Quiz 1: CS 215

Name: ____ **Roll Number:** _________________________

Attempt all five questions, each carrying 10 points. Clearly mark out rough work.

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 with finite variance is approximately Gaussian distributed. The approximation accuracy is better when n is larger.

3. Defining $\Phi(x) = \int_{-\infty}^{x} \frac{e^{-x^2/2}}{\sqrt{2\pi}} dx$, we have the following table:

n	$\Phi(n) - \Phi(-n)$
1	68.2%
2	95.4%
2.6	99%
2.8	99.49%
3	99.73%

4. For a non-negative random variable X, we have $P(X \geq a) \leq E(X)/a$ where $a > 0$.

5. For a random variable X with mean μ and variance σ^2, we have $P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$.

6. Integration by parts: $\int u dv = uv - \int v du$.

7. Gaussian pdf: $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}$

8. Poisson pmf: $P(X = i) = \frac{e^{-\lambda}\lambda^i}{i!}$

Additional space

1. Consider two random variables X_1 and X_2 which are identically distributed. Verify whether each of the following statements is true or false with justification. No credit without correct justification. (a) $X_1 - X_2$ and $X_1 + X_2$ are uncorrelated. (b) $aX_1 - bX_2$ and $aX_1 + bX_2$ are uncorrelated where $a > 0, b > 0$ are constants. [5+5=10 points]
(a) We need to check whether $E[(X_1 - X_2)(X_1 + X_2)] = E[X_1 - X_2]E[X_1 + X2]$. The LHS equals $E[X_1^2 - X_2^2] = E[X_1^2] - E[X_2^2] = 0$ as X_1, X_2 are identically distributed. The RHS is also 0 for the same reason. Hence $X_1 - X_2$ and $X_1 + X_2$ are uncorrelated.
(b) We need to check whether $E[(aX_1 - bX_2)(aX_1 + bX_2)] = E[aX_1 - bX_2]E[aX_1 + bX2]$. The LHS equals $E[a^2X_1^2 - b^2X_2^2] = (a^2 - b^2)E[X_1^2]$ as X_1, X_2 are identically distributed. The RHS equals $(a^2 - b^2)[E(X_1)]^2$ for the same reason. Hence one cannot conclude that $X_1 - X_2$ and $X_1 + X_2$ are uncorrelated unless $E[X_1^2] = [E(X_1)]^2$.

2. Here is a nice application of basic probability. I hope you will like it! 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 will test positive is p, and let us assume that the test results across all n individuals are independent. 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 groups, 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 group are mixed together. This mixture is tested, thus leading to n/g independent tests, one per group. (2) If the mixture tests negative (non-infected), then all group members are declared negative. If the mixture tests positive (infected), then each member of the group is individually tested in a second round of tests.
It is known that the mixture of different mucus samples, or using small portions of the sample, has no influence on the probability p or on test accuracy. This procedure is called Dorfman pooling and it was widely used during the COVID-19 pandemic.
(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.
(b) Now suppose that exactly $k \ll n$ individuals are infected. In this scenario, what is the number of tests required in Dorfman's method in the worst case? For what value of g, expressed in terms of n and k, will this worst case number of tests be minimized? What is the number of tests in that case? [5 + (2.5 + 2.5) = 10 points]
(a) The total number of tests is n/g in the first round. The probability of any one individual being infected is p. Hence the probability of any one individual being non-infected is $1 - p$. The probability of obtaining a group of g individuals who are all non-infected is $(1 - p)^g$, and hence the probability of obtaining a group of g individuals containing at least one infected individual is $1 - (1 - p)^g$. As the total number of groups is n/g, the expected number of groups with at least one infected member is $n/g \times [1 - (1 - p)^g]$. Hence the expected number of total tests is $n/g + g \times n/g \times [1 - (1 - p)^g] = n/g + n[1 - (1 - p)^g]$.
(b) In the worst case, each of the k infected people will lie in a different group. So each of these k groups will have to be tested in the second round, and each member of these k groups will be tested individually. This will give rise to $k \times g$ more tests. So the total number is $n/g + kg$ for the worst case number of tests. If the worst case number has to be optimal, we set the derivative of this number (w.r.t. g) to zero, giving rise to $-n/g^2 + k = 0$, that is $g = \sqrt{n/k}$. The number of tests in this case will be $2\sqrt{nk}$.

3. Consider a set of independent random variables $X_1, X_2, ..., X_n$ from a distribution with parameter α. A quantity V is said to be a **sufficient statistic for the parameter** α if the conditional distribution of $X_1, X_2, ..., X_n$ given any value $V = v$ is independent of α (i.e. if the formula for the conditional distribution does not contain any terms in α). Now suppose $X_1, X_2, ..., X_n$ are independent random variables from a Gaussian distribution with known variance σ^2 but unknown mean μ. Let $\tilde{V} = \sum_{i=1}^n X_i/n$. Determine with appropriate justification using the conditional distribution whether $\tilde{V}$ is a sufficient statistic for parameter μ. [10 points]
We have $p(X_1, X_2, ..., X_n|\tilde{V}) = p(X_1, X_2, ..., X_n, \tilde{V})/p(\tilde{V})$. Since $X_1, X_2, ..., X_n$ are independent random variables from $\mathcal{N}(\mu, \sigma^2)$, we know that $\tilde{V} \sim \mathcal{N}(\mu, \sigma^2/n)$. Thus $p(\tilde{V} = \bar{x}) = \dfrac{e^{-n(\bar{x}-\mu)^2/2\sigma^2}}{\sqrt{2\pi}\sigma/\sqrt{n}}$ where $\bar{x}$ stands

for the arithmetic mean. On the other hand, $p(X_1, X_2, ..., X_n, \tilde{V}) = p(X_1, X_2, ..., X_n) = \Pi_{i=1}^n p(X_i) = \dfrac{e^{-\sum_{i=1}^n (x_i-\mu)^2/(2\sigma^2)}}{(2\pi\sigma^2)^{n/2}} = \dfrac{e^{-[\sum_{i=1}^n (x_i-\bar{x})+n(\bar{x}-\mu)^2]/(2\sigma^2)}}{(2\pi\sigma^2)^{n/2}}$. This simplifies to $p(X_1, X_2, ..., X_n, \tilde{V}) = p(X_1, X_2, ..., X_n) = \dfrac{e^{-\sum_{i=1}^n (x_i-\bar{x})^2/(2\sigma^2)} e^{-n(\bar{x}-\mu)^2/(2\sigma^2)}}{(2\pi\sigma^2)^{n/2}}$.

Thus $p(X_1, X_2, ..., X_n|\tilde{V}) = p(X_1, X_2, ..., X_n, \tilde{V})/p(\tilde{V})$ will contain terms with $\sigma, \bar{x}$ but not μ because the exponential term $e^{-n(\bar{x}-\mu)^2/(2\sigma^2)}$ will cancel out from both the numerator and the denominator. As the conditional density has no term in μ, we see that $\tilde{V}$ is a sufficient statistic for μ.

4. Here is a nice application of some of the concepts we have studied in class to basic machine learning. I hope you will like it! The Greater Mumbai Region is home to thousands of flamingoes, which belong to one of two predominant species: greater and lesser flamingo. Let us suppose you know that the population of greater flamingoes is thrice that of lesser flamingoes. Furthermore, suppose it is known that the height of a greater flamingo is Gaussian distributed with mean and variance both equal to 4, and that the height of the lesser flamingo is Gaussian distributed with mean 3 and variance 5. Let x be the height of a given flamingo, and you want to classify it as greater or lesser. Determine the conditional probability that it is a greater flamingo given x, and the conditional probability that it is a lesser flamingo given x. For what values of x is the former probability equal to the latter? [3.5+3.5+3=10 points]
Let G and L be the events that the flamingo under consideration is a greater and lesser flamingo respectively. Then $P(G|x) = p(x|G)P(G)/p(x)$ and likewise $P(L|x) = p(x|L)P(L)/p(x)$. As $P(G) = 3P(L)$ and $P(G) + P(L) = 1$, we have $P(L) = 0.25, P(G) = 0.75$. Thus we have $P(G|x) = \dfrac{e^{-(x-4)^2/(2*4)}}{2\sqrt{2\pi}} \times 0.25/p(x)$ and $P(L|x) = \dfrac{e^{-(x-3)^2/(2*5)}}{\sqrt{5}\sqrt{2\pi}} \times 0.75/p(x)$.
For the last part, we equate $P(G|x) = P(L|x)$. Cancelling out common terms and taking logarithms on both sides, we obtain $\dfrac{(x-4)^2}{8} = \dfrac{(x-3)^2}{10} + \log(\sqrt{5}/6)$.

5. If $X \sim \mathcal{N}(\mu, \sigma^2)$, then express the CDF of $Y = aX + b$ in terms of the CDF of X. Also write down the PDF of Y. Here a, b are non-zero constants. If the PDF of X is $f_X(.)$ and $Y = aX + b$ as before, write down an expression for the PDF of Y, i.e. $f_Y(.)$ in terms of $f_X(.)$. [6+1+3=10 points]
In the case where $a > 0$, we have $F_Y(y) = P(Y \leq y) = P(aX + b \leq y) = P(X \leq (y-b)/a) = F_X((y-b)/a)$. In the case that $a < 0$, we have $F_Y(y) = P(Y \leq y) = P(aX+b \leq y) = P(X \geq (y-b)/a) = 1-F_X((y-b)/a)$. This is the expression for the CDF of Y.
The PDF is obtained by taking the derivative w.r.t. x, giving $f_Y(y) = f_X((y-b)/a)/a$ when $a > 0$ and $f_Y(y) = -f_X((y-b)/a)/a$ when $a < 0$. This yields $f_Y(y) = f_X((y-b)/a)/|a|$.
Now since $f_X(x) = e^{-(y-\mu)^2/(2\sigma^2)}/(\sigma\sqrt{2\pi})$, we have $f_Y = f_X((y-b)/a)/|a| = e^{-((y-b)/a-\mu)^2/(2\sigma^2)}/(|a|\sigma\sqrt{2\pi}) = e^{-(y-b-a\mu)^2/(2a^2\sigma^2)}/(|a|\sigma\sqrt{2\pi})$.