CS 215 : Data Analysis and Interpretation

(Instructor : Suyash P. Awate)

Mid-Semester Examination (Closed Book)

Roll Number: _____

Name: _____

For all questions, if you feel that some information is missing, make justifiable assumptions, state them clearly, and answer the question.

Relevant Formulae

- Poisson: $P(k|\lambda) := \lambda^k \exp(-\lambda)/(k!)$
- Exponential: $P(x;\lambda) = \lambda \exp(-\lambda x); \forall x > 0$
- Univariate Gaussian:

$$P(x;\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-0.5\frac{(x-\mu)^2}{\sigma^2}\right)$$

---

1. (15 points)
   - [2 points] Define the bias of an estimator. Give a mathematical expression.
   - [3 points] Define the variance of an estimator. Give a mathematical expression.
   - [5 points] Mathematically derive the bias-variance decomposition for an estimator.

Please check the class notes.

- [5 points] Suppose the statistic T is an unbiased estimator of the variance parameter σ^2 of a Gaussian distribution.

  Suppose you define $U := \sqrt{T}$ as an estimator of the standard-deviation parameter σ.

  Is U an unbiased estimator ? If yes, prove it. If not, prove so.

Because T is unbiased, $E[T] = \sigma^2_{\text{true}}$

Jensen's inequality gives: $E[U] = E[\sqrt{T}] \leq \sqrt{E[T]} = \sigma_{\text{true}}$

`https://en.wikipedia.org/wiki/Unbiased_estimation_of_standard_deviation`

---

2. (10 points: 5 + 5)

   Mathematically derive the maximum likelihood estimators for the parameters of the Gaussian probability density function (PDF).

Please check the class notes.

---

3. (20 points) Consider two dependent continuous random variables V and D, where V models petrol volume in the tank of a car and D models the distance travelled by the car. While collecting data, the observations of distance D, i.e., $\{d_i\}_{i=1}^N$, are virtually error-free (because the odometer has high accuracy and precision) but the observations about V have measurement errors (because the system to measure petrol volume in petrol tank has high accuracy and low precision). Thus, instead of the error-free values $\{v_i\}_{i=1}^N$, the measurements are the perturbed values $\{\widehat{v}_i := v_i + \epsilon_i\}_{i=1}^N$, where ϵ_i is a random perturbation caused by measurement error.

• [5 points] Propose a statistical model for the measurement errors ϵ_i in each v_i, parameterized by a set of parameters θ_1.

Gaussian with zero mean and positive variance, i.e., each ϵ_i is independently drawn from $G(0, \sigma^2)$

$\theta_1 = \sigma$

• [5 points] For the functional dependency between V and D, propose a model parameterized by a set of parameters θ_2.

---

linear relationship between distance covered and petrol used, i.e., $V = mD + c$ with $0 < m < \infty$

$\theta_1 = m, c$

• [2 points] Using the proposed model for measurement errors and the proposed model for the functional dependency, based on concepts covered in class, formulate a strategy to infer the model parameters $\theta := \theta_1 \cup \theta_2$ from the observations $\{(d_i, \widehat{v}_i)\}_{i=1}^N$.

$\arg\max_{m,c,\sigma} \prod_{i=1}^N P(\widehat{v}_i, d_i)$

$= \arg\max_{m,c,\sigma} \prod_{i=1}^N G(\widehat{v}_i; md_i + c, \sigma^2)$

• [8 points: 4 + 4] Mathematically derive and propose an implementable algorithm to solve for θ_1 and θ_2 given the observations $\{(d_i, \widehat{v}_i)\}_{i=1}^N$.

---

for m, c, irrespective of σ: linear system of equations obtained by assigning partial derivatives of log-likelihood function to zero; 2 equations; 2 unknowns.

given estimates $\widehat{m}, \widehat{c}$: solve 1 equation in 1 unknown σ

`https://www.math.arizona.edu/~jwatkins/n-mle.pdf`

---

4. (20 points)

Let S be a continuous random variable modeling the examination scores of students in a class.

Scores are real-valued within the range $[0, 100]$.

The mean of S is 60.

• [4 points] Mathematically derive a non-trivial upper bound or lower bound (using concepts covered in class) for $P(S \geq 90)$.

Markov inequality on S gives: $P(S \geq 90) \leq 60/90$

• [7 points] Mathematically derive a non-trivial upper bound or lower bound (using concepts covered in class) for $P(S \leq 20)$.

Let $S' := 100 - S$. Then S' also lies within $[0, 100]$.

$E[S'] = 100 - E[S] = 40$

Applying Markov inequality to S' gives: $P(S \leq 20) = P(S' \geq 80) \leq 40/80$

• [9 points] Using an additional fact that the standard deviation of S is 10,

mathematically derive a non-trivial upper bound or lower bound (using concepts covered in class) for $P(S \leq 20)$.

Let $S' := 100 - S$. Then S' also lies within $[0, 100]$.

$E[S'] = 100 - E[S] = 40$

$SD(S') = SD(S) = 10$

$P(S \leq 20) = P(S' \geq 80) = P(S' - 40 \geq 40) = P(|S' - 40| \geq 40)$ (because we know $S' \geq 0$)

Chebyshev's inequality gives $P(|S' - 40| \geq 40) \leq 10^2/40^2$
