Midterm Exam: CS 215

Attempt all six questions. Each question carries 10 points for a total of 60. You have a time of 120 minutes for this exam. It is your responsibility to clearly mark out rough work. No calculators or phones are allowed (or required :-)).

Useful Information

1. Binomial theorem: $(x+y)^n = \sum_{k=0}^{n} C(n,k) x^k y^{n-k}$
2. The empirical mean of n independent and identically distributed random variables is approximately Gaussian distributed. The approximation accuracy is better when n is larger.
3. For a non-negative random variable X, we have $P(X \geq a) \leq E(X)/a$ where $a > 0$.
4. For a random variable X with mean μ and variance σ^2, we have $P(|X-\mu| \geq k\sigma) \leq \dfrac{1}{k^2}$.
5. Integration by parts: $\int u dv = uv - \int v du$.
6. Gaussian pdf: $f(x) = \dfrac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}$
7. Poisson pmf: $P(X=i) = \dfrac{e^{-\lambda}\lambda^i}{i!}$

---

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 cases: (a) $g(\mu) = a\mu + b$ where $a \neq 0$ and b are constants, (b) $g(\mu) = \mu^2$, assuming $\mu > 0$ for simplicity. In both cases, also determine whether the estimate $\hat{\nu}$ is unbiased. You must provide proper reasoning for all answers (no credit otherwise). [10 points]

2. (a) A student is trying to design a procedure to generate a sample from a distribution function F which we will assume to be invertible. For this, (s)he generates a sample u_i from a $[0,1]$ uniform distribution using the 'rand' function of MATLAB, computes $v_i = F^{-1}(u_i)$. This is repeated n times for $i = 1...n$. Prove that the values $\{v_i\}_{i=1}^{n}$ follow the distribution F.

   (b) Let $Y_1, Y_2, ..., Y_n$ represent data from a continuous distribution F. The empirical distribution function F_e of these data is defined as $F_e(x) = \dfrac{\sum_{i=1}^{n} \mathbf{1}(Y_i \leq x)}{n}$ where $\mathbf{1}(z) = 1$ if the predicate z is true and 0 otherwise. Now define $D = \max_x |F_e(x) - F(x)|$. Also define $E = \max_{0 \leq y \leq 1} \left| \dfrac{\sum_{i=1}^{n} \mathbf{1}(U_i \leq y)}{n} - y \right|$ where $U_1, U_2, ..., U_n$ represent data from a $[0,1]$ uniform distribution. Now prove that $P(E \geq d) = P(D \geq d)$. (This is a surprising result, as it proves that the distribution of D does not depend upon F!). [4+6 = 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 $P(X = \mathbf{x}; n, \{p_i\}_{i=1}^{k}) = \dfrac{n!}{x_1! x_2! ... x_k!} p_1^{x_1} p_2^{x_2} ... p_k^{x_k}$ where $\mathbf{x} = (x_1, x_2, ..., x_k)$ and $\forall i, 0 \leq p_i \leq 1, \sum_{i=1}^{k} p_i = 1, \sum_{i=1}^{k} x_i = n$. [10 points]

4. Verify whether true or false with justification: (a) The minimum of n iid Bernoulli random variables is also a Bernoulli random variable. (If your answer is in the affirmative, what is the parameter of the Bernoulli random variable?). (b) The minimum of n iid geometric random variables is also a geometric random variable. (If your answer is in the affirmative, what is the parameter of the geometric random variable?). Recall that the pmf of a geometric random variable has the form $P(X = i) = (1 - p)^{i-1}p$. [5+5=10 points]

5. A Laplace random variable X with parameters μ, b has the pdf $f_X(x) = \dfrac{1}{2b} \exp\left(- \dfrac{|x - \mu|}{b} \right)$, and cdf $F_X(x) = \dfrac{1}{2} \exp\left(\dfrac{x - \mu}{b} \right)$ when $x < \mu$ and $F_X(x) = 1 - \dfrac{1}{2} \exp\left(- \dfrac{x - \mu}{b} \right)$ when $x \geq \mu$. Derive the maximum likelihood estimate $\hat{\mu}$ for μ given n independent samples, assuming n to be odd. Deduce the pdf of $\hat{\mu}$ in terms of F_X and f_X. What happens to $\hat{\mu}$ when n is even? [3+6+1=10 points]

6. Consider you are given a set of n_1 samples of a Gaussian random variable with unknown mean μ_1 and unknown variance σ^2, a set of n_2 samples of a Gaussian random variable with unknown mean μ_2 and unknown variance σ^2, ..., and a set 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 which sample belongs to which set. 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. Now, if $\mu_2, \mu_3, ..., \mu_k$ were known but not μ_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. [1+4+2+3=10 points]