Probability Intro

Random Variables

- A random variable is a mathematical formalization of a quantity or object which depends on random events
  - The full formalization is beyond the scope of this course
- For example, a random variable X capturing a fair coin takes a value of $heads$ with probability 0.5 and $tails$ w.p. 0.5
  - written $\mathbb{P}[X = heads] = \mathbb{P}[X = tails] = 0.5$
- Similarly, a random variable Y capturing a fair die can take a value in $\{1, \dots, 6\}$, each w.p. 1/6
  - written $\mathbb{P}[Y = 1] = \cdots = \mathbb{P}[Y = 6] = 1/6$
- For mathematical convenience, we map discrete event names to numbers, e.g., $heads = 1, tails = 0$

Philosophy and Probability

- The probability of a given event does not tell us what will happen in a specific realization
  - E.g., we don't know what the next coin toss will be
- Let's say we have an event A (e.g., $A = \{X = heads\}$)
  - Suppose $\mathbb{P}[A] = p$
- Interpretation: if we ran the same experiment N times, we would expect A to occur pN times
  - A bit confusing because if we ran the *exact* same experiment, we *should* see the same outcome
  - But there are random factors beyond our control, e.g., wind
- We won't talk about philosophy too much
  - Probability is a nice formalization that has served us well

Random Variables, cont'd

- A random variable can be discrete or continuous
- A discrete variable can take on a finite number of values
  - Coin tosses and dice are discrete variables
- A continuous variable can take on infinitely many values
  - For example, stock prices are continuous

Probability Distribution

- A probability distribution characterizes the probabilities of all values that a variable can take
  - E.g., a coin toss has a binary (aka Bernoulli) distribution with probability 0.5
- Suppose we have a variable $weather$ that can take on values $sun, rain, snow$
  - The probability distribution of $weather$ in Troy is

$$\mathbb{P}[weather = sun] = 0.2$$
$$\mathbb{P}[weather = rain] = 0.2$$
$$\mathbb{P}[weather = snow] = 0.6$$

- Note that all probabilities must sum up to 1

Joint Distributions

- The joint distribution of two random variables X, Y characterizes the probabilities of all pairs of values
- E.g., suppose you have a variable $traffic$ that takes values in $\{low, medium, high\}$
  - The joint distribution of $weather$ and $traffic$ in Troy is

$$\mathbb{P}[weather = sun, traffic = low] = 0.1$$
$$\mathbb{P}[weather = sun, traffic = medium] = 0.06$$
$$\mathbb{P}[weather = sun, traffic = high] = 0.04$$
$$\ldots$$

- All probabilities need to sum up to 1 again
- In some sense, you can imagine we created a new variable $weathertraffic$ that can take all combinations of values

Independent Variables

- Intuitively, two variables X, Y are independent if their probabilities are unaffected by each other
  - E.g., if we toss two coins, we expect one coin to not affect the other
- Mathematically, X, Y are independent if
$$\mathbb{P}[X = a, Y = b] = \mathbb{P}[X = a]\mathbb{P}[Y = b]$$
  - for all possible values a and b
  - E.g., the probability that both coins are $heads$ is the same as the product of each coin being $heads$ independently
- Independence is a critical property in ML and statistics!

Conditional Distribution

- The conditional distribution of X given Y characterizes the probabilities of different values of X for a given value of Y
  - written $\mathbb{P}[X = a|Y = b]$
- For example, we know that

$$\mathbb{P}[weather = sun, traffic = low] = 0.1$$
$$\mathbb{P}[weather = sun, traffic = medium] = 0.06$$
$$\mathbb{P}[weather = sun, traffic = high] = 0.04$$

- This means

$$\mathbb{P}[traffic = low|weather = sun] = 0.5$$
$$\mathbb{P}[traffic = medium|weather = sun] = 0.3$$
$$\mathbb{P}[traffic = high|weather = sun] = 0.2$$

Conditional Distribution, cont'd

- Mathematically, the relationship between conditional and joint distributions is the following:

$$\mathbb{P}[X|Y] = \frac{\mathbb{P}[X,Y]}{\mathbb{P}[Y]}$$

- If Y has occurred, what proportion of the time does X also occur?

Marginalization

- Marginalization is a very useful tool when deriving properties in RL
- Suppose you have two discrete random variables, X and Y
  - i.e., $X \in \{x_1, \dots, x_N\}, Y \in \{y_1, \dots, y_M\}$
- Marginalization is the following property

$$\mathbb{P}[X = x_i] = \sum_{j=1}^{M} \mathbb{P}\left[X = x_i, Y = y_j\right]$$

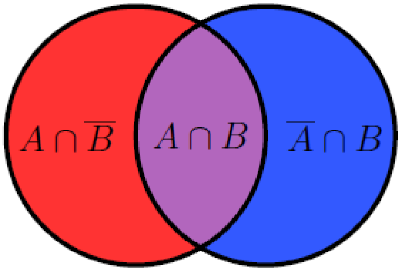
- Intuitively, the probability that X is equal to x_i is the sum of the probabilities of all events where $X = x_i$
  - for all possible values of Y

IID Variables

- Two variables are identically distributed if they have the same distribution
  - Two fair coins are identically distributed
  - A fair coin and a biased coin (e.g. $\mathbb{P}[heads = 0.6]$) are not identically distributed
  - A coin and a die are not identically distributed
- Two variables X, Y are IID if they are independent and identically distributed
- Two fair coin tosses are IID
  - Any number of fair coin tosses are IID
- If you tie two coins with a string, they are not independent, but they are identically distributed

Union Bound

Rensselaer



- Recall the union bound from set theory
- What is the size of $|A \cup B|$?

$$|A \cup B| = |A| + |B| - |A \cap B|$$

- In particular, $|A \cup B| \leq |A| + |B|$
- In general

$$|\cup_i A_i| \leq \sum_i |A_i|$$

- Similarly,

$$\mathbb{P}[\cup_i A_i] \leq \sum_i \mathbb{P}[A_i]$$

Expected value

- Consider a random variable X that can take k possible values, $x_1, \ldots, x_k$, each with probability $p_1, \ldots, p_k$
- The expected value of X is defined as
$$\mathbb{E}[X] = p_1 x_1 + \cdots + p_k x_k$$
  - i.e., it is a weighted average
- If X can take on infinitely many values, the expectation is a bit more involved
- In the case where X is continuous, one may be able to describe $\mathbb{E}[X]$ in terms of its probability density function (pdf):
$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x p(x) dx$$
  - where $p(x)$ is the pdf of X

Expected value, cont'd

- The expected value is a linear operator, i.e.,
$$\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$$
- The expected value of the product of independent variables is just the product of the expectations:
$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$$
- What if the variables are not independent?
  – There is no closed form expression, need to know the joint probabilities:
$$\mathbb{E}[XY] = \sum_{x,y} xy\mathbb{P}[X=x, Y=y]$$

Expectation Examples

Rensselaer

- Suppose you have a fair coin that produces values 0 and 1

$$\mathbb{E}[X] = 0.5 * 0 + 0.5 * 1 = 0.5$$

- Suppose you have a fair die that produces values 1-6

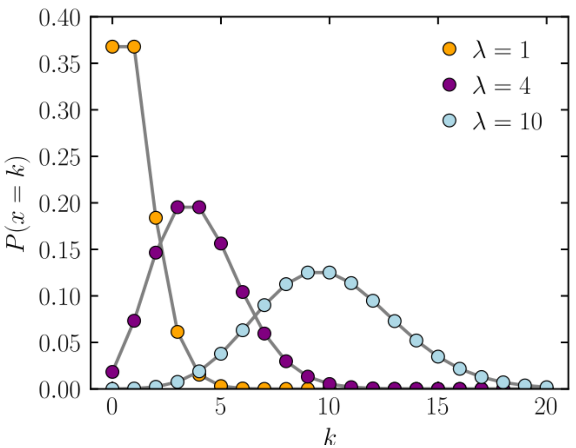
$$\mathbb{E}[X] = \frac{1}{6} * 1 + \cdots + \frac{1}{6} * 6 = \frac{1}{6} * 21 = 3.5$$

- Suppose you have a Poisson distribution where the probability of integer k is

$$\mathbb{P}[X = k] = \frac{\lambda^k e^{-\lambda}}{k!}$$

  – for a given parameter $\lambda > 0$

$$\mathbb{E}[X] = \sum_{k=0}^{\infty} \frac{\lambda^k e^{-\lambda}}{k!} k$$



Source: wikipedia

Expectation Examples, cont'd

Rensselaer

- The expected value of a Poisson distribution is

$$
\begin{aligned}
\mathbb{E}[X] &= \sum_{k=0}^{\infty} \frac{\lambda^k e^{-\lambda}}{k!} k \\
&= \sum_{k=1}^{\infty} \frac{\lambda^k e^{-\lambda}}{(k-1)!} \\
&= \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} \\
&= \lambda e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} \\
&= \lambda e^{-\lambda} e^{\lambda} = \lambda
\end{aligned}
$$

First term is 0!

Expectation Examples, cont'd

Rensselaer

- Suppose you have a uniform distribution on $[0,1]$

$$\mathbb{E}[X] = 0.5$$

  - But why?
  - Uniform distribution has density $p(x) = 1$

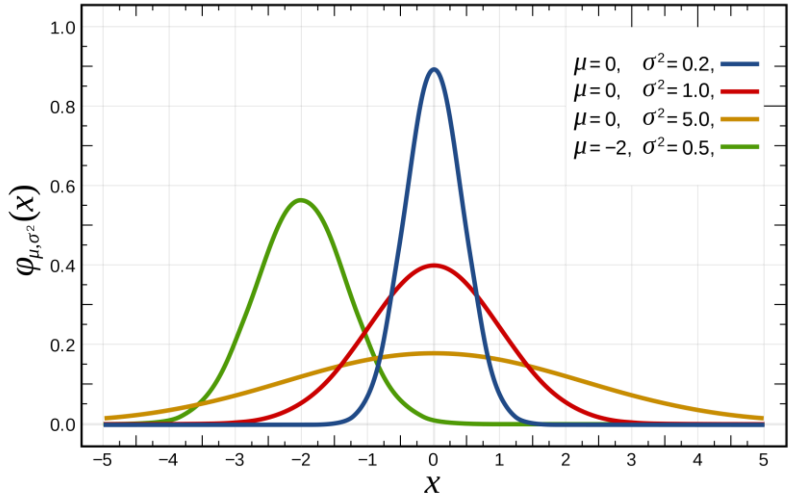
$$\mathbb{E}[X] = \int_0^1 x dx$$

$$= \frac{x^2}{2}\Big|_0^1 = 0.5$$

- For general intervals $[A, B]$, the density is $p(x) = \frac{1}{B-A}$

Expectation Examples, cont'd

- Suppose you have a normal distribution



- The pdf is

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- Looks intimidating but it's actually quite easy to work with
- One of the most popular distributions for many reasons
  - Central limit theorem, etc.

Variance

- The variance of a random variable X is defined as

$$\mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$$

- Measures how much X deviates from its mean
  - Very similar to the definition of squared error
- When $\mathbb{E}[X] = 0$, the variance is just $\mathbb{E}\left[X^2\right]$
  - This is called the second moment of X
  - Higher moments defined similarly: $\mathbb{E}\left[X^3\right]$, etc.
  - For complex distributions, higher moments provide even more information about the distribution spread

Variance Examples

- What's the variance of the fair coin?

$$\mathbb{E}\left[(X-0.5)^2\right] = 0.5 \ * (-0.5)^2 + 0.5 \ * (0.5)^2 = 0.25$$

- What's the variance of the fair die?

$$\mathbb{E}\left[(X-3.5)^2\right] = \frac{1}{6}\left((-2.5)^2 + (-1.5)^2 + (-0.5)^2 + (2.5)^2 + (1.5)^2 + (0.5)^2\right) \approx 2.92$$

- What's the variance of the uniform distribution?

$$\mathbb{E}\left[(X-0.5)^2\right] = \int_0^1 (x-0.5)^2 dx = \left[\frac{x^3}{3} - \frac{x^2}{2} + 0.25x\right]_0^1 = \frac{1}{12}$$

Entropy and Cross-Entropy

Rensselaer

- The entropy of a discrete random variable X is defined as

$$H(X) = -\sum_x p(x) \log[p(x)] = -\mathbb{E}[\log[p(X)]]$$

  – Measures the level of “surprise” or “information” in X

  – Similar to variance but with subtle differences

  – E.g., entropy is invariant to scale

- The cross-entropy between two distributions p and q is

$$H(p,q) = -\sum_x p(x) \log[q(x)]$$

  – Measures the similarity between the two distributions

KL Divergence

- The Kullback-Leibler divergence between two distributions is

$$D_{KL}(p||q) = \sum_x p(x) \log \left[\frac{p(x)}{q(x)}\right]$$

  - Another measure of difference between distributions

- Cross-entropy can defined in terms of entropy and KL divergence

$$H(p,q) = H(p) + D_{KL}(p||q)$$

- KL divergence can be thought of as a distance metric between distributions (although it's not symmetric)
- Cross-entropy is not a distance metric since $H(P,P) \neq 0$

Law of Large Numbers

- Let $X_1, \ldots, X_n$ be n IID random variables
- Let $S_n = X_1 + \cdots + X_n$
- (Weak) Law of Large Numbers:

$$\mathbb{P}\left[\left|\frac{S_n}{n} - \mathbb{E}[X_1]\right| < \epsilon\right] \to 1 \text{ as } n \to \infty$$

  - for any positive ϵ
- As we collect more data, the sample mean S_n/n converges to the expected mean $\mathbb{E}[X_1]$
  - Since the X_i are IID, $\mathbb{E}[X_1] = \mathbb{E}[X_i]$ for any i
- Practically speaking, as our dataset gets larger, the law of large numbers is more likely to apply
  - E.g., for accuracy, parameter estimates, etc.