Data Science (CSE558)

End-Sem

Time: 2 Hours **Maximum Marks: 50**

Instructions

1. **Please read all questions carefully before writing your answers. The 1-mark questions need to answered with a single word. An incorrect answer to these questions fetches negative marks.**
2. **Please write neat, short and meaningful answers. Untidy answers will not be evaluated.**
3. **All parts of a single question should be answered collectively, i.e., all at one place.**

All The Best! .

1. **Consider a weighted coin such that probability of getting head is 0.2.**
   a. **What is the probability of getting at least 40 heads from 100 random tosses of this coin? (3 Marks)**
   b. **Use this coin to define a random variable that can be used to design a family of random matrices of size k x d such that a matrix randomly sampled from the family ensure preservation of distance between 2^k pairs of points. (5 Marks)**
2. **Let A and B be two events in a sample space such that Pr(A) = 6/11 and Pr(B) = 6/13.**
   a. **Can the events be disjoint? (+1/-0.5 Marks)**
   b. **Prove or disprove your answer. (2 Marks)**
3. **Recall the hypothesis tests for population means, i.e., z-test and t-test. For a sample of size n, after computing the test statistics why degree of freedom is taken into account in t-test whereas not in z-test? (2 Marks)**
4. **Consider a bag with four unbiased dice as {d1, d2, d3, d4}. For every 1 ≤ i ≤ 4, di has i-faces numbered 1 to i.**
   a. **Let one die be selected uniformly at random from the bag and it has been rolled 3 times. The outcome is {1,3,1}. Use maximum likelihood estimation to determine the most likely dice selected from the bag. (2 Marks)**
   b. **Use Bayes' rule to compute the probability of the most probable die for the given outcome after every roll. Compute Pr(die|{1}) after 1st roll, compute Pr(die|{3}) after 2nd roll and so on. (4 Marks)**
   c. **Let a die be selected from the bag to play snakes and ladder between three friends. Each player can start her game only when she rolls 1. Let player-1, Player-2, and Player-3 take 3, 7, and 2 rolls to start their respective games. Using maximum likelihood estimation determine, the most likely die selected from the bag. (3 Marks)**
5. **A random point v on the surface of a unit sphere in a d-dimensional space, centered at origin is generated as follow: for every i ϵ [1, …, d], v(i) is either +($d^{-0.5}$) or -($d^{-0.5}$) with equal probability. Let x and y be two such points.**
   a. **Prove that as d tends to infinity, the points tend to be orthogonal. (2 Marks)**
   b. **For a fixed d, use Chebyshev's to bound the probability of the event $|x^T y| \geq 0.1$. (3 Marks)**

c. **For a fixed d, use Bernstein to prove that there are at most $O(2^{0.01d})$ points such the following event E is true with probability at least 0.9.**

E: For all pair of points (x, y) from the set, $|x^T y| < 0.1$. (5 Marks)

6. **Let A be a matrix of size n x m such that its rank is 10, where 10 < min(n,m). Let $A = U\Sigma V^T$ where U and V are two orthonormal square matrices of size n x n and m x m respectively.**
   a. **How many non-zero singular values are there in Σ? (+1/-0.5 Marks)**
   b. **Let, k < 10 and $Z = [U_{11-k}, U_{12-k}, \ldots, U_9, U_{10}]$ consisting of k singular vectors from U. $B = ZZ^T A$ be the low rank representation of A. Then calculate the exact Frobenius norm of the difference matrix between the original data and the projected data i.e., $\|A - B\|_F^2$. (3 Marks)**
7. **Continuing the above question, consider another matrix B of size m x n. The running time of AB is $O(n^2 m)$, which worries you because it has a quadratic dependence on n. So, your friend proposed the following steps to improve the running time of the matrix-matrix product.**

   - Sample a random vector 'q' in R^m such that, every index of q is -1 or +1 with equal probability.

   - Compute C = Aq and $D = q^T B$.

   - Return X = CD.

   a. **What is the running time of the above algorithm? (+1/-0.5 Marks)**
   b. **You noticed that, $E[qq^T] = I$ (m x m identity matrix). Hence, E[X] = AB and it is an unbiased estimator. If the rank(A) = rank(B) = rank(AB) = 10 then what is the rank of X? Why? (2 Marks)**
8. **Let A be a matrix of size n x m. Let C be a set of 2k columns in A and R be set of k rows in A. With the selected C and R design an efficient algorithm to compute U for the CUR decomposition. State its running time? (5 Marks)**
9. **Let palyer1 and player2 are playing for car1 and car2 in two different Monty Hall game separately. Let player1 did not change her preference upon given a choice and she lost the car1. In a separate game let player2 decides to change her preference upon given the option. With this strategy is the player2 guaranteed to win the car2 in her game? (+1/-0.5 Marks)**
10. Let H_m be a n x n Hadamard matrix, where n = 2^m for some positive integer m. Let $H_m = \begin{pmatrix} H_{m-1} & H_{m-1} \\ H_{m-1} & -H_{m-1} \end{pmatrix}$ such that H_0 = [1]**. Let v be a n-dimensional vector. Write an efficient pseudocode for the matrix vector product** $H_m v$. What is its running time? **(5 Marks)**

11. **(Bonus) Recall that sparsity in JL matrix has an advantage in the running time. Is this always useful? Explain your reason with an analysis. (3 Marks)**

Rubric

Q1.

Q1 (a) Random Variable r

$$r = \begin{cases} 1 & \text{we assign } 0.2 \\ 0 & \text{else} \end{cases}$$

$$\left.\begin{aligned} 100 \cdot E[r] &= 100 \cdot (1 \times 0.2 + 0 \times 0.3) \\ &= 20 \text{ expected heads.} \end{aligned}\right\} \text{0.5 Mark}$$

$$R = \sum_{i=1}^{100} r \qquad \boxed{E[R] = 20}$$

Markov — we look at 40

$$Pr(R \geq 40) \leq \frac{E[R]}{40} = \boxed{\frac{1}{2}}$$

Markov → 1 Mark

$$\text{Variance}(R) = \sum_{i=1}^{100} Var(r_i) = 100 \cdot Var(r)$$

$$= 100\left[E[r^2] - [E[r]]^2\right]$$

$$\boxed{= 100\,[0.2 - 0.04] = 16}$$ 0.5 Mark

0.5 Chebyshev

$$\boxed{Pr(|R - E[R]| \geq 20)} \leq \frac{Var(R)}{(20)^2} = \frac{16}{20 \cdot 20} = \boxed{\frac{1}{25}}$$

$R \geq 40$

0.5 Mark

$\#$ of heads ≥ 40

Prob $\leq \frac{1}{25}$ Full Marks

OR

$$\left.\sum_{i=40}^{100} \binom{100}{i} \cdot 0.2^{i}\, 0.8^{60}\right\} \text{0.5 Mark}$$

show

(b) $\boxed{E[r]=0}$ → 1 Mark

$\boxed{Var(r) = 1/K}$ → 2 Marks

Unbiased coin

$$\boxed{\begin{array}{llll} HT \to \tilde{H} & \to \text{Prob} & 0.16 \\ TH \to \tilde{T} & \to \text{Prob} & 0.16 \end{array}} \leftarrow \Omega$$

new sample space

$$\times \begin{cases} TT & \to \;\; '' \;\; 0.64 \\ HH & \quad\quad\; 0.04 \end{cases}$$

$$\left. \begin{array}{l} Pr_{\Omega}(\tilde{H}) = \dfrac{0.16}{0.32} = 0.5 \\ Pr_{\Omega}(\tilde{T}) = 0.5 \end{array} \right\} \text{1 Mark}$$

$$r = \left. \begin{cases} \dfrac{1}{\sqrt{K}} & \text{if } \tilde{H} \;\; 0.5 \\ \dfrac{-1}{\sqrt{K}} & \text{if } \tilde{T} \end{cases} \right\} \text{1 Mark}$$

Prove:-

$\boxed{E[r]=0}$

$$Var(r) = \frac{1}{K}(0.16) + \frac{1}{K}(0.16) - (E[r])^2$$

$$0$$

we need

$$\frac{0.5}{K} + \frac{0.5}{K}$$

Prove:-

$$\boxed{Var = \frac{1}{K}}$$

Que.3

1.For Z-test - No degree of freedom because population variance is known and fixed.

2.For T-test - DoF is critical because the test relies on sample variance and smaller sample size increases uncertainty. **2 Marks if the above points are mentioned in the answer.**

Que.7

Time complexity = O(n^2 + nm) —--- **1 M / -0.5M**

Rank of X = 1 —---- **1 M**

Reason(C, D only 1,1 vectors) —----**1 M**

Que.8

SVD method :

Running time —- **1M** Running time = O(K^3)

ntersection —-- **2M**

Pseudo Inverse — **2M**

Other method :

Calculation of U —- **1M**

C+, R+ calculation —-- **2M**

Efficient Algorithm —--- **1M**

Time Complexity —--- **1M** Running time = O(k^2n + kmn + k^2m)

Marks awarded for both methods

Que.9

Answer : **No, 1 Mark for correct, -0.5 for incorrect.**

Que-2

Q2.

a) No

b) $P(A) = \frac{6}{11} = \frac{6 \times 13}{11 \times 13} = \frac{78}{143}$

$P(B) = \frac{6}{13} = \frac{6 \times 11}{13 \times 11} = \frac{66}{143}$

If disjoint then A & B cannot have a common sample --- (i)

If (i) true then $P(A) + P(B) \leq 1$ --- (ii)

$P(A) + P(B) = \frac{144}{143} > 1$

$\therefore$ Contradiction

Que.4

Q4/ Given :-

$d_1 = \{1\}$

$d_2 = \{1, 2\}$

$d_3 = \{1, 2, 3\}$

$d_4 = \{1, 2, 3, 4\}$

- Each die ~~has~~ is equally likely to choosen from the bag initially, so the prior probability of selecting

(a) Maximum Likelihood Estimation for outcome $\{1, 3, 1\}$

d_1 :- $P(\{1,3,1\} \mid d_1) = 1 \times 0 \times 1 = 0$

d_2 :- $P(\{1,3,1\} \mid d_2) = \frac{1}{2} \times 0 \times \frac{1}{2} = 0$

d_3 :- $P\{1,3,1\} \mid d_3) = (\frac{1}{3})(\frac{1}{3})(\frac{1}{3}) = \frac{1}{27}$

d_4 :- $P\{1,3,1\} \mid d_4) = (\frac{1}{4})(\frac{1}{4})(\frac{1}{4}) = \frac{1}{64}$

$\boxed{\text{Ans:- } d_3}$

- Ans = d_3 (involves conditional probability) ~~2 Marks~~ – 2 Marks.

(or)

(if conditional probability

(i) Geometric distribution formula is used and correct answer — 3 Marks

(or)

(ii) eq^n (1) is written and substitution of $\theta = 1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}$ is done — 3 Marks.

(or)

(iii) directly 'd_4' is written as answer — 1 Mark

~~(or)~~

(or)

(iv) if $\theta = \frac{n}{\sum x_i} = \frac{3}{12} = \frac{1}{4}$ is computed only ———— 2.5 Marks.

$$= \frac{1 \cdot 1/4}{\left(1\cdot\cancel{\frac{1}{4}}\right)+\left(\frac{1}{2}\cdot\cancel{\frac{1}{4}}\right)+\left(\frac{1}{3}\cdot\cancel{\frac{1}{4}}\right)+\left(\frac{1}{4}\cdot\cancel{\frac{1}{4}}\right)}$$

$$P(d_1|\{1\}) = \frac{1}{1+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}} = 0.48 \quad , \quad P(d_2|\{1\}) = \frac{1/2}{1+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}} = 0.24$$

$$P(d_3|\{1\}) = \frac{1/3}{1+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}} = 0.16 \quad , \quad P(d_4|\{1\}) = \frac{1/4}{1+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}} = 0.12$$

After the first roll (1), the posterior probablities are:-

- $P(d_1|\{1\}) = 0.48$
- $P(d_2|\{1\}) = 0.24$
- $P(d_3|\{1\}) = 0.16$
- $P(d_4|\{1\}) = 0.12$

After the second Roll (Outcome = 3)

~~$P(d_1|\{1,3\}) =$~~

$$P(d_1|3) = \frac{P(3|d_1)\cdot P(d_1)}{P(3)} = \frac{0\cdot\left(\frac{1}{1+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}}\right)}{(\quad)}$$

$$\boxed{P(d_1|3) = 0}$$

$$P(d_2|3) = 0 \quad (\text{since } P(3|d_2) = 0)$$

$$P(d_3|3) = \frac{P(3|d_3)\cdot P(d_3)}{P(3)}$$

$$= \frac{\frac{1}{3}\cdot\frac{1/3}{1+1/2+1/3+1/4}}{\frac{1}{3}\cdot\left(\frac{1}{3}\right)\Big/1+\frac{1}{2}+\frac{1}{3}+\frac{1}{4} + \frac{1}{4}\cdot\frac{1/4}{1+1/2+\frac{1}{3}+\frac{1}{4}}}$$

$$P(d_3|3) = 16/25$$

$$P(d_4|3) = \frac{P(3|d_4)\cdot P(d_4)}{P(3)}$$

$$= \frac{\frac{1}{4}\cdot\frac{1/4}{1+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}}}{\frac{1}{3}\cdot\left(\frac{1}{3}\right)\Big/1+\frac{1}{2}+\frac{1}{3}+\frac{1}{4} + \frac{1}{4}\cdot\frac{1/4}{1+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}}}$$

$$P(d_4|4) = 9/25$$

After the second roll (3), the posterior probablities are:-

- $P(d_1|\{1,3\}) = 0$
- $P(d_2|\{1,3\}) = 0$
- $P(d_3|\{1,3\}) = 16/25$
- $P(d_4|\{1,3\}) = 9/25$

$\cancel{P(d_3|1)} =$

After the third roll (outcome = 1) :-

$$P(d_3|1) = \frac{P(1|d_3)\cdot P(d_3)}{\quad}$$

$$= \frac{\frac{1}{3}\cdot 16/25}{\frac{1}{3}\cdot\frac{16}{25}+\frac{1}{4}\cdot\frac{9}{25}} = \frac{64}{91}$$

$$P(d_4|1) = \frac{\frac{1}{4}\cdot 9/25}{\frac{1}{3}\cdot\frac{16}{25}+\frac{1}{4}\cdot\frac{9}{25}} = \frac{27}{91}$$

After the third roll (1), the posterior probablities are:

- $P(d_3|\{1,3,1\}) = 64/91$
- $P(d_4|\{1,3,1\}) = 27/91$

(i) all validations are shown + correct answer — 4 Marks

(or)

(ii) only d3 calculation is done completely — 1 Mark

(or)

(iii) Iteration 1 only — 1 mark

" (1+2) — 2 marks

" (1+2+3) — 4 marks + (updated probablities)

Iteration (1 and 2) — 3 marks + (updated probablities)

<u>Que.5</u>

Q5. $i \in [1, 2, \cdots, d]$

$v_i \in \{ \frac{1}{\sqrt{d}}, \frac{-1}{\sqrt{d}} \}$

$x, y \in \mathbb{R}^d$

(a)

$$E[x^T y] = E[\sum_{i=1}^{d} x_i y_i] = \sum_{i=1}^{d} E[x_i y_i]$$

Since x & y are independent. 0.5

$$E[x_i y_i] = E[x_i] . E[y_i]$$

$$x_i, y_i = \begin{cases} 1/\sqrt{d} & p = 1/2 \\ -1/\sqrt{d} & 1/2 \end{cases} \Rightarrow E[x_i] = E[y_i] = 0$$ 0.5

$$\Rightarrow E[x^T y] = \sum_{i=1}^{d} E[x_i y_i] = \sum_{i=1}^{d} E[x_i] E[y_i] = 0$$ 0.5

$$Var[x^T y] = Var(\sum_{i=1}^{d} x_i y_i) = \sum_{i=1}^{d} Var(x_i y_i)$$

$$= \sum_{i=1}^{d} \left(E[(x_i y_i)^2] - (E[x_i y_i])^2 \right)$$ 1

$$= \sum_{i=1}^{d} \left(E[x_i^2] . E[y_i^2] - 0 \right)$$

$$= \sum_{i=1}^{d} 1/d \times \frac{1}{d} = 1/d \qquad (E(x_i^2) = 1/d)$$

as $d \to \infty$ $\quad Var[x^T y] = \lim_{d \to \infty} 1/d \to 0$

$Var[x^T y] \to 0$ implies $x^T y$ approaches one point value

and as $E[x^T y] = 0$, This point is $\cancel{\geq}$ 0. 0.5

$\Rightarrow x^T y \to 0$ for $d \to \infty$ i.e x, y are orthogonal

(b) Let Z be the R.v for representing $x^T \cdot y$

$\mu = E[x^T \cdot y] = 0$ } 0.5

$Var(Z) = Var(x^T y) = 1/d = \sigma^2$

Chebyshev's Inequality :-

$P(|Z - \mu| \geq k\sigma) \leq \frac{1}{k^2}$ } 1

$\Rightarrow P(|Z - 0| \geq 0.1) \leq \frac{\sigma^2}{(0.1)^2} = \frac{1}{d(0.1)^2}$

$= \frac{100}{d}$ } 0.5 $\quad \because \left(k = \frac{0.1}{\sigma}\right)$

(C) Bernstein's Inequality:

$$P(|S - E[S]| \geq t) \leq \exp\left(-\frac{t^2}{\sigma^2 + bt}\right) \Big\} 0.5$$

$$X_i = x_i y_i = \begin{cases} 1/d & p = 1/2 \\ -1/d & p = 1/2 \end{cases} \Big\} 0.5$$

$$S = \sum_{i=1}^{d} X_i, \quad |X_i - E[X_i]| = |x_i y_i - 0|$$

$$E[S] = E[x^T y] = 0 \qquad = |x_i y_i| = \frac{1}{d}$$

$$\Rightarrow b = 1/d \Big\} 0.5$$

$$\sigma^2 = Var(X_i) = Var(x^T y) = 1/d \Big\} 0.5$$

$\cancel{|x^T y|} \geq, \quad t = 0.1 \Big\} 0.5$

Using the inequality

$$\Rightarrow P(|S - 0| \geq 0.1) \leq \exp\left(-\frac{(0.1)^2}{1/d + \frac{0.1}{d}}\right) \Big\} 0.5$$

$$\Rightarrow P\left(\left|\sum_{i=1}^{d} x_i y_i\right| \geq 0.1\right) \leq \exp\left(-\frac{d}{110}\right)$$

$$\Rightarrow P(|x^T y| \geq 0.1) = 1 - P(|x^T y| < 0.1) \Big\} 0.5$$

$$\leq \exp\left(-\frac{d}{110}\right)$$

$$\Rightarrow P(|x^T y| < 0.1) \geq 1 - \exp\left(-\frac{d}{110}\right)$$

If there are m points which form the set,

no. of pairs $(x, y) = \frac{m(m-1)}{2}$

1

$$P\left(P\left(|x^T y| < 0.1\right) \text{ for all pairs}\right)$$

$$= \left(1 - e^{d/110}\right)^{\frac{m(m-1)}{2}} \geq 0.9$$

as $e^{-d/110} \to \ll 1$

$$\Rightarrow 1 - \frac{m(m-1)}{2} e^{-d/110} \geq 0.9$$

$$\Rightarrow m(m-1)e^{-d/110} \leq \frac{(0.1)2}{\cancel{m(m}}$$

$$\Rightarrow m(m-1) \leq (0.2)\, e^{d/110}$$

0.5

$$\Rightarrow m \leq O\left(e^{d/220}\right) \leq O\left(2^{d/100}\right)$$

Que.6

6.a: **10**, **1 Mark for correct, -0.5 for incorrect.**

6.b: **1 Mark** for expression B, **2 marks** for exact frobenius norm.

ZZ^T is projection Matrix which is spanned by 'Z', $Z = [U_{11-k}, U_{12-k}, \ldots, U_{10}]$.

and, All the vectors in 'U' matrix is Orthonor-mal i.e. every pair wise vector is zero.

that, effectively

$$B = \sigma_8 u_8 v_8^T + \sigma_9 u_9 v_9^T + \sigma_{10} u_{10} v_{10}^T \text{ — 1 Mark}$$

Assuming $k=3$.

After $\|A-B\|_F^2 =$

$$\left\| \sum_{i=1}^{10} \sigma_i u_i v_i^T - \sum_{j=8}^{10} \sigma_j v_j^T u_j \right\|_F^2$$

$\because$ v_i v_j are orthonormal from 10 to 8 get Cancel out to zero

$$\therefore \left\| \sum_{i=1}^{7} \sigma_i u_i v_i^T \right\|_F^2$$

$$\Rightarrow \sum_{i=1}^{7} \sigma_i^2 \longrightarrow \text{2 Mark.}$$

<u>Que-10</u>

Q.10

Base case: $H_0 = [1]$

$H_1 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$ and so on

for a vector $\bar{v}_{n \times 1} \to \begin{bmatrix} \bar{v}_{1\, \frac{n}{2} \times 1} \\ \bar{v}_{2\, \frac{n}{2} \times 1} \end{bmatrix}$ $\forall\ n > 1$

this split is always true as $n = 2^m$ & $m \geq 0$

until m = 0 where $\bar{v} = [v_0]$

for a general $v_{n \times 1} = \begin{bmatrix} \bar{v}_1 \\ \bar{v}_2 \end{bmatrix}$ where $n = 2^m$

& a ~~general~~

$H_m = \begin{bmatrix} H_{m-1} & H_{m-1} \\ H_{m-1} & -H_{m-1} \end{bmatrix}$

$$H_m v = \begin{bmatrix} H_{m-1} & H_{m-1} \\ H_{m-1} & -H_{m-1} \end{bmatrix} \begin{bmatrix} \bar{v}_1 \\ \bar{v}_2 \end{bmatrix} = \begin{bmatrix} H_{m-1} v_1 + H_{m-1} v_2 \\ H_{m-1} v_1 - H_{m-1} v_2 \end{bmatrix}$$

~~Solve each of the 2~~

→ Divide each of the 2 terms further until H_0 & ~~any~~ v length = 1 (log n operations)

→ Perform addition backwards from m = 0 to m = 2^n (n operations on merge step/conquer step)

* Idea similar to merge sort & can cache at intermediate steps for future calculations

Time Complexity: nLog(n)

Note: In Q10. of the endsem. If you have used a recursive algorithm. Partial marks have been awarded for correct divide and merge steps. And another mark for the correct time complexity mentioned.There are no marks for a rank approximation or any other randomized prediction of the result, an exact computation was required.