Assignment 2

15-359 Probability and Computing

Name: Shashank Singh
Email: sss1@andrew.cmu.edu
Section: B

Due: Friday, February 3, 2012

Problem 1: Cell Block (exercise)

By definition of expected value,

$$\begin{aligned} E\left(\frac{X}{Y} \middle| X^2 + Y \leq 10\right) &= \frac{\left(\frac{1}{3} * \frac{1}{12} + \frac{2}{3} * \frac{2}{12} + \frac{1}{8} * \frac{3}{12}\right)}{\frac{6}{12}} \\ &= \boxed{\frac{49}{144}.} \end{aligned}$$

Problem 2: Friend of a friend (exercise)

By definition of conditional probability and conditional independence,

$$\begin{aligned} P(A|B \cap C) &= \frac{P(A \cap B \cap C)}{P(B \cap C)} \\ &= \frac{P(A \cap B|C) \cdot P(C)}{P(B \cap C)} \\ &= \frac{P(A|C) \cdot P(B|C) \cdot P(C)}{P(B \cap C)} \\ &= \frac{P(A|C) \cdot P(B \cap C) \cdot P(C)}{P(B \cap C) \cdot P(C)} \\ &= P(A|C). \quad \blacksquare \end{aligned}$$

Problem 3: Expecting something different? (exercise)

Let X be a non-negative, discrete, integer-valued random variable. Let $A = \{(i,j) \in \mathbb{N} \times \mathbb{N} \mid i \geq 1, j < i\} = \{(i,j) \in \mathbb{N} \times \mathbb{N} \mid j \geq 0, i > j\}$. By definition of expected value,

$$\begin{aligned} E(X) &= \sum_{i=0}^{\infty} iP(X=i) = \sum_{i=1}^{\infty} iP(X=i) \\ &= \sum_{i=1}^{\infty}\sum_{j=0}^{i-1} P(X=i) = \sum_{(i,j)\in A} P(X=i) = \sum_{j=0}^{\infty}\sum_{i=j+1}^{\infty} P(X=i) \\ &= \sum_{j=0}^{\infty} P(X>j). \qquad \blacksquare \end{aligned}$$

Problem 4: Big data (exercise)

A. Let n be the number of files in the database, and let $\{F_i\}_{i=1}^{n}$ be a decreasing sequence of the file sizes in the database. Suppose, for sake of contradiction, that, for some $m \geq \frac{n}{2}$, for $i \in \{1, 2, \ldots, m\}$, $F_i > 12K$. Then, the average file size A of the files in the database is given by:

$$A = \frac{1}{n}\sum_{i=1}^{n} F_i \geq \frac{1}{n}\sum_{i=1}^{m} F_i > \frac{1}{n}\sum_{i=1}^{m} 12K = \frac{1}{n} m \cdot 12K \geq \frac{1}{n}\frac{n}{2} 12K = 6K,$$

contradicting the given that $A = 6K$. $\blacksquare$

B. Let n and $\{F_i\}_{i=1}^{n}$ be as in the solution to part A. Suppose, for sake of contradiction, that, for some $m \geq \frac{n}{3}$, for some $i \in \{1, 2, \ldots, m\}$, $F_i > 12K$. Then, the average file size A of the files in the database is given by:

$$\begin{aligned} A &= \frac{1}{n}\sum_{i=1}^{n} F_i = \frac{1}{n}\left(\sum_{i=1}^{m} F_i + \sum_{i=m+1}^{n} F_i\right) \\ &> \frac{1}{n}\left(\sum_{i=1}^{m} 12K + \sum_{i=m+1}^{n} 3K\right) \\ &= \frac{1}{n}\left(m \cdot 12K + (n-m) \cdot 3K\right) = \frac{1}{n}\left(m \cdot (9K) + n \cdot 3K\right) \\ &\geq \frac{1}{n}\left(\frac{n}{3}(9K) + n \cdot 3K\right) = 6K, \end{aligned}$$

contradicting the given that $A = 6K$. Therefore, at fewer than a third of the files can have size $> 12K$. $\blacksquare$

Problem 5: Making a stack of coins out of fish

Let $\lambda = pn$, so that $p = \lambda/n$. Let X be a random variable such that $X \sim \text{Binomial}(n, p)$. Since

$$\lim_{t\to\infty}\left(\frac{n!}{n^{k+i}(n-k-i)!}\right) = 1,$$

by the Binomial Theorem,

$$\begin{aligned}
\lim_{t\to\infty}\left(\frac{n!}{(n-k)!n^k}(1-p)^{n-k}\right) &= \lim_{t\to\infty}\left(\frac{n!}{(n-k)!n^k}\left(1-\frac{-\lambda}{n}\right)^{n-k}\right) \\
&= \lim_{t\to\infty}\left(\frac{n!}{(n-k)!n^k}\sum_{i=0}^{n-k}\binom{n-k}{i}\left(\frac{-\lambda}{n}\right)^i\right) \\
&= \lim_{t\to\infty}\left(\sum_{i=0}^{n-k}\frac{(n-k)!}{(n-k)!}\frac{n!}{n^{k+i}(n-k-i)!}\frac{(-\lambda)^i}{i!}\right) \\
&= \sum_{i=0}^{\infty}\frac{(-\lambda)^i}{i!} = e^{-\lambda}.
\end{aligned}$$

$$\begin{aligned}
\lim_{t\to\infty}(P_X(k)) &= \lim_{t\to\infty}\left(\binom{n}{k}p^k(1-p)^{n-k}\right) \\
&= \lim_{t\to\infty}\left(\frac{\lambda^k n!}{k!(n-k)!n^k}(1-p)^{n-k}\right) \\
&= \frac{\lambda^k e^{-\lambda}}{k!}.
\end{aligned}$$

Thus, for large n, the binomial distribution $\text{Binomial}(n, p)$ is well-approximated by the Poisson distribution $\text{Poisson}(np)$. ■

Problem 6: Coffee-theorem automata

Let X_1 be a random variable denoting the time (in hours) the student takes to get home from work, and let X_2 be a random variable denoting the time (in hours) the student takes to get home from the coffee shop. Let $W = E(X_1)$, and let $C = E(X_2)$. Then, conditioning on whether the student goes home or goes to get coffee,

$$W = 1 * \frac{1}{3} + (1 + C) * \frac{2}{3}.$$

Similarly, conditioning on whether the student goes back to work or stays at the coffee house,

$$C = (1 + W) * \frac{1}{3} + (1 + C) * \frac{2}{3}.$$

Since this gives a system of two linear equations in two variables, we can solve the system, yielding $W = 9, C = 12$. Thus, the expected time until the student goes home is $\boxed{\text{9 hours.}}$

Problem 7: Expecting to be astonished

A. Suppose that X is a constant random variable, taking only a single value x_1 with $P(X = x_1) = 1$. Then, $a(P(X = x_1)) = a(1) = 0$, so that $C(X) = P(X = x_1) \cdot a(P(X = x_1)) = 0$. Suppose, on the other hand, that X takes at least two values, including some distinct x_1 and x_2, so that $0 < P(X = x_1), P(X = x_2) < 1$. Then, $C(X) \geq P(X = x_1) \cdot a(P(X = x_1)) + P(X = x_2) \cdot a(P(X = x_2)) > 1$, since $a > 0$ on $(0, 1)$. Thus, $C(X)$ is non-negative, and zero if and only if X is a constant random variable. ■

B. Since

$$\sum_{j=1}^{m}\sum_{i=1}^{n} P(X = x_i, Y = y_i)a(P(Y = y_i)) = \sum_{j=1}^{m} P(Y = y_i)a(P(Y = y_i)) = C(Y),$$

$$C(X, Y) = \sum_{i=1}^{n}\sum_{j=1}^{m} P(X = x_i, Y = y_i) \cdot a(P(X = x_i, Y = y_i))$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{m} P(X = x_i, Y = y_i) \cdot a(\frac{P(X = x_i, Y = y_i)P(Y = y_i)}{P(Y = y_i)})$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{m} P(X = x_i, Y = y_i) \cdot \log_2(\frac{P(Y = y_i)}{P(X = x_i, Y = y_i)P(Y = y_i)})$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{m} P(X = x_i, Y = y_i) \cdot \left(\log_2(P(Y = y_i)) + \log_2\left(\frac{P(X = x_i, Y = y_i)}{P(Y = y_i)}\right)\right)$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{m} P(X = x_i, Y = y_i) \cdot (a(P(Y = y_i)) + a(P(X = x_i|Y = y_i)))$$

$$= \sum_{j=1}^{m}\sum_{i=1}^{n} P(X = x_i, Y = y_i) \cdot (a(P(Y = y_i)) + a(P(X = x_i|Y = y_i)))$$

$$= \sum_{j=1}^{m}\sum_{i=1}^{n} P(X = x_i, Y = y_i) \cdot a(P(X = x_i|Y = y_i)) + \sum_{i=1}^{n} P(X = x_i, Y = y_i) \cdot a(P(Y = y_i))$$

$$= C(X|Y) + C(Y). \quad ■$$

C. $1/p(X)$ is a random variable because it is the outcome the experiment of applying the function $X \mapsto \log_2 1/p(X)$ to X, the outcome of an experiment. Furthermore, $1/p(X)$ can take at most as many values as X, and thus can take only finitely (and thus countably) many values, so that $1/p(X)$ is discrete. Since $X \log_2 1/p(X) = \log_2 1/p(x)$ whenever $X = x$, $E(\log_2 1/p(X)) = \sum_{x\in\Omega} P(\log_2 1/p(X) = \log_2 1/p(x)) \cdot (\log_2 1/p(x)) = \sum_{x\in\Omega} P(X = x) \cdot (\log_2 1/p(x)) = C(X)$. ■

D. Since $y \mapsto \log_2 1/y$ is a convex, decreasing function, by Jensen's Inequality, $C(X) = E(\log_2 1/p(X)) \geq \log_2 1/E(p(X))$, so that $C(X)$ is maximized when $X = \arg\min_{x \in \Omega} p(x)$, the result in the sample space of least probability. ■

Problem 8: You may call a k-clause a Klaus

Let $n < 2^{k-1}$ be the number of clauses in the given k-CNF. Suppose we randomly assign values to each of the variables in the given k-CNF. Let X be the event that some clause contains contains all only true or only false variables. For $i \in \{1, 2, \ldots, n\}$, let X_i be the event that the i^{th} k-clause contains only true or only false variables, so that $X = \bigcup_{i=1}^{n} X_i$. Then, for each $i \in \{1, 2, \ldots, n\}$,

$$P(X_i) = \frac{2}{2^k} = \frac{1}{2^{k-1}}.$$

Thus,

$$P(X) = P\left(\bigcup_{i=1}^{n} X_i\right) \leq \sum_{i=1}^{n} P(X_i) = \sum_{i=1}^{n} \frac{1}{2^{k-1}} = n\frac{1}{2^{k-1}} < \frac{2^{k-1}}{2^{k-1}} = 1.$$

If Y is the event that every k-clause includes at least one true variable and at least one false variable. Then, $Y = X^c$, so that $P(Y) = 1 - P(X) > 0$. Since there is a non-zero probability that, under a random assignment of the variables in the given k-CNF, every clause includes at least one true variable and at least one false variable, there exists an assignment of the variables in the given k-CNF with the desired condition. ■

Problem 9: Shearing off projections (extra credit)
