Chris Piech
CS109

Lecture #13
Oct 23$^{\text{rd}}$, 2017

Properties of Joint Distributions

Expectation with Multiple RVs

Expectation over a joint isn't nicely defined because it is not clear how to compose the multiple variables. However, expectations over functions of random variables (for example sums or multiplications) are nicely defined: $E[g(X,Y)] = \sum_{x,y} g(x,y)p(x,y)$ for any function $g(X,Y)$. When you expand that result for the function $g(X,Y) = X+Y$ you get a beautiful result:

$$\begin{aligned}
E[X+Y] = E[g(X,Y)] &= \sum_{x,y} g(x,y)p(x,y) = \sum_{x,y}[x+y]p(x,y) \\
&= \sum_{x,y} xp(x,y) + \sum_{x,y} yp(x,y) \\
&= \sum_{x} x \sum_{y} p(x,y) + \sum_{y} y \sum_{x} p(x,y) \\
&= \sum_{x} xp(x) + \sum_{y} yp(y) \\
&= E[X] + E[Y]
\end{aligned}$$

This can be generalized to multiple variables:

$$E\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} E[X_i]$$

Independence with Multiple RVs

Discrete

Two discrete random variables X and Y are called independent if:

$$P(X = x, Y = y) = P(X = x)P(Y = y) \text{ for all } x, y$$

Intuitively: knowing the value of X tells us nothing about the distribution of Y. If two variables are not independent, they are called dependent. This is a similar conceptually to independent events, but we are dealing with multiple *variables*. Make sure to keep your events and variables distinct.

Continuous

Two continuous random variables X and Y are called independent if:

$$P(X \leq a, Y \leq b) = P(X \leq a)P(Y \leq b) \text{ for all } a, b$$

This can be stated equivalently as:

$$F_{X,Y}(a,b) = F_X(a)F_Y(b) \text{ for all } a, b$$
$$f_{X,Y}(a,b) = f_X(a)f_Y(b) \text{ for all } a, b$$

More generally, if you can factor the joint density function then your continuous random variable are independent:

$$f_{X,Y}(x,y) = h(x)g(y) \text{ where } -\infty < x, y < \infty$$

Example 2

Let N be the # of requests to a web server/day and that $N \sim Poi(\lambda)$. Each request comes from a human (probability = p) or from a "bot" (probability = $(1–p)$), independently. Define X to be the # of requests from humans/day and Y to be the # of requests from bots/day.

Since requests come in independently, the probability of X conditioned on knowing the number of requests is a Binomial. Specifically:

$$(X|N) \sim Bin(N, p)$$
$$(Y|N) \sim Bin(N, 1-p)$$

Calculate the probability of getting exactly i human requests and j bot requests. Start by expanding using the chain rule:

$$P(X = i, Y = j) = P(X = i, Y = j | X + Y = i + j) P(X + Y = i + j)$$

We can calculate each term in this expression:

$$P(X = i, Y = j | X + Y = i + j) = \binom{i+j}{i} p^i (1-p)^j$$
$$P(X + Y = i + j) = e^{-\lambda} \frac{\lambda^{i+j}}{(i+j)!}$$

Now we can put those together and simplify:

$$P(X = i, Y = j) = \binom{i+j}{i} p^i (1-p)^j e^{-\lambda} \frac{\lambda^{i+j}}{(i+j)!}$$

As an exercise you can simplify this expression into two independent Poisson distributions.

Symmetry of Independence

Independence is symmetric. That means that if random variables X and Y are independent, X is independent of Y and Y is independent of X. This claim may seem meaningless but it can be very useful. Imagine a sequence of events $X_1, X_2, \ldots$. Let A_i be the event that X_i is a "record value" (eg it is larger than all previous values). Is A_{n+1} independent of A_n? It is easier to answer that A_n is independent of A_{n+1}. By symmetry of independence both claims must be true.

Conditional Distributions

Before we looked at conditional probabilities for events. Here we formally go over conditional probabilities for random variables. The equations for both the discrete and continuous case are intuitive extensions of our understanding of conditional probability:

Discrete

The conditional probability mass function (PMF) for the discrete case:

$$p_{X|Y}(x|y) = P(X = x | Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)} = \frac{P_{X,Y}(x,y)}{p_Y(y)}$$

The conditional cumulative density function (CDF) for the discrete case:

$$F_{X|Y}(a|y) = P(X \leq a | Y = y) = \frac{\sum_{x \leq a} p_{X,Y}(x,y)}{p_Y(y)} = \sum_{x \leq a} p_{X|Y}(x|y)$$

Continuous

The conditional probability density function (PDF) for the continuous case:

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$$

The conditional cumulative density function (CDF) for the continuous case:

$$F_{X|Y}(a|y) = P(X \leq a|Y = y) = \int_{-\infty}^{a} f_{X|Y}(x|y)dx$$

Example 2

Let's say we have two independent random Poisson variables for requests received at a web server in a day: X = # requests from humans/day, $X \sim Poi(\lambda_1)$ and Y = # requests from bots/day, $Y \sim Poi(\lambda_2)$. Since the convolution of Poisson random variables is also a Poisson we know that the total number of requests $(X+Y)$ is also a Poisson $(X+Y) \sim Poi(\lambda_1+\lambda_2)$. What is the probability of having k human requests on a particular day given that there were n total requests?

$$\begin{aligned}
P(X = k|X+Y = n) &= \frac{P(X = k, Y = n-k)}{P(X+Y = n)} = \frac{P(X = k)P(Y = n-k)}{P(X+Y = n)} \\
&= \frac{e^{-\lambda_1}\lambda_1^k}{k!} \cdot \frac{e^{-\lambda_2}\lambda_2^{n-k}}{(n-k)!} \cdot \frac{n!}{e^{1(\lambda_1+\lambda_2)}(\lambda_1+\lambda_2)^n} \\
&= \binom{n}{k} \left(\frac{\lambda_1}{\lambda_1+\lambda_2}\right)^k \left(\frac{\lambda_2}{\lambda_1+\lambda_2}\right)^{n-k} \\
&\sim Bin\left(n, \frac{\lambda_2}{\lambda_1+\lambda_2}\right)
\end{aligned}$$