Midterm Exam: CS 215

Write your roll number on the answer sheet. Attempt all questions. You have 120 minutes 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. The empirical mean of n independent and identically distributed random variables is approximately Gaussian distributed. The approximation accuracy is better when n is larger. If the random variables are Gaussian, the empirical mean is exactly Gaussian distributed.
2. Markov's inequality: For a non-negative random variable X, we have $P(X \geq a) \leq E(X)/a$ where $a > 0$.
3. Chebyshev's inequality: For a random variable X with mean μ and variance σ^2, we have $P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$.
4. 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 $\phi_X(t) = e^{\mu t + \sigma^2 t^2/2}$
5. Poisson PMF: $P(X = i) = \frac{e^{-\lambda}\lambda^i}{i!}$, MGF $\phi_X(t) = e^{\lambda e^t - 1}$
6. Integration by parts: $\int u dv = uv - \int v du$
7. 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. Let $X_1, X_2, ..., X_n$ be independent random variables from the same Gaussian distribution with unknown mean μ. Let $\nu = g(\mu)$ where g is a bijective function. Let $\hat{\nu}$ and $\hat{\mu}$ denote the maximum likelihood estimates for μ and ν respectively. Determine whether $\hat{\nu} = g(\hat{\mu})$ in the following two cases: (a) $g(\mu) = a\mu + b$ where $a \neq 0$ and b are constants; (b) $g(\mu) = \mu^3$, assuming $\mu \neq 0$. In both cases, also determine whether the estimate $\hat{\nu}$ is unbiased. You must provide proper reasoning for all answers (no credit otherwise). [4+3+4+3=14 points]
2. A biologist has taken the picture of a bird in SGNP. She/He knows the bird in the picture belongs to one of two similar-looking species A and B. Based on collected statistics, (s)he knows that (1) the population of B in SGNP is twice that of A, and that (2) the average beak-length (henceforth referred to as x) is distributed as $\mathcal{N}(1, 1)$ for species A and as $\mathcal{N}(2, 2)$ for species B. Write an expression for the conditional probability that the bird belongs to species A given a value of x, and repeat this for B. For what values of x, will it be impossible for the biologist to classify the bird as belonging to A or B, based on these conditional probabilities using feature x alone? [5+5=10 points]
3. Derive the covariance matrix for a multinomial distribution that models outcomes from n trials and k categories, with success probability $p_1, p_2, ..., p_k$ respectively. The multinomial pmf is given as $P(X = \mathbf{x}; n, \{p_i\}_{i=1}^k) = \frac{n!}{x_1!x_2!...x_k!} p_1^{x_1} p_2^{x_2} ... p_k^{x_k}$ where $\mathbf{x} = (x_1, x_2, ..., x_k)$ and $\forall i, 0 \leq p_i \leq 1, \sum_{i=1}^k p_i = 1, \sum_{i=1}^k x_i = n$. [14 points]
4. If $Y \sim$ Uniform(a, b) where $0 < a < b$, derive the mean, median, variance, PDF and CDF of $Z = \frac{1}{Y}$. [14 points]

5. Let $X \sim \text{Poisson}(\lambda)$. Define $h(u) = 2\dfrac{(1+u)\log(1+u)-u}{u^2}$. Your task here is to prove that $P(X \geq \lambda+x) \leq e^{-\frac{x^2}{2\lambda}h(x/\lambda)}$ and $P(X \leq \lambda - x) \leq e^{-\frac{x^2}{2\lambda}h(-x/\lambda)}$.
For the first inequality, do as follows: (i) Consider that $P(X \geq \lambda + x) = P(e^{tX} \geq e^{t(\lambda+x)})$ for some t. Use this to show that $P(X \geq \lambda + x) \leq E[e^{t(X-\lambda-x)}]$. (ii) Now take the minimum of the RHS over t and deduce the first inequality.
Carry out similar steps for the second inequality as well.
(Not needed for this question, but just to provide some context: These inequalities are useful in deriving tail bounds for the Poisson distribution.) [3+4+3+4=14 points]

6. Consider you are given a set S_1 of n_1 samples of a Gaussian random variable with unknown mean μ_1 and unknown variance σ^2, a set S_2 of n_2 samples of a Gaussian random variable with unknown mean μ_2 and unknown variance σ^2, ..., and a set S_k of n_k samples of a Gaussian random variable with unknown mean μ_k and unknown variance σ^2. Derive a maximum likelihood estimate for σ^2 assuming all $n = n_1 + n_2 + ... + n_k$ samples are mutually independent, and assuming that you know the set-identity of every sample. Note that your estimate should be derived from samples of all k Gaussians. Is the estimate unbiased? Justify. If not, state a feasible correction to the estimate to make it unbiased and justify it. *A feasible correction means a correction that will not require knowledge of unknown parameters.*
Now, if $\mu_1, \mu_2, ..., \mu_{k-2}$ were known but not μ_k, μ_{k-1}, how does this maximum likelihood estimate change? Is the estimate unbiased? Justify. If not, state a feasible correction to the estimate to make it unbiased and justify it. [2+5+3+4=14 points]

7. Consider the empirical distribution function $F_n(x) = \sum_{i=1}^{n} \mathbf{1}(X_i \leq x)/n$ where $\{X_i\}_{i=1}^n$ are n iid random variables with the (true) CDF $F(x)$, and $\mathbf{1}(z)$ is an indicator function which yields 1 if the predicate z is true and 0 if z is false. $F_n(x)$ is an estimator of $F(x)$ and uses the values of $\{X_i\}_{i=1}^n$. Answer the following questions:
   (a) Determine the bias, variance and MSE of the estimator $F_n(x)$. [4+4 = 8 points]
   (b) Using Chebyshev's inequality (CI), derive an upper bound for $P(|F_n(x) - F(x)| \geq \epsilon)$ for any $\epsilon > 0$. [2 points]
   (c) Derive an upper bound for $P(|F_n(x) - F(x)| \geq \epsilon)$ for any $\epsilon > 0$, using the central limit theorem (CLT). [4 points]
   (d) How does the CLT-based upper bound compare with the CI-based one? Briefly (i.e. in 1-2 sentences) state pros and cons. [2 points]
   (e) There exists a very well-known result called the Dvoretzky-Kiefer-Wolfowitz (DKW) inequality, which states that $P(\max_{x \in \mathbb{R}} |F_n(x) - F(x)| \geq \epsilon) \leq 2e^{-2n\epsilon^2}$ for any $\epsilon > 0$. Using this inequality, a confidence interval for $F_n(x)$ of the following form can be constructed: $P(\forall x \in \mathbb{R}, L(x) \leq F(x) \leq U(x)) \geq 1 - \alpha$. Your task is specify what $L(x)$ and $U(x)$ would be, given a value of α. [4 points]