Lecture 2, Jan 12, 2024

Joint Random Variables

- A *random variable* is a function that assigns one or more numbers to the outcome of an experiment
  - Random numbers can be multi-dimensional: $\boldsymbol{X}\colon s \mapsto \mathbb{R}^2 \iff \boldsymbol{X}(s) = (X(s), Y(s))$
- The probability mass function is denoted $P[X = x_i, Y = y_i] = p_{\boldsymbol{X}}(x_i, y_i)$ for discrete random variables
  - Probability of a set/event is the sum of the PMF over the events
- The probability density function is denoted $P[x < X < x + \mathrm{d}x, y < Y < y + \mathrm{d}y] \approx f_{\boldsymbol{X}}(x, y)\,\mathrm{d}x\,\mathrm{d}y$ for continuous random variables
  - Probability of a set/event is the integral of the PDF over the continuous region that defines the event
  - Note we denote PMFs by p, PDFs by f
- *Marginal probabilities* can be computed as $p_Y(y_j) = \sum_j p_{\boldsymbol{X}}(x_i, y_j), p_X(x_i) = \sum_j p_{\boldsymbol{X}}(x_i, y_j)$ (discrete)
  - $f_X(x) = \int_{-\infty}^{\infty} f_{\boldsymbol{X}}(x, y')\,\mathrm{d}y', f_Y(y) = \int_{-\infty}^{\infty} f(x', y)\,\mathrm{d}x'$
  - In isolation the marginals don't have all the information that the joint PMF provides
- Conditional probabilities are given by $p_{Y|X}(y_j|x_i) = \dfrac{p_{\boldsymbol{X}}(x_i, y_j)}{p_X(x_i)}, f_{Y|X}(y|x) = \dfrac{f_{\boldsymbol{X}}(x, y)}{f_X(x)}$
  - The discrete version follows directly from the definitions
  - The continuous version requires a limiting procedure
  - Rearranging gives the *product rule*: $p_{\boldsymbol{X}}(x_i, y_j) = p_{Y|X}(y_j|x_i)p_X(x_i) = p_{X|Y}(x_i|y_j)p_Y(y_j)$ (same with continuous version)

Expectation, Mean and Variance

- The *expected value* of a function $Z = g(X, Y)$ is $E[Z] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x', y')f_{\boldsymbol{X}}(x', y')\,\mathrm{d}x'\,\mathrm{d}y'$
  - For a function dependent on only one of the variables, this is equivalent to integrating on the marginal
- The *mean* is simply $m_X = E[X] = \int_{-\infty}^{\infty} x' f_X(x')\,\mathrm{d}x'$
- The *variance* is defined as $\sigma_X^2 = \mathrm{Var}[X] = E[(X - E[X])^2]$
  - This is a measure of spread
  - Expanding this out gives $\sigma_X^2 = E[X^2] - (E[X])^2$
- $$\begin{aligned} E[g(Y)] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(y')f_{Y|X}(y'|x')f_X(x')\,\mathrm{d}y\,\mathrm{d}x \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(y')f_{Y|X}(y'|x')\,\mathrm{d}y f_X(x')\,\mathrm{d}x \\ &= \int_{-\infty}^{\infty} E[g(Y)|X = x]f_X(x')\,\mathrm{d}x' \\ &= E[E[g(Y)|X]] \end{aligned}$$
  - In other words we can find the expectation of $g(Y)$ assuming X is known, and then find the expectation of that over X, to find the overall expectation of $g(Y)$
  - Special case: if $g(Y) = Y$ then $E[Y] = E[E[Y|X]]$
  - Example: picking X from a uniform $[0, 1]$, and then picking Y from a uniform $[0, x]$
    - $E[Y] = E[E[Y|X]] = E\left[\dfrac{X}{2}\right] = \dfrac{E[X]}{2} = \dfrac{1}{4}$
- The *covariance* of X and Y is defined as $\sigma_{XY} = \mathrm{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])]$
  - If X and Y tend to vary positively together, the covariance is positive; if one varies positively while the other varies negatively,the covariance is negative; if there is no relation, the covariance is zero
  - Expanding gives $E[XY] - E[X]E[Y]$ ($E[XY]$ is known as the *correlation*)

  - Normalizing gives the *correlation coefficient* $\rho_{XY} = E\left[\left(\frac{X - m_X}{\sigma_x}\right)\left(\frac{Y - m_Y}{\sigma_Y}\right)\right] = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y}$
  - X and Y are *uncorrelated* if $\rho_{XY} = 0 \iff \mathrm{Cov}(X, Y) = 0$ (note uncorrelated does not always imply independent)
  - Note covariance is bilinear (i.e. linear in each argument)
- X, Y are independent if $f_{\boldsymbol{X}}(x, y) = f_X(x) f_Y(y)$ or $p_{\boldsymbol{X}}(x_i, y_j) = p_X(x_i) p_Y(y_j)$
  - Independence means $E[g(X)h(Y)] = E[g(X)]E[h(Y)]$
  - This also means $\mathrm{Cov}(X, Y) = 0$ (i.e. independence implies uncorrelated)
  - $f_{X|Y}(x|y) = \frac{f_X(x) f_Y(y)}{f_Y(y)} = f_X(x)$
    * The *a posteriori* distribution is the same as the *a priori* distribution
    * i.e. knowing one does not give any information about the other