CS215 **Data Analytics and Interpretation** **Post-Midsem Quiz2**

2/11/23 **Time- 02.45 pm to 3.30 pm** **Marks: 20**

Solution

(There are 10 questions. For the questions, more than one answer-choice may be applicable, in which case all of the choices have to be given. You get full marks for ticking ALL correct options, else you get 0)

Q1. Consider the ML-technique spectrum of **Table-lookup (T)→rules (R)→Statistical-ML (S)→Deep-Learning (D)**. The two techniques that need data to train are:

(a) *T and R*
(b) *R and S*
(c) *S and D*
(d) *D and T*

Ans: (c)

Justification: both T and R are completely manual effort driven. Humans create tables and rules. S and D involve architectures and parameters which have to be fixed with data (for example, SVMs and Neural Nets). Hence (c).

Q2. Suppose an ML model has not been trained on adequate amount of DATA, as well as for adequate amount of time. The task for the model is to detect if a presented input pattern is the alphabet 'A' (output 1) or not (output 0). Now, two error cases were observed: (I) The model is presented with a hand written 'A' and it outputs 0; (II) the model is presented a 'B' and it outputs 1. Which of the following is/are true?

(a) I is a case of false NEGATIVE
(b) II is a case of false NEGATIVE
(c) I is a case of false POSITIVE
(d) II is a case of false POSITIVE

Ans: (a), (d)

Justification:

False negative occurs when the input is wrongly decided as NOT belonging to the concept; False positive occurs when the input is wrongly decided as belonging to the concept. Hence (a) and (d).

Q3. If $M_X(t)$ is the moment generating function (MGF) of a random variable X, then the MGF of (X+1) is

(a) Same as $M_X(t)$
(b) e^t times $M_X(t)$

(c) $M_X(t)/e^t$
(d) $\log_e M_X(t)$

Ans: (b)

Justification:

$M_{X+1}(t)=E(e^{t(X+1)})=e^t \cdot E(e^{tX})=e^t M_X(t)$, since e^t is constant when expectation is taken over X.

Q4. "Suppose X and Y are random variables having $M_X(t)$ and M(t) as moment generating functions respectively. Then X and Y have the same probability distributions if and only if $M_X(t)=M_Y(t)$ identically". This statement is called

(a) *Law of Large Numbers*
(b) *Uniqueness theorem*
(c) *Cauchy's theorem*
(d) *Gauss's theorem*

Ans: (b)

Justification: The statement given in the question is indeed the statement of the Uniqueness theorem.

Q5. A random variable *X* has the density function $f(x)=1/[c(x^2+1)]$, where *x* lies between minus and plus infinity. The value of '*c*' is

(a) 2
(b) 1/2
(c) π (i.e., pi)
(d) 2π

Ans: (c)

Justification: integration of f(x) from –inf to +inf is 1. But integration of $1/(x^2+1)$ is $\tan^{-1}(x)$. Hence $[\pi/2-(-\pi/2)].1/c=1$. So $c=\pi$.

Q6. Use the z-score data sheet below for answering this question:

Standard Normal Probabilities

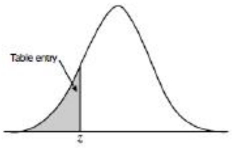


Table entry for z is the area under the standard normal curve to the left of z.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
−1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
−1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
−1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
−1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
−1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
−1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
−0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
−0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
−0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
−0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
−0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
−0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
−0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
−0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
−0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
−0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

Standard Normal Probabilities

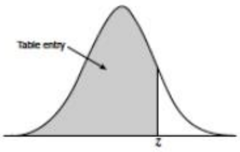


Table entry for z is the area under the standard normal curve to the left of z.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

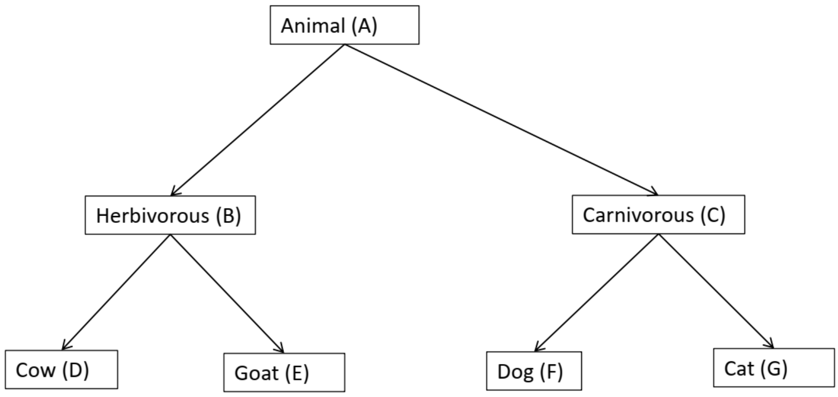
The probability P(0<=Z<=1.65)

(a) Lies in the range [0.4, 0.5]
(b) Lies in [0, 0.3]
(c) Is equal to 0.5-P(Z<=-1.65)
(d) None of the above

Ans: **(a), (c)**

Justification: consulting the data sheet, ϕ(Z=-1.65)= P(-inf<=Z<=1.65)=value under row 1.6 and col 0.05=0.0495; hence P(0<=Z<=1.65)=0.5-0.0495 which is the same as what is written under (a) and (c).

Q7. Consider the “animal hierarchy tree” given below:



We want to find something called the “probability of the tree” which has many applications in data science. By convention the probability of the root is 1. The probability of a tree is the joint probability of the nodes in the tree. It is assumed that the probability of a node depends only on its IMMEDIATE PARENT AND NOTHING ELSE. Then the probability of the animal tree is

(a) P(A,B).P(A,C). P(D|B).P(E|A).P(F,G|C)
(b) P(A,B).P(A,C). P(D|B).P(E|B).P(F|C). P(G|C)
(c) P(A,B).P(A,C). P(D,B).P(E,A).P(F,G|C)
(d) None of the above

Ans: (b)

Justification: Given P(A)=1.

Required probability,
P(.)=P(A,B,C,D,E,F,G)=P(A).P(B|A).P(C|A,B).P(D|A,B,C).P(E|A,B,C,D).P(G|A,B,C,D,E,F)

Because of independence assumption, P(.)=P(A).P(B|A).P(C|A).P(D|B).P(E|B).P(F|C).P(G|C)

=P(A,B).P(A).P(C|A).P(D|B).P(E|B).P(F|C).P(G|C)

=P(A,B).P(A,C). P(D|B).P(E|B).P(F|C).P(G|C)= expression in (b)

Q8. Assume that the speeds at which cars pass through a checkpoint are normally distributed with mean 61 units and standard deviation 4 units. Then the expected number of cars out of 500, passing through the checkpoint with speed more than 66 units is

(a) in the range 30 to 40
(b) close to 50
(c) less than 20
(d) greater than 65

Ans: (b)

Justification:

For X (=speed)=66, Z=(66-61)/4=1.25, from the data sheet in Q6, $\phi(1.25)=0.89$ (approx.). Therefore, P(Z>1.25)=1-0.89=0.11. Hence the required expected number of cars=500 X 0.11= 55 (approx); hence (a) and (b) are the correct answers.

Q9. Suppose the level of significance is fixed at α (alpha). Given a sample of size *36* and mean *30*, and the standard deviation of population as 3, the population mean μ satisfies which of the following statement(s)?

(a) With probability α, μ lies between $30-1.5z_\alpha$ and $30+1.5z_\alpha$
(b) With probability $1-\alpha$, μ lies between $30-0.5z_{\alpha/2}$ and $30+0.5z_{\alpha/2}$
(c) With probability α, μ lies between $30-0.5z_{\alpha/2}$ and $30+0.5z_{\alpha/2}$
(d) None of the above

Ans: (b)

Justification:

Sample mean $X_m=30$, $\sigma=3$, n=36. Hence standard error= σ/sqrt(n)=3/6=0.5. Since the significance level is α, the confidence interval is from $-z_{\alpha/2}$ to $+z_{\alpha/2}$

$$P(-z_{\alpha/2}<=(X_m-\mu)/(\sigma/\mathrm{sqrt}(n))<= +z_{\alpha/2})=1-\alpha$$

$$i.e., P(-z_{\alpha/2}<=(30-\mu)/0.5<= +z_{\alpha/2})=1-\alpha$$

Hence, (b) is correct.

Q10. Use the following table for this question:

test-type (col)			
vs.	Two-Tail	1 sided to +inf	1 sided from -inf
Confidence Interval (significance level)			
90% (0.10)	(- and +) 1.65	-1.28 to +inf	-inf to +1.28
95% (0.05)	(- and +) 1.96	-1.65 to +inf	-inf to +1.65
99% (0.01)	(- and +) 2.58	-2.33 to +inf	-inf to 2.33

The mean height of 50 male students who showed above average participation in college athletics was 173.3 cm with a standard deviation of 6.4 cm, while 50 male students who showed no interest in such participation had a mean height of 171.5 cm with a standard deviation of 7.1cm. z_o is 1.33. It is required to examine if male students who participate in college athletics are taller than other male students. Choose the correct option(s).

(a) Have to accept the alternative hypothesis at 90% confidence interval
(b) Have to accept the null hypothesis at 95% confidence interval
(c) Have to accept the null hypothesis at 99% confidence interval
(d) None of the above

Ans: (a), (b), (c)

Justification:

Let the mean height of athletics participants' group be M_1 and that of the other group be M_2.

H_0: M_1-M_2=0, H_A: M_1>M_2

Based on this and the standard deviations of two groups, Zo is given as 1.33.

H_0 will be rejected and H_A accepted at significance level α if Zo falls outside the range –inf to Zα.

Since 1.33>1.28, but <1.65 and 2.33, (a),(b), (c) hold.

===================end=================