Midterm Exam #1

MATH 60061/70061: Mathematical Statistics I

October 7, 2020

This is a take-home exam. Please submit your answers as a **PDF** file to Blackboard by **11:59 p.m. on October 8**. Please show your work and write legibly. Your grade will be based on the correctness of your answers and the clarity with which you express them. Collaboration, copying, and cheating are not allowed.

Problems

1. (10 points) You divide your emails into 2 categories: "spam" and "not spam", denoted by C_1 and C_2, respectively. From previous experience you find that $P(C_1) = 0.7$ and $P(C_2) = 0.3$. Based on the observation that some words (e.g., "free") are more likely to appear in a spam email than in a non-spam email, you create a list of 50 words that are more likely to be used in spam than in non-spam. Let W_j be the event that an email contains the jth word on the list, and
$$p_j = P(W_j \mid C_1), \quad r_j = P(W_j \mid C_2),$$
for $j = 1, \ldots, 50$. With the 50-word list, you design a spam filter based on the *naïve Bayes* classifier, which assumes that $W_1, \ldots, W_{50}$ are *conditionally independent* given that an email is spam, and *conditionally independent* given that it is not spam. A new email has just arrived, and it includes the first 3 words on the list (but not the other 47). Given the information, what is the probability that the new email is spam?

2. (10 points) Let $X_1, X_2, \ldots, X_n$ be *independent and identically distributed* (i.i.d.) random variables, with expectation $E(X_i) = \mu$ and variance $\text{Var}(X_i) = \sigma^2$. The sample mean is defined as
$$\bar{X}_n = \frac{1}{n}(X_1 + X_2 + \cdots + X_n).$$
   a. What is the expected value of $\bar{X}_n$?

   b. What is the variance of $\bar{X}_n$?

3. (25 points) Let X and Y be independent random variables, each exponentially distributed with mean $1/5$.

   a. (5 points) Find the probability $P(X \geq 5Y)$.

   b. (10 points) Let Z be the maximum (i.e., larger) of X and Y. Find the probability density function (PDF) of Z.

c. (10 points) Let $Q = \sqrt{X}$. Find the PDF of Q.

4. (30 points) Consider the hierarchical model

$$
\begin{aligned}
X_i \mid P_i &\sim \text{Bern}(P_i), \quad i = 1, \ldots, n, \\
P_i &\sim \text{Beta}(a, b).
\end{aligned}
$$

a. Show that for $Y = \sum_{i=1}^{n} X_i$, $E(Y) = na/(a+b)$ and $\text{Var}(Y) = nab/(a+b)^2$.

b. Show that Y is a $\text{Bin}(n, \frac{a}{a+b})$ random variable.

c. Suppose now that the model is

$$
\begin{aligned}
X_i \mid P_i &\sim \text{Bin}(n_i, P_i), \quad i = 1, \ldots, k, \\
P_i &\sim \text{Beta}(a, b).
\end{aligned}
$$

Show that for $Y = \sum_{i=1}^{k} X_i$, $E(Y) = \frac{a}{a+b} \sum_{i=1}^{k} n_i$, and $\text{Var}(Y) = \sum_{i=1}^{k} \text{Var}(X_i)$, where

$$
\text{Var}(X_i) = n_i \frac{ab(a+b+n_i)}{(a+b)^2(a+b+1)}
$$

5. (25 points) Consider an experiment where we observe the value of a random variable X, and estimate the value of an unknown parameter θ. As in the Bayesian perspective, we assume that X and θ have a joint distribution. Let $\hat{\theta}$ be the estimator (which is a function of X). Then $\hat{\theta}$ is said to be *unbiased* if $E(\hat{\theta} \mid \theta) = \theta$, and $\hat{\theta}$ is said to be the *Bayes procedure* if $E(\theta \mid X) = \hat{\theta}$.

a. (10 points) Let $\hat{\theta}$ be unbiased. Find $E(\hat{\theta} - \theta)^2$ (the average squared difference between the estimator and the true value of θ), in terms of marginal moments of $\hat{\theta}$ and θ.

b. (10 points) Let $\hat{\theta}$ be Bayes procedure. Find $E(\hat{\theta} - \theta)^2$ in terms of marginal moments of $\hat{\theta}$ and θ.

c. (5 points) Show that it is *impossible* for $\hat{\theta}$ to be both the Bayes procedure and unbiased, except in trivial problems where we get to know θ perfectly by observing X.