answer

1	2	3	4	5	6	7	8	9	10	11	12	13	14	Total

CS 215: Data Interpretation and Analysis 2024, End-Semester exam

Roll: ______________________________

November 14, 2024.
1:30–4:30 pm

Name: ______________________________

Write all your answers in the space provided. Do not spend time/space giving irrelevant details or details not asked for. Use the marks as a guideline for the amount of time you should spend on a question. The exam is close book. You are only allowed to refer to five pages of hand-written notes.

1. Annie thinks she just discovered a very efficient algorithm for solving a challenging graph theory task. Her advisor asked her to run her algorithm 20 times and a baseline algorithm 30 times, and record the time taken. Let $A_1 \ldots, A_{20}$ denote the recorded times of Annie's algorithms and $B_1, \ldots, B_{30}$ be the recorded times of the baseline. You need to suggest a test statistic to establish if Annie's algorithm is faster than the baseline under the assumption that both running times follow a Gaussian distribution with possibly different means but the same variance. [Running time cannot be negative! But we will ignore this aberration for now.]

   (a) Write down the null and alternative hypothesis for this test. ..1

   $H_0 : \mu_A = \mu_B, H_1 : \mu_A < \mu_B$. The null hypothesis can also be written as $H_0 : \mu_A \geq \mu_B$.

   (b) Write down the test statistic in terms of A_is and B_js, and the condition under which you will accept Annie's claim that her algorithm is faster than the baseline. Assume 10% significance level. ..3

   $M_A = \sum_i A_i/20, M_B = \sum_i B_i/30, S = (\sum_i (A_i - M_A)^2 + \sum_j (B_j - M_N)^2)/48$. $T = (M_A - M_B)/\sqrt{S(1/20 + 1/30)}$.

   $P_{H_0}(T) \sim t_{48}$ follows a t-distribution with 48 degrees of freedom. We will accept Annie's claim if observed T is less than $-t_{0.1,48}$ where $t_{0.1,48}$ is the upper 10 percentile of the t-distribution with 48 df.

   (c) However, Annie suspects that possibly p% of her timings are abnormally high because of the server facing an unusually high load. When $0 < p < 0.2$, suggest a different test statistic that is less affected by a few abnormal measurements. Make sure that the test statistic you suggest is both robust to abnormal values, while also being as precise as possible at normal data. ..4

   We make use of the information that at most p% of the observations may be unusually high — that is, outliers are one-sided. Instead of sample mean use winsorized or trimmed mean where only the top p% largest running times are dropped from both the A and B observations.

   Note: If anyone mentioned trimmed or winsorized mean with dropping both largest and smallest values, the credit should be reduced since that estimate is not as efficient. Median would be even less efficient than these.

   Instead of computing the shared variance on the entire data we compute it on the subset after removing the p% largest values in each set.

2. A very large number of students took a challenging test. You do not know the distribution of marks in the test. Your friend Aditya, who obtained 70 absolute marks in the test, claims

3. What is the breakdown value of the test statistic used in the signed rank hypothesis test of a distribution being symmetric about a median value m_0 given n data samples $X_1, \ldots, X_n$? Recall that the break value of a statistic is defined as the largest number m such that if we replace $m-1$ values in the data by arbitrary values, the statistic still remains within a bounded set. ..2

   The test statistic used in signed rank is always bounded irrespective of the values. So, breakdown value is n.

4. Consider a two-dimensional data distribution with mean μ at $\binom{0}{0}$ and covariance $\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$ where $\rho > 0$.

   (a) Write the Mahalonobis distance from μ of a point $p = \binom{p_1}{p_2}$ when $\rho = 0.5$. ..1

   Simple application of the formula yields $\sqrt{4/3(p_1^2 + p^2 - p_1p_2)}$

   Rubrics: 1 mark for correct expression. 0.5 for minor error (no square root, error in constant factor etc.)

   (b) Among the set of points $S : x_1^2+x_2^2 = 1$, identify the subset whose Mahalonobis distance from μ is the largest, and the set from where it is smallest. Interpret your answer. ..3

   The Mahalonobis distance from mean can be written as $(x_1^2 + x_2^2 - 2\rho x_1x_2)/(2-2\rho^2)$. If we restrict to points in S, the minimum will be when $x_1 = x_2 = 1/\sqrt{2}$ or $x_1 = x_2 = -1/\sqrt{2}$ and maximum when $x_1 = 1/\sqrt{2}, x_2 = -1/\sqrt{2}$ or vice-versa. Among all points in S, the point $[1/\sqrt{2}, 1/\sqrt{2}]$ is most expected given the positive correlation, and the point $[1/\sqrt{2}, -1/\sqrt{2}]$ expresses negative correlation and is least expected.

   Rubrics: 1 mark for maxima. 1 mark for minima. 1 mark for correct interpretation - either in terms of correlation, variance of data sampled etc.

5. Consider a p dimension Normal distribution $X \sim \mathcal{N}(\mu, \Sigma)$ where μ is of size $p \times 1$ and Σ is of size $p \times p$. Assume $p = 2$.

   (a) Let $Z_1 = X_1 + X_2$, $Z_2 = X_1 - X_2$ $Z_3 = X_1$ What is the joint distribution of $Z = [Z_1, Z_2, Z_3]$? Provide both the form of the distribution and the parameters in terms of parameters of X. ..4

   Gaussian distribution

   For $Z_i = \mathbf{a}_i'X$ where $\mathbf{a}_1 = [1,1]', \mathbf{a}_2 = [1,-1]', \mathbf{a}_3 = [1,0]'$. With these we can define $Cov(Z_i, Z_j) = \mathbf{a}_i\Sigma\mathbf{a}_j$, and define mean of each Z_i as $\mathbf{a}_iX$

   Rubrics: 2 marks for right mean, 2 marks for right covariance matrix Common mistakes: Add variances. Wrong since the Xs are not independent.

   (b) Write this conditional distribution $P(X_1|X_2 = 0)$. ..1

   Simple application of formula of conditional Gaussian gives $P(X_1|X_2 = 0) \sim \mathcal{N}(\mu_1 + \frac{\sigma_{12}}{\sigma_{22}^2}(-\mu_2); \sigma_{11}^2 - \frac{\sigma_{12}^2}{\sigma_{22}^2})$.

   Rubrics: 1 mark for substitution in conditional distribution formula. Here notation for variance of X_1 is σ_{11}^2.

$$\mu_{1|2} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2)$$

$$\Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

(c) Show that this conditional distribution $P(X_2|X_1 \geq 0)$ is not guaranteed to be Gaussian with an example. ..3

Consider a Gaussian distribution with $\Sigma = \left(\begin{smallmatrix}1 & 1\\ 1 & 1\end{smallmatrix}\right)$ In this case $X_1 = X_2$. We know that $P(X_2|X_2 \geq 0)$ is the half Gaussian distribution, not the Gaussian distribution.

Rubrics: 3marks for valid example and justification that not gaussian. 1mark only if example not given.

6. Given two uncorrelated zero-mean random variables $X_1 \sim \mathcal{N}(0, 100)$ and $X_2 \sim \mathcal{N}(0, 1)$, what are the principal components of (X_1, X_2) ..3

We need to find the covariance matrix of the random vector $X = [X_1, X_2]^T$.

The covariance matrix is given by:

$$C = \begin{bmatrix} \text{Var}(X_1) & \text{Cov}(X_1, X_2) \\ \text{Cov}(X_1, X_2) & \text{Var}(X_2) \end{bmatrix} \tag{1}$$

Since, X_1 and X_2 are uncorrelated, $\text{Cov}(X_1, X_2) = 0$. Hence, the covariance matrix is given by:

$$C = \begin{bmatrix} 100 & 0 \\ 0 & 1 \end{bmatrix} \tag{2}$$

The eigen values of the matrix can be calculated by equating the determinant of the following matrix to 0:

$$\begin{vmatrix} 100 - \lambda & 0 \\ 0 & 1 - \lambda \end{vmatrix} = 0 \tag{3}$$

Therefore, the eigen values are $\lambda_1 = 100$ and $\lambda_2 = 1$. The eigen vectors are given by:

The corresponding eigen vectors are given by $[1, 0]^T$ and $[0, 1]^T$. Clearly one of the eigen values is much larger than the other. Therefore, the principal component of the data is given by projection of the data on the first eigen vector which is X_1.

Rubrics: 1 mark for covariance matrix and 2 marks for the final solution. If the covariance matrix is not mentioned/incorrect, then 0 marks have been awarded.

7. Suppose you are trying to design a projection method for a creature that can only visualize in one-dimension. For that we can just use the SNE or T-SNE algorithm but with learning just a single coordinate Y_i for each point X_i in the high-dimensional space.

(a) If the original dataset is 2-dimensional and consists of points uniformly distribution on the surface of a unit circle, what will be the likely SNE projection? ..2

The points are likely to be arranged in a line where two halves of the circle will collapse into each other. For example, if you consider a unit circle, then (1,0) and (-1,0) will collapse to the origin, while (0,1) maps to $\pi/2$ and (0, -1) maps to $-\pi/2$. This way all near neighbors in 2-D stay close together in 1-D. The major distortion is that points on the diametrically opposite semi-circle will be pulled in close.

(b) What is the main reason for favoring the T-SNE projection over the SNE projection? Illustrate with an example if needed. ..3

SNE leads to over-crowding of points because for far-away points the penalty of being far away is a drastic drop in probability with the Gaussian distribution. Heavy-tailed distributions like the t-distribution used in T-SNE, is better capable of spreading out the far away points.

8. Express an MA(1) model as an AR(q) model for an adequate value of q. ..2

$$\begin{aligned} x_1 &= \eta + w_1 \\ x_2 &= \eta + \theta w_1 + w_2 = \eta + \theta(x_1 - \eta) + w_2 \\ &. \\ x_t &= \eta + \theta w_{t-1} + w_t \\ &= \eta(1 - \theta + \theta^2 - \ldots (-1)^{t-1}\theta^{t-1}) + \theta x_{t-1} - \theta^2 x_{t-2} + \ldots (-1)^{t-1} x_1 + w_t \end{aligned}$$

9. Suppose the amount a person A saves in an year is 90% of the amount he saved the previous year and a random amount caused by fluctuations in the market conditions. Let S_t denote the market condition at year t, and let it follow a Gaussian distribution with mean 10 and variance 16 Lakhs. Assume the market condition is random and uncorrelated from one year to the next. Assume the person started with zero savings in year 2000, let the amount saved in year 2000+t be x_t.

   (a) Express x_t with an appropriate time-series model. Is his annual rate of savings showing an increasing or decreasing trend along time? ..3

   This is an AR(1) series of the form $x_t = 10 + 0.9x_{t-1} + W_t$ where $W_t \sim \mathcal{N}(0, 16)$. Since the coefficient of x_{t-1} is of magnitude less than 1, this is a stationary series.

   (b) What is the expected savings at the end of 2030? ..2

   We need to sum up the expected value from 2000 to 2030. Since the series is stationary this becomes $30E(x_t) = 30 * 10/(1 - 0.9) = 3000$ lakhs.

   (c) Now consider a different person B where the amount B saves per year is S_t the market valuation this year, 90% of market condition in previous year S_{t-1}. Express x_t with an appropriate time-series model. Is his annual rate of savings showing any increasing or decreasing trend along time? ..2

   This is an MA(1) series of the form $x_t = 10 + 9 + 0.9w_{t-1} + w_t$ where $w_t \sim \mathcal{N}(0, 16)$. This is a stationary model.

10. Suppose you have n observations as a dataset $D = \{(x_i, y_i) : i = 1 \ldots n\}$ where the response variable $y = \beta x + \alpha + \epsilon_i$ where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. You partition the dataset D into k random disjoint partitions of size n/k each, estimate parameters $(B_1, A_1) \ldots (B_k, A_k)$ from each of the k partitions, and return the estimate as $\bar{B} = \sum_j B_j/k$, $\bar{A} = \sum_j A_j/k$. [Recall that in class we had analyzed that for a standard linear regression model on n samples, the variance of the estimated parameters is given as $Var(B) = \frac{\sigma^2}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}, Var(A) = \frac{\sigma^2 \sum_{i=1}^n x_i^2/n}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}$]

    (a) Write the expression for the distribution of $\bar{B}, \bar{A}$. ..4

    Each A_j and B_j is Gaussian with mean as α, β respectively and variance given above. We know that average of k independent Gaussians is Gaussian with average mean and average variance divided by k. Thus, $\bar{A} \sim \mathcal{N}(\alpha, \sum_{p=1}^k \frac{\sigma^2 \sum_{i \in D_p} kx_i^2/n}{k^2(\sum_{i \in D_p} x_i^2 - \frac{n}{k}\bar{x}_p^2)})$ where D_p denotes the n/k instances in the p-th data partition,

    Expression for $\bar{B}$ is similarly $\bar{B} \sim \mathcal{N}(\beta, \sum_{p=1}^k \frac{\sigma^2}{k^2(\sum_{i \in D_p} x_i^2 - \frac{n}{k}\bar{x}_p^2)})$

    Rubric: 1 mark for mentioning Gaussian distribution for both, 1 mark for mentioning mean for both, 2 marks for correct variances. 1 mark deducted if variance expressed in terms of A_j or B_j without explicit computation. 0.5 marks deducted for basic errors WITHIN the variance term A_j or B_j.

(b) Contrast the above with the estimator A, B of normal linear regression over the entire data. Which estimator has a smaller risk? Justify. ..3

The bias is the same for both estimators. Let us consider the B parameter. We need to contrast variance these two variance terms:
$\sum_{p=1}^{k} \frac{\sigma^2}{k(\sum_{i \in D_p} x_i^2 - \frac{n}{k}\bar{x}_p^2)})$ with $\frac{\sigma^2}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$
Let SS denote $\sum_{i=1}^{n} x_i^2 - n\bar{x}^2$ and let SS_p denote $(\sum_{i \in D_p} x_i^2 - \frac{n}{k}\bar{x}_p^2)$. Each of $SS_p \leq SS$. Thus, the average of their reciprocal will also be greater than $1/SS$. Thus, variance of $\bar{B}$ will be greater.
Rubric: 1 mark for correct conclusion, 1 mark for writing risk term (identifying 0 bias), 1 mark for explicit comparison of variance terms. Other correct justifications given 2 or 3 marks dependent on rigor of proof.

11. Suppose you want to estimate the effect of fan speed on the temperature of a GPU. You have software controls to change the fan speed to any value between 1000 RPM and 3000 RPM. You have time to try 30 different fan speed settings, and record the temperature at each setting. Thereafter, you will estimate a linear regression model to estimate temperature y as a linear function of rpm x. At what 30 fan speeds will you perform your experiment, to reduce the variance of your estimated linear regression parameters. Provide justification. ..3

Uniform distributed in the space between 1000 and 3000 to maximize the value of X-variance.
Rubrics: 3 marks for correct distribution with justification, 1 mark for only mentioning either the distribution or giving correct justification, 0 otherwise

12. In standard linear regression we assumed that the response variable Y follows a Gaussian distribution with mean $E[Y|x] = \mu_x = \alpha + \beta x$. Instead, assume Y is a count variable that follows a Poisson distribution and we model its dependence on x as $\log E[Y|x] = \log \lambda_x = \alpha + \beta x$. You are given a dataset $D = \{(x_i, y_i) : i = 1 \dots n\}$. Write the expression for the maximum likelihood estimation of the parameters. ..4

The MLE is $\sum_i y_i \log \lambda_{x_i} - \lambda_{x_i} - \log y_i$

In terms of parameters this becomes (ignoring terms that do not contain parameters) $\sum_i y_i(A + Bx_i) - e^{A+Bx_i}$

There was a typo in the question where the log before λ_x was missing in the original question. The fix was announced in the class, but if anyone missed, we will consider this alternative solution too:

$\sum_i y_i \log(A + Bx_i) - (A + Bx_i)$

Rubrics: 4 marks for the correct expression and 2 marks for partially correct expression
Compute gradients with respect to any one parameter. ..2

Rubrics: 2 marks for the correct gradient calculation either for A or B but no marks if MLE expression in previous part is incorrect

13. We have four datapoints $x_1 = 0.1, x_2 = 0.4, x_3 = 0.3, x_4 = -0.3$.

Consider the following kernel: $K(t) = \frac{1}{2}I(-1 \leq t \leq 1)$. The bandwidth is $h = 0.3$. What is the value of the kernel density estimator at $x = 0.2$? ..1

$\hat{e}$xtcolorblue5/4

Rubrics: 1/2 mark for answer and 1/2 mark for steps

What is the empirical cumulative density function value at $x = 0.35$ ..1

3/4

Rubrics: 1/2 mark for answer and 1/2 mark for steps

14. Analyze the bias and variance of a kernel density estimator with the Beta density $f(x) = \frac{x^{m-1}(1-x)^{n-1}}{B(m,n)}$ when $0 \leq x \leq 1$ and when the kernel is a uniform kernel $K(t) = \frac{1}{2}I(-1 \leq t \leq 1)$ of width h, and the estimator is made on n samples $X_1, \ldots, X_n$ ..4

Just apply the formula discussed in class.

Bias is approximated as $\frac{1}{2}\sigma_K^2 h^2 f''(x)$. For uniform kernel $\sigma_K^2 = \int_t t^2 k(t) = 1/3$, and for the given density function $f''(x)$ can be calculated easily by double differentiation.

The approximate variance formula is $\frac{f(x)\int_t K^2(t)dt}{nh}$. For the given kernel, $\int_t K^2(t)dt = 1/2$, and $f(x)$ is what is given.

Total: 65