TA Lecture 12 - Statistical Inference 2

May 29 - 30

School of Information Science and Technology,
ShanghaiTech University

Outline

HW Problems

Problem 1

Show the following orthogonality properties of MMSE: (a) For any function $\phi(\cdot)$, one has

$$E[(Y - E[Y \mid X])\phi(X)] = 0.$$

(b) If the function $g(X)$ satisfied

$$E[(Y - g(X))\phi(X)] = 0, \forall \phi(\cdot).$$

then $g(X) = E(Y \mid X)$.

Problem 1 Solution

Problem 2

Let X and Y be independent, positive r.v.s. with finite expected values.

(a) Give an example where $E\left(\frac{X}{X+Y}\right) \neq \frac{E(X)}{E(X+Y)}$, computing both sides exactly. Hint: Start by thinking about the simplest examples you can think of!

(b) If X and Y are i.i.d., then is it necessarily true that $E\left(\frac{X}{X+Y}\right) = \frac{E(X)}{E(X+Y)}$?

(c) Now let $X \sim \text{Gamma}(a, \lambda)$ and $Y \sim \text{Gamma}(b, \lambda)$. Show without using calculus that

$$E\left(\frac{X^c}{(X+Y)^c}\right) = \frac{E\left(X^c\right)}{E\left((X+Y)^c\right)}$$

for every real $c > 0$.

Problem 2 Solution

Problem 3

A DNA sequence can be represented as a sequence of letters, where the "alphabet" has 4 letters: A,C,T,G. Suppose such a sequence is generated randomly, where the letters are independent and the probabilities of $\mathrm{A}, \mathrm{C}, \mathrm{T}, \mathrm{G}$ are p_1, p_2, p_3, p_4 respectively.

(a) In a DNA sequence of length 115 , what is the expected number of occurrences of the expression "CATCAT" (in terms of the p_j)? (Note that, for example, the expression "CATCATCAT" counts as 2 occurrences.)

(b) What is the probability that the first A appears earlier than the first C appears, as letters are generated one by one (in terms of the p_j)?

Problem 3 cont.

(c) For this part, assume that the p_j are unknown. Suppose we treat p_2 as a Unif $(0, 1)$ r.v. before observing any data, and that then the first 3 letters observed are "CAT". Given this information, what is the probability that the next letter is C ?

Problem 3 Solution

Problem 4

A coin with probability p of Heads is flipped repeatedly. For (a) and (b), suppose that p is a known constant, with $0 < p < 1$.

(a) What is the expected number of flips until the pattern HT is observed?

(b) What is the expected number of flips until the pattern HH is observed?

(c) Now suppose that p is unknown, and that we use a $\text{Beta}(a, b)$ prior to reflect our uncertainty about p (where a and b are known constants and are greater than 2). In terms of a and b, find the corresponding answers to (a) and (b) in this setting.

Problem 4 Solution

Problem 5

(a) Let $p \sim \text{Beta}(a, b)$, where a and b are positive real numbers. Find $E\left(p^2(1-p)^2\right)$, fully simplified (Γ should not appear in your final answer).

Two teams, A and B, have an upcoming match. They will play five games and the winner will be declared to be the team that wins the majority of games. Given p, the outcomes of games are independent, with probability p of team A winning and $(1-p)$ of team B winning. But you don't know p, so you decide to model it as an r.v., with $p \sim \text{Unif}(0, 1)$ a priori (before you have observed any data).

Problem 5 Continued

To learn more about p, you look through the historical records of previous games between these two teams, and find that the previous outcomes were, in chronological order, $AAABBAABAB$. (Assume that the true value of p has not been changing over time and will be the same for the match, though your beliefs about p may change over time.)

(b) Does your posterior distribution for p, given the historical record of games between A and B, depend on the specific order of outcomes or only on the fact that A won exactly 6 of the 10 games on record? Explain.

Problem 5 Continued

(c) Find the posterior distribution for p, given the historical data.

The posterior distribution for p from (c) becomes your new prior distribution, and the match is about to begin!

(d) Conditional on p, is the indicator of A winning the first game of the match positively correlated with, uncorrelated with, or negatively correlated of the indicator of A winning the second game of the match? What about if we only condition on the historical data?

(e) Given the historical data, what is the expected value for the probability that the match is not yet decided when going into the fifth game (viewing this probability as an r.v. rather than a number, to reflect our uncertainty about it)?

Problem 5 Solution

Problem 6

(Optional Challenging Problem) Use two different methods to show that if X and Y are jointly Normal random variables, then

$$E[Y \mid X] = L[Y \mid X] = E(Y) + \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(X)}(X - E(X)).$$

Problem 6 Solution

(Optional Challenging Problem) Use two different methods to show that if X and Y are jointly Normal random variables, then

$$E[Y \mid X] = L[Y \mid X] = E(Y) + \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(X)}(X - E(X)).$$