Stat 230

Homework Problem Set 7

Due Jun 21 7 pm EST.

Please justify your answers and show the steps that lead you to your answer. Without a proper explanation, even a perfectly correct answer will receive a low score. Doubts regarding the problem sets/ Problem setup can be posted on Piazza. Simplify the expressions whenever you can.

Pr. 1 (20 points, 5 each)

Let (X, Y) follow the bivariate normal distribution with $E[X] = -1, E[Y] = 1, \mathrm{Var}(X) = 4, \mathrm{Var}(Y) = 25, \mathrm{Cov}(X, Y) = -5$.

Find:

a) The marginal distributions of X and Y

b) The correlation of X and Y

c) The conditional distribution of X given Y

d) The conditional distribution of Y given X

Pr. 2 (20 points)

Gibbs' inequality is an important inequality related to information theory.

Theorem 0.1 (Gibbs' inequality)**.** *For discrete probability distributions $P = (p_1, \ldots, p_n)$, with $p_i := \Pr(X_p = i)$ and $Q = (q_1, \ldots, q_n)$, with $q_i := \Pr(X_q = i)$, the following inequality holds:*

$$-\sum_{i=1}^{n} p_i \log(p_i) \leq -\sum_{i=1}^{n} p_i \log(q_i).$$

a) (15 points) Prove Gibbs' inequality. Hint: Jensen

b) (5 points) The Kullback-Leibler divergence is a measure of how different two probability distributions are. Show that the Kullback-Leibler divergence, defined as

$$D_{KL}(P, Q) := \sum_{i=1}^{n} p_i \log\left(\frac{p_i}{q_i}\right)$$

of two discrete distributions P and Q is bounded from below by 0.

Pr. 3 (20 points)

a) (5 points) Let $X_i \sim^{\text{iid}} \mathrm{Exp}(1)$, $i = 1, 2, \ldots,$, and $\bar{X}_n := \frac{\sum_{i=1}^{n} X_i}{n}$ argue that for any x

$$\Pr\left(\frac{\bar{X}_n - 1}{1/\sqrt{n}} \leq x\right) \to \Pr(Z \leq x),$$

where Z follows the standard normal distribution.

b) (15 points) By differentiating the equation in a) with respect to x, and evaluating the resulting expression at $x = 0$, prove the *Stirling approximation* to the factorial:

$$n! \approx \sqrt{2\pi n}\left(\frac{n}{e}\right)^n.$$

Hint: The sum of iid exponential random variables is a distribution we are familiar with. Also remember the fundamental theorem of calculus: if $F(0) = 0$ and $f(x)$ is the derivative of F, then for $x \geq 0$

$$\frac{d}{dx}F(g(x)) = \frac{d}{dx}\int_0^{g(x)} f(t)\ dt = f(g(x))g'(x)$$

Pr. 4 (15 points)

Markov himself used his namesake chains to analyze Russian literature back in 1913[1]. He took an extract from "Eugene Onegin"by A. S. Pushkin. He found that a consonant is followed by a vowel about 87.2 percent of the time, while a vowel is followed by a consonant 66.3 percent of the time.

a) (5 points) What does this problem have to do with Markov chains?

b) (10 points) What is the proportion of vowels in Russian high quality literature?

Pr. 5 (25 points)

Consider a game with the following structure: There are 3 levels. To advance to level $n+1$,you have to win at level n. If you win at level 3, you win the game and get a fabulous prize. If you fail at level 1, the game ends completely. If you are at level 2 or 3 and fail, then you move to the beginning of preceding level (i.e. level 1 or 2, respectively). There is no skill involved: winning a level is determined by chance. Starting at level 1, the probability of winning level 1 is $1/2$. Starting at level 2, the probability of winning at level 2 is $1/4$. Starting at level 3, the probability of winning at level 3 (and getting the prize) is $1/8$. If you have to repeat a level, the probability of winning at that level is unchanged (nothing is gained from your previous experience with the level).

a) (10 points) Identify this game with a Markov chain. Find the transition probability matrix and represent the chain graphically

b) (5 points) What are the absorbing states of this Markov chain?

c) (10 points) What is the probability that a person starting at level 1 will eventually win the prize? A good approximation suffices.

References

[1] Марков, А.А. *Пример статистического исследования над текстом "Евгения Онегина", иллюстрирующий связь испытаний в цепь* `http://books.e-heritage.ru/book/10086570`