AM207 - Lecture 01 Probability Review

What is a probability? There are two ways to answer:

① It is a number that tells you how often some event will happen

⇒ Frequentist view

② It is a number that expresses our belief that something would happen

⇒ Bayesian view

Both views are useful and we will take whichever is more convenient. No need to join the Bayes vs. Frequentist debate.

For an event X: $P(X) = 1$ ⇒ it will happen

$P(X) = 0$ ⇒ it will not happen

$0 \leq P(X) \leq 1$ ⇒ it might or might not happen

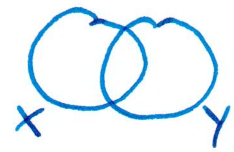
X^- : opposite of event X

$P(X) + P(X^-) = 1$ ⇒ Either X happens or not

⇒ Normalized

For two events X and Y:

$$P(X+Y) = P(X) + P(Y) - P(X,Y)$$



$P(X,Y)$: joint probability

$P(X)$: marginal probability

$P(X|Y)$: conditional probability

Marginal sum rule: $P(X) = \sum_{Y} P(X,Y)$

Product rule: $$P(X,Y) = P(X|Y) \cdot P(Y)$$
$$= P(Y|X) \cdot P(X)$$

Note: In general it is wrong to assume $P(X) \approx P(X|Y)$

Example: X: Patient dies of disease

Y: Patient goes to hospital

A doctor might think $P(X)$ is high, but he/she is only observing $P(X|Y)$.

For three events: X, Y, Z

$$P(X,Y,Z) = P(X|Y,Z) \cdot P(Y,Z) = P(X|Y,Z) \cdot P(Y|Z) \cdot P(Z)$$

This is going to be handy when we talk about graphical models.

Bayes rule:

$$P(X,Y) = P(X,Y)$$

$$P(Y|X) \cdot P(X) = P(X|Y) \cdot P(Y)$$

$$P(Y|X) = \frac{P(X|Y) \cdot P(Y)}{P(X)}$$

If you don't have P(X):

$$P(X) = \sum_Y P(X,Y) = \sum_Y P(X|Y) \cdot P(Y)$$

Some vocabulary: $\text{posterior} = \frac{\text{likelihood} \cdot \text{prior}}{\text{evidence}}$

We will later often omit the evidence and use

$$P(Y|X) \propto P(X|Y) \cdot P(Y)$$

In 1999 Sally Clark was found guilty of the murder of her two infant sons. Both sons died within a few weeks of their birth, one in 1996 and one in 1998.

Prosecuter's argument:

- prob. of SIDS: $\frac{1}{8500}$
- prob. of two deaths: $\left(\frac{1}{8500}\right)^2$
- prob. of Clark's innocense: one in 73 million

Mistakes:

① Independence assumption. There might have been a genetic cause

② $P(\text{innocense} \mid \text{evidence}) \neq P(\text{evidence} \mid \text{innocense})$!

Bayes theorem:

$$P(\text{innocense} \mid \text{evidence}) = \frac{P(\text{evidence} \mid \text{innocense}) \cdot P(\text{innocense})}{P(\text{evidence})}$$

While the likelihood is low, the prior should be high!

Unfortunately Clark spen three years in jail before her conviction was overturned.

Independence:

$$P(X, Y) = P(X) \cdot P(Y)$$

Note: X and Y being uncorrelated does not imply independent.

Example: $X \sim N(0,1)$ and $Y = X^2$ then

$$\mathrm{Cov}(X,Y) = E[XY] - E[X] \cdot E[Y]$$

$$= 0 - 0 = 0$$

as $E[XY] = E[X^3] = 0$

But, X and Y are dependent by definition! (But the dependence is not linear)

If X and Y are independent, then

$$P(X \mid Y) = \frac{P(X,Y)}{P(Y)} = \frac{P(X) \cdot P(Y)}{P(Y)} = P(X)$$

$\Rightarrow$ Y tells us nothing about X.

Conditional independence:

$$P(X, Y \mid Z) = P(X \mid Z) \cdot P(Y \mid Z)$$

Note: Conditional independence does not imply independence!

Example: Play online game against two possible opponents

opponent A: evenly matched

opponent B: 0.75 chance of you winning

X: You win the first game

Y: You win the second game

Z: playing against A

Then $X|Z=1$ and $Y|Z=1$ are both independent and Bernoulli (0.5).

Same holds for $X|Z=0$ and $Y|Z=0$ with Bernoulli (0.75)

But X and Y are not independent! Observing $X=1$ makes it more likely that you played against B.

$P(Y=1 \mid X=1) > P(Y=1)$ as the past game gave us some information who your opponent might have been.

Let's talk more about distributions:

For the discrete case, a distribution is a function that takes an event and gives us its probability.

Bernoulli: Probability of a binary outcome. Experiment can be a success or a failure.

$\Rightarrow$ two outcomes: $K=0$ or $K=1$

$$f(K,p) \begin{cases} p & \text{if } K=1 \text{ (success)} \\ 1-p & \text{if } K=0 \text{ (failure)} \end{cases}$$

$$f(K,p) = p^K (1-p)^{1-K}$$

Bernoulli considers only one event. If we have a sequence of events and want the probability that we have K successes out of n trials, we get:

Binomial:

$$P(X=k;n,p) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$$

Poisson: Probability of a number of events within a fixed interval

$$P(X=k;\lambda) = \frac{\lambda^K e^{-\lambda}}{k!}$$

Catch for discrete PMFs: Random variables are not the same as PMFs!

- If we sample from Bernoulli(0.75) we do not get 0.25 or 0.75. We only get 1s and 0s!
- Given a random variable X how do we get the PMF of $2X$? Not by multiplying the PMF of X with a factor of 2. The resulting function would no longer be normalized.

  If X takes on values x_j with probability p_j, then $2X$ takes on values $2x_j$ with probability p_j.

Let's move on to continuous distributions:

One difficulty is that for continuous distributions the probability for a single event collapses to 0. Hence, it only makes sense to talk about ranges.

Normal distribution (Gaussian):

$$f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

Many, many phenomena in nature follow this distribution. Central limit theorem: the distribution of a sum of random variables can be approximated by a normal distribution.

⇒ If you have noisy measurements and the noise has multiple sources then your error bars will be Gaussian.

Exponential: Describes the time between events in a Poisson process. ⇒ Time between events which occur continuously and independently at a constant average rate

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & \text{for } x \geq 0 \\ 0 & \text{for } x < 0 \end{cases}$$

Other PDFs worth looking at: Beta, Gamma