difference
1	18	22	+4
2	21	25	+4
3	16	17	+1
4	22	24	+2
5	19	16	-3
6	24	29	+5
7	17	20	+3
8	21	23	+2
9	23	19	-4
10	18	20	+2
11	14	15	+1
12	16	15	-1
13	16	18	+2
14	19	26	+7
15	18	18	0
16	20	24	+4
17	12	18	+6
18	22	25	+3
19	15	19	+4
20	17	16	-1

$n = 20$ students.

Calculating the mean and standard deviation of the differences gives:

$\bar{d} = 2.05$ and $s_d = 2.837$. Therefore, $SE(\bar{d}) = \frac{s_d}{\sqrt{n}} = \frac{2.837}{\sqrt{20}} = 0.634$

So, we have:

$$t = \frac{2.05}{0.634} = 3.231 \qquad \text{on 19 df}$$

Looking this up in tables gives $p = 0.004$. Therefore, there is strong evidence that, on average, the module does lead to improvements.

Practice Question 4

The table below shows the lifetimes under controlled conditions, in hours in excess of 1000 hours of samples of 60W electric light bulbs of three different brands.

Table 2 (Q. No. 5)

Brand		
1	2	3
16	18	26
15	22	31
13	20	24
21	16	30
15	24	22

a) Identify the factor(s) and level(s) in the above-mentioned problem statement.
b) Calculate the sample mean and variance for each level.
c) Calculate the pooled estimate of variance.
d) Calculate the variability between samples.
e) Assuming all lifetime to be normally distributed with common variance, test, at the 1% significance level, the hypothesis that there is no difference between the brands with respect to mean lifetime.

a) Hence, there is one factor (i.e. brand) at three levels (i.e. 1, 2 and 3). There are three samples each of size 5.

b)

	Brand		
	1	2	3
Sample size	5	5	5
Sum	80	100	135
Sum of squares	1316	2040	3689
Mean	16	20	27
Variance	9	10	11

c) $\widehat{\sigma_w}^2 = \frac{(5-1)*9 + ((5-1)*10 + (5-1)*11}{(5-1)+(5-1)+(5-1)} = 10$

d)

	Brand		
	1	2	3
Sample mean	16	20	27
Sum		63	
Sum of squares		1385	
Mean		21	
Variance		31	

Practice Question 4

The table below shows the lifetimes under controlled conditions, in hours in excess of 1000 hours of samples of 60W electric light bulbs of three different brands.

Table 2 (Q. No. 5)

Brand		
1	2	3
16	18	26
15	22	31
13	20	24
21	16	30
15	24	22

a) Identify the factor(s) and level(s) in the above-mentioned problem statement.
b) Calculate the sample mean and variance for each level.
c) Calculate the pooled estimate of variance.
d) Calculate the variability between samples.
e) Assuming all lifetime to be normally distributed with common variance, test, at the 1% significance level, the hypothesis that there is no difference between the brands with respect to mean lifetime.

e)

$H_0 : \mu_1 = \mu_2 = \mu_3 = 20$
$H_1 : \mu 1 \neq \mu_2 \neq \mu_3 \neq 20$

Significance level, $\alpha = 0.01$
Degree of freedom $v_1 = 2$, $v_2 = 12$
$(k-1)$ $(n-k)$
Critical region is $F > 6.926$
Test statistics is $F > \frac{5.\widehat{\sigma_B}^2}{\widehat{\sigma_w}^2} = 155/10 = 15.5$
Since, the value lies in the critical region, thus, there is evidence that, at the 1% significance level, the true mean lifetime of three bulbs from three brands do differ.

Thank You