Week 5 Live Session

w203 Instructional Team

Homework 4 Presentation

Joint Distributions Practice

Professorial Mistakes

In a live session, the number of questions that students ask is a random variable, X. X takes on three values: 0, 1, and 2, each with probability 1/3. Every time a student asks a question, Professor Paul Laskowski answers incorrectly with probability 1/4, independently of other questions. Let Y be the random variable representing the number of incorrect responses in the live session.

a. Compute the expectation of Y, conditional on X.
b. Using the law of iterated expectations, compute $E(Y)$.
c. Describe the joint probability distribution of X and Y. (You may find it easiest to use a table)
d. Compute the expectation of the product of X and Y, $E(XY)$
e. Using the previous result, compute $cov(X, Y)$.

Triangular Regions

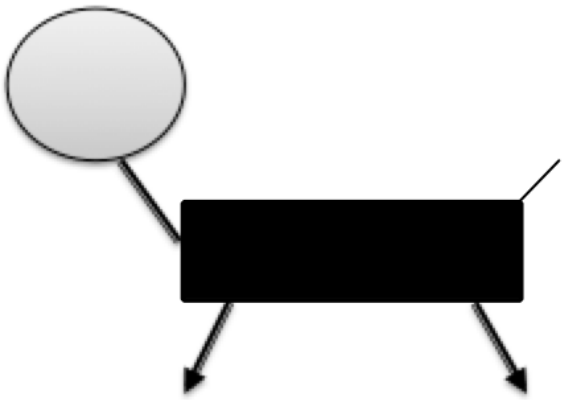
Continuous random variables X and Y have a joint distribution with probability density function,

$$f(x, y) = \begin{cases} 1, & 0 < y < 1, a \cdot y < x < a \cdot y + 1 \\ 0, & otherwise. \end{cases}$$

where a is a constant.

a. Choose 2 example values for a and draw a graph of the region for which X and Y have positive probability density.
b. Derive the marginal distribution of Y.
c. Compute the conditional expectation of X, conditional on Y.
d. As a slight variation on the previous part, compute $E(XY|Y)$. Note that since we're conditioning on Y, the Y inside the expectation is just a constant.
e. Derive $cov(X, Y)$. Hint: an nice way to do this is to use the law of iterated expectations. Write down the definition of covariance, then break the expectation up into two expectations. The inner expecation should be conditional on Y, and the outer expectation should be unconditional. The results you derived above should help you finish the proof.
f. Check what $cov(X, Y)$ equals when $a = 0$. What is $cov(X, Y)$ when $a = -1$?
g. Choose an example value for a, then use R to simulate 100 draws from the given joint distribution and plot them.

Discussion on Models



Reading:

The End of Theory: The Data Deluge Makes the Scientific Method Obsolete
http://archive.wired.com/science/discoveries/magazine/16-07/pb_theory

> All models are wrong, but some are useful. - George Box, Statistician
>
> All models are wrong, and you can increasingly succeed without them. - Peter Norvig, Google's Research Director

Setting aside the over-dramatic tone of Anderson's piece, there are many elements of his argument worth discussing:

1. Are the fancy algorithms Anderson points to actually model-free? For example, is the page rank algorithm model-free? What is a model?
2. What is the real difference between parametric statistics and learning algorithms that originate in a computer science tradition?
3. In part, Anderson argues that we don't need models that are understandable by humans. When is it enough to have a model that "fits" the data well, and when do we really need it to be human-readable?
4. When does context matter in data science? Is domain knowledge important or can we just rely on the numbers in a data set?
5. Anderson seems to imply that we don't need to test hypotheses today. Are hypotheses still relevant to data science, or should our focus be on estimating effect sizes, or using our models to classify and predict?