Midterm Exam, Fall 2023: CS 215

Write your name and roll number on the answer sheet. Attempt all questions. You have 2 hours for this exam. Clearly mark out rough work. No calculators or phones are allowed (or required). You may directly use results/theorems we have stated or derived in class, unless the question explicitly mentions otherwise. **Avoid writing lengthy answers.**

Useful Information

1. Markov's inequality: For a non-negative random variable X, we have $P(X \geq a) \leq E(X)/a$ where $a > 0$.
2. Chebyshev's inequality: For a r.v. X with mean μ and variance σ^2, we have $P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$.
3. Gaussian PDF: If $X \sim \mathcal{N}(\mu, \sigma^2)$, then $f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}$, MGF $\phi_X(t) = e^{\mu t + \sigma^2 t^2/2}$
4. Poisson PMF: $P(X = i) = \frac{e^{-\lambda}\lambda^i}{i!}$, MGF $\phi_X(t) = e^{\lambda e^t - 1}$
5. Integration by parts: $\int udv = uv - \int vdu$
6. Gaussian tail bound: If $X \sim \mathcal{N}(0, 1)$, then $P(X > x) \leq \frac{e^{-x^2/2}}{x\sqrt{2\pi}}$.

---

1. In a certain town, there exist 100 rickshaws out of which 1 is red and 99 are blue. A person XYZ observes a serious accident caused by a rickshaw at night and remembers that the rickshaw was red in color. Hence, the police arrest the driver of the red rickshaw. The driver pleads innocence. Now, a lawyer decides to defend the hapless rickshaw driver in court. The lawyer ropes in an opthalmologist to test XYZ's ability to differentiate between the colors red and blue, under illumination conditions similar to those that existed that fateful night. The opthalmologist suggests that XYZ sees red objects as red 99% of the time and blue objects as red 2% of the time. What will be the main argument of the defense lawyer? (That is, what is the probability that the rickshaw was really a red one, when XYZ observed it to be red?) **Show clearcut steps for your answer.** [10 points]

2. If X and Y are two independent continuous random variables with PDFs f_X and f_Y respectively, then prove that the PDF of $Z = XY$ is given by $f_Z(z) = \int_{-\infty}^{+\infty} f_X(x) f_Y(z/x) \frac{1}{|x|} dx$. Clearly point out where the independence is used. Also take care that X or Y may be negative-valued. [10 points]

3. (a) Prove that the sum of n independent Gaussian random variables with means $\mu_1, \mu_2, ..., \mu_n$ and variances $\sigma_1^2, \sigma_2^2, ..., \sigma_n^2$ respectively is also a Gaussian random variable even when n is small (thus, do not use the central limit theorem). What is its mean and variance? We may have derived this result in class, but re-derive it. [4+1=5 points]
   (b) If $X \sim \mathcal{N}(0, 1)$, then derive the CDF, PDF and mode of $Z = X^2$. [3+1+1=5 points]

4. Let $X_1, X_2, ..., X_n$ be a sequence of n independent Bernoulli random variables with success probability p. We have seen how to obtain the maximum likelihood estimate of p in class. Suppose now we have the additional knowledge that $p \in \{1/2, 1\}$, i.e. p could be either 1 or 1/2. Define $Y = \sum_{i=1}^{n} X_i$. Consider the following estimator U for p:

$$U = \begin{cases} 1, \text{ if } Y = n \\ 0.5 \text{ if } Y < n \end{cases} \tag{1}$$

   Now answer the following questions [2.5 + 2.5 + 2.5 + 2.5 = 10 points]:

(a) Derive $E(U)$ if $p = 1$ and if $p = 1/2$.

(b) Is U a biased estimator for finite n? Is it biased as $n \to \infty$? Justify.

(c) Derive the MSE for U and show that it goes to 0 as $n \to \infty$.

(d) Is U a maximum likelihood estimator for p? Why (not)?

5. Suppose some n individuals have arrived at a lab for RTPCR testing. An RTPCR test involves extracting the nasal mucus of the individual and testing it within an RTPCR machine. Suppose that the probability that any individual is infected is p. Instead of individually testing each person, we follow the two-step procedure described below to save on the number of tests: (1) We divide the people into n/g pools, each of size g where we assume that g divides n. Small, equal-volume portions of the mucus samples of all individuals belonging to the same pool are mixed together. This mixture is tested, thus leading to n/g independent tests, one per pool. (2) If the mixture tests negative (non-infected), then all pool members are declared negative. If the mixture tests positive (infected), then each member of the pool is individually tested in a second round of tests. This procedure is called Dorfman pooling. In addition, suppose we assume that each test on a pool containing at least one infected sample has a fixed probability u of correctly returning a positive result, and that each test on a pool of entirely non-infected samples has a fixed probability v of correctly returning a negative result, both independently of the outcomes of all other tests, and both independently of the size of the pool (including a pool-size of 1). As per this protocol, we declare an individual to be infected if and only if both his/her individual test and pooled test are positive. Now answer the following, and **provide some steps/calculations for your answers**:

   (a) What is the expected total number of tests? Note that an individual test counts as one test, and the test of a mixture also counts as one test. [3 points]

   (b) What is the expected number of false negatives for Dorfman pooling? How does this compare to the expected number of false negatives of individual testing? [1+1+1 = 3 points]

   (c) What is the expected number of false positives for Dorfman pooling? How does this compare to the expected number of false positives of individual testing? [3+1+1=5 points]

6. Pooled testing of a different form than Dorfman's method from the previous question has also been advocated as a method for saving resources in COVID-19 testing. Instead of testing the samples of n people separately (one sample per person), we create $m < n$ pools and test the m pools. Each pool is obtained by mixing fixed, small portions of a randomly chosen subset of the n samples. Let $\boldsymbol{y}$ be a binary vector of m elements, where $y_i = 1$ if the i^{th} pool is positive (i.e. at least one of the contributing samples is infected) and 0 if the pool is negative (i.e. non-infected which means that none of the contributing samples are infected). Let $\boldsymbol{x}$ be the binary vector of n elements where $x_j = 1$ if the j^{th} sample is infected and 0 otherwise. We have the relationship $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$ where $\boldsymbol{A}$ is a $m \times n$ pooling matrix, where $A_{ij} = 1$ if the j^{th} sample contributed to the i^{th} pool and 0 otherwise. (In this special form of matrix-vector multiplication, for any i, we have $y_i = A_{i1}x_1 \vee A_{i2}x_2 \vee ... \vee A_{in}x_n$ where $\vee$ is the logical OR operator.) The aim of pooled testing is to determine $\boldsymbol{x}$ directly from $\boldsymbol{y}$ and $\boldsymbol{A}$. It turns out that such a procedure is feasible if the number of infected people k is small compared to n and $\boldsymbol{A}$ is carefully designed. Note that we have no knowledge of k beforehand. This question seeks to explore a technique to estimate k directly from $\boldsymbol{y}$ and $\boldsymbol{A}$, assuming that the entries of $\boldsymbol{A}$ are drawn independently from Bernoulli(0.5) (i.e. $P(A_{ij} = 0) = P(A_{ij} = 1) = 0.5$). To this end, answer the following questions: [2.5+2.5+2.5+2.5=10 points]

   (a) Let d_i be the number of entries for which A_{ij} and x_j are both unequal to 0, where $1 \leq j \leq n$. What is the distribution of d_i, if the number of non-zero elements of $\boldsymbol{x}$ is k, since the entries of $\boldsymbol{A}$ are drawn independently from Bernoulli(0.5)?

   (b) Prove that $P(y_i = 0) = P(d_i = 0)$.

   (c) Let H be a random variable for the number of non-zero elements in $\boldsymbol{y}$. Then what is the distribution of H, if the number of non-zero elements of $\boldsymbol{x}$ is k?

   (d) Express k in terms of $P(d_i = 0)$ and hence write the maximum likelihood estimate of k given $\boldsymbol{y}$.