CO245 - Probability and Statistics

15th January 2020

Probability is a mathematical formalism used to describe and quantify uncertainty.

Sample Spaces and Events

- **sample space** S or Ω

  a set containing the possible outcomes of a random experiment

  for example; sample space of two coin tosses $S = \{(H,H),(H,T),(T,H),(T,T)\}$
- **event** E $(E \subseteq S)$

  any subset of the sample space (collection of some possible events)

  for example; event of the first coin being heads in two tosses $E = \{(H,H),(H,T)\}$

  the extremes are $\varnothing$ (the null event) which will never occur, or S (the universal event) which will always occur - there is only uncertainty when the events are strictly between the events, such that $\varnothing \subset E \subset S$
- **elementary event** singleton subset containing exactly one element from S

When performing a random experiment, the outcome will be a single element $s^* \in S$. Then an event $E \subseteq S$ has **occurred** iff $s^* \in E$. If it has not occurred, then $s^* \notin E \Leftrightarrow s^* \in \bar{E}$ (can be read as not E).

With a set of events $\{E_1, E_2, \dots\}$, we can have the following set operations;

- $\bigcup_i E_i = \{s \in S \mid \exists i.\ [s \in E_i]\}$ will only occur if at least one of the events E_i occurs ("or")
- $\bigcap_i E_i = \{s \in S \mid \forall i.\ [s \in E_i]\}$ will only occur if all of the events E_i occurs ("and")
- $\forall i, j.\ E_i \cap E_j = \varnothing$ $(i \neq j)$ if they are mutually exclusive (at most one can occur)

σ-algebra

In an uncountably infinite set, the event set you are assigning probabilities to cannot be every subset, as the probabilities cannot be made to sum to 1 under reasonable axioms.

We define the σ-algebra as the subset of events which we can assign probabilities to. We want to define a probability function P that corresponds to the subsets of S that we wish to **measure**. This set of subsets is referred to as $\mathfrak{S}$ (the event space), with the following three properties (corresponding to the axioms of probability);

- nonempty $S \in \mathfrak{S}$
- closed under complements $E \in \mathfrak{S} \Rightarrow \bar{E} \in \mathfrak{S}$
- closed under countable union (therefore any countable set is fine) $E_1, E_2, \dots \in \mathfrak{S} \Rightarrow \bigcup_i E_i \in \mathfrak{S}$

A probability measure on the pair $(S, \mathfrak{S})$ is a mapping $P : \mathfrak{S} \to [0,1]$, satisfying the following three axioms;

- $\forall E \in \mathfrak{S}.\ [0 \leq P(E) \leq 1]$
- $P(S) = 1$
- countably additive, for **disjoint subsets** $E_1, E_2, \dots \in \mathfrak{S}$ $P\left(\bigcup_i E_i\right) = \sum_i P(E_i)$

From these, we can derive the following;

- $P(\bar{E}) = 1 - P(E)$

$$\underbrace{P(E) + P(E)}_{\text{disjoint}} = P(\underbrace{E \cup \bar{E}}_{E \cup \bar{E} = S}) = P(S) = 1$$

- $P(\varnothing) = 0$ special case of the above, when $E = S$
- for any events E and F $P(E \cup F) = P(E) + P(F) - P(E \cap F)$

16th January 2020

Independent Events

It's important to note that independent events are **not** the same as disjoint events. Two events E and F are independent iff $P(E \cap F) = P(E)P(F)$ - sometimes written as $E \perp F$. Generally, a set of events $\{E_1, E_2, \dots\}$ are set to be independent if for any finite subset $\{E_{i_1}, E_{i_2}, \dots, E_{i_n}\}$;

$$P\left(\bigcap_{j=1}^{n} E_{i_j}\right) = \prod_{j=1}^{n} P(E_{i_j})$$

Where we have $\{i_j \mid 1 \leq j \leq n\}$ is any set of distinct positive integers. Note that independence is more than just pairwise independence.

We propose that if events E and F are independent, then $\bar{E}$ and F are also independent. Note that E and $\bar{E}$ form a partition of S (they are disjoint, and union to S). $F = (E \cap F) \cup (\bar{E} \cap F)$ is a disjoint union (and also a partition of F), this gives us $P(F) = P(E \cap F) + P(\bar{E} \cap F) \Rightarrow P(\bar{E} \cap F) = P(F) - P(E \cap F)$;

$$\begin{aligned}
P(\bar{E} \cap F) &= P(F) - P(E \cap F) && E \text{ and } F \text{ are independent}, \Rightarrow \\
&= P(F) - P(E)P(F) && \Rightarrow \\
&= (1 - P(E))P(F) && \text{probability of complement}, \Rightarrow \\
&= P(\bar{E})P(F) && \text{hence independent}, \blacksquare
\end{aligned}$$

Interpretations of Probability

In order to assign meaning to P, we need to have some interpretation of probability, such as the following;

- **classical**

  If S is finite, and the elementary events are "equally likely", then for an event $E \subseteq S$, the probability is the number of outcomes in E out of the total number of possible outcomes (S);

  $$P(E) = \frac{|E|}{|S|}$$

  This idea of "equally likely" (uniform) can be extended to infinite spaces. Instead of taking the set cardinality, another standard measure (such as area or volume) can be used instead.

- **frequentist**

  The idea is that if someone were to perform the same experiment (E may or may not occur) in identical random situations many times, then the proportion of times E occurs will tend to some limiting value, which would be $P(E)$.

- **subjective**

  Not assessed. Probability is the degree of belief held by an individual (see *De Finetti*) - suppose a random event $E \subseteq S$ is to be performed, and an individual enters a game regarding this experiment, with two choices;

- gamble — if E occurs they win \$1, otherwise if $\bar{E}$ occurs they win \$0
- stick — regardless of the outcome, the individual receives $\$P(E)$

The critical value of $P(E)$, where the individual is indifferent between the choices, is their probability of E.

Dependent Probabilities and Conditional Probability

For the standard example of flipping a coin and rolling a die (assuming both fair), we have independence - the probability of each elementary event is $\frac{1}{2} \cdot \frac{1}{6} = \frac{1}{12}$.

However, consider the case where we have two die, where the first is fair, and the second is a "top", where we only have odd numbers (such that a roll of a 2 is mapped to a 5, 4 to 3, and 6 to 1). When we now flip the coin, if it is heads, we use the normal die, otherwise if it is tails, we use the "top". As expected, this is no longer independent.

For two events E and F in S, where $P(F) \neq 0$, we can define the probability of E occurring, given that we know F has occurred to be;

$$P(E|F) = \frac{P(E \cap F)}{P(F)}$$

Note that this also holds for independence ($P(E)$ doesn't change, as expected);

$$P(E|F) = \frac{P(E \cap F)}{P(F)} = \frac{P(E)P(F)}{P(F)} = P(E)$$

An example of this is as follows - suppose we roll two normal dice, with one from each hand. The sample space is all the ordered pairs of possible values $S = \{(1,1),(1,2),\ldots,(6,6)\}$. Let the event E be defined as the die from the left hand has a higher value than the die from the right hand. Looking at all possible combinations, we have;

$$P(E) = \frac{15}{36}$$

Suppose we now know F, the value of the left die being 5, has occurred. Since we know F has occurred, the only events that could have happened are $F = \{(5,1),(5,2),\ldots,(5,6)\}$. Similarly, the only sample space elements in E that could've occurred are $E \cap F = \{(5,1),(5,2),(5,3),(5,4)\}$. Our probability is as follows;

$$\frac{|E \cap F|}{|F|} = \frac{4}{6} = \frac{\frac{4}{36}}{\frac{1}{6}} = \frac{P(E \cap F)}{P(F)} \equiv P(E|F)$$

One way to think about probability conditioning as a shrinking of the sample space, with events being replaced by intersections with the reduced space, and a rescaling of the probabilities. For example, with $F = S$, we have the following;

$$P(E) = \frac{P(E)}{1} = \frac{P(E \cap S)}{P(S)} = P(E|S)$$

Furthermore, we can extend the idea of independence of events with respect to a probability measure P to conditional probabilities. $P(\cdot|F)$ is a valid probability measure which obeys the axioms of probability on the set F. For three events E_1, E_2, F, the event pair E_1 and E_2 are conditionally independent given F (sometimes written as $E_1 \perp E_2|F$) if and only if;

$$P(E_1 \cap E_2|F) = P(E_1|F)P(E_2|F)$$

Bayes Theorem

For two events E and F in S, we have $P(E \cap F) = P(F)P(E|F)$, and $P(E \cap F) = P(E)P(F|E)$ (interchanging, and noting commutativity of $\cap$). Hence we have Bayes Theorem;

$$P(E|F) = \frac{P(E)P(F|E)}{P(F)}$$

Partition Rule

Consider a set of events $\{F_1, F_2, \dots\}$, which form a partition of S (they are disjoint, and union together to form S). Then for any event $E \subseteq S$, the partition rule states;

$$P(E) = \sum_i P(E|F_i)P(F_i)$$

The proof is as follows;

$$\begin{aligned}
E &= E \cap S \\
&= E \cap \bigcup_i F_i && \text{by definition of partitions} \\
&= \bigcup_i (E \cap F_i) && \text{by distributivity of intersection} \\
P(E) &= P\left(\bigcup_i (E \cap F_i)\right) \\
&= \sum_i P(E \cap F_i) && \text{disjoint union} \\
&= \sum_i P(E|F_i)P(F_i)
\end{aligned}$$

Note that $\{E \cap F_1, E \cap F_2, \dots\}$ is disjoint if $\{F_1, F_2, \dots\}$ is. Assume there is an element $s \in E \cap F_i$ and $s \in E \cap F_j$ (where $i \neq j$), if it is in both, then $s \in F_i$ and $s \in F_j$, which is not possible.

Note that $\{F, \bar{F}\}$ forms a partition of S, therefore by the Law of Total Probability we have;

$$P(E) = P(E \cap F) + P(E \cap \bar{F}) = P(E|F)P(F) + P(E|\bar{F})P(\bar{F})$$

Terminology

- conditional probabilities $P(E|F)$
- joint probabilities $P(E \cap F)$
- marginal probabilities (margins of a table) $P(E)$

  margins of a table

Likelihood and Posterior Probability

Suppose we have a probability model with parameters θ, that define a model instance (such as μ and σ), and a set of observations (or evidence) X.

- **likelihood function** (probability of the evidence, given the parameters) $P(X|\theta)$

  what is the probability our model will predict that evidence?
- **posterior probability** (probability of the parameters, given the evidence) $P(\theta|X)$

  what is the probability the actual parameters are θ, given our evidence?
- **prior probability** (not taking into account the evidence) $P(\theta)$

This is related by Bayes theorem;

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)}$$

$$\text{posterior probability} \propto \text{likelihood} \times \text{prior probability}$$

This is then divided by the normalising constant;

$$\sum_\theta P(X|\theta)P(\theta) = P(X)$$

22nd January 2020

Example Questions

1. There are 5000 VLSI chips, 1000 from company X (which has a 10% chance of being defective), and 4000 from company Y (which has a 5% chance of being defective). If a chip is defective, what is the probability it came from company X?

   Let E be the event that the randomly selected chip was made by X, and F be the event that the chip is defective.

   $$\begin{aligned} P(E) &= \frac{1000}{5000} \\ &= 0.2 \\ P(\bar{E}) &= \frac{4000}{5000} \\ &= 0.8 \\ P(F|E) &= 0.1 && \text{given} \\ P(F|\bar{E}) &= 0.05 && \text{given} \\ P(E \cap F) &= P(F|E)P(E) \\ &= 0.02 \\ P(\bar{E} \cap F) &= P(F|\bar{E})P(\bar{E}) \\ &= 0.04 \end{aligned}$$

   This gives us enough to fill in the table, as well as the missing entries with the law of total probabilities;

	E	$\bar{E}$	
F	0.02	0.04	0.06
$\bar{F}$	0.18	0.76	0.94
	0.2	0.8	

   $$P(E|F) = \frac{P(E \cap F)}{P(F)} = \frac{0.02}{0.06} = \frac{1}{3}$$

2. A multiple choice question has c available choices. Let p be the probability the student knows the right answer. When he doesn't know, he chooses an answer at random. Given that the answer the student chooses is correct, what is the probability that the student knew the correct answer?

   Let A be the event that the question is answered correctly, and K be the event that the student knew the correct answer. We therefore want to find $P(K|A)$.

   $$P(K|A) = \frac{P(A|K)P(K)}{P(A)}$$

   We know $P(A|K) = 1$ (given that they don't purposely choose a wrong answer), and $P(K) = p$. By the partition rule, we have $P(A) = P(A|K)P(K) + P(A|\bar{K})P(\bar{K})$. Substituting values we get;

   $$P(A) = 1 \cdot p + \frac{1}{c} \cdot (1-p) = p + \frac{1-p}{c}$$

   Therefore,

   $$P(K|A) = \frac{p}{p + \frac{1-p}{c}} = \frac{cp}{cp + 1 - p}$$

3. A new HIV test is claimed to correctly identify 95% of people who are really HIV positive and 98% of people who are really HIV negative.

(a) If only 1 in a 1000 of the population are HIV positive, what is the probability that someone who tests positive actually has HIV?

Let H be the event that someone has the virus ($P(H) = 0.001$), and T be the event that someone tests positive. Similar to above, we want to find the following, and can use the partition rule again.

$$P(H|T) = \frac{P(T|H)P(H)}{P(T)} = \frac{P(T|H)P(H)}{P(T|H)P(H) + P(T|\bar{H})P(\bar{H})} \approx 0.045$$

Therefore, less than 5% of those who test positive really have HIV.

(b) Is this acceptable? no

(c) Would a repeat test be appropriate for someone who tests positive?

Let T_i denote the event that the i^{th} test is positive. Suppose that the correctness of the test stays the same, and the test results are conditionally independent.

$$\begin{aligned}
P(H|T_1 \cap T_2) &= \frac{P(T_1 \cap T_2|H)P(H)}{P(T_1 \cap T_2)} \\
&= \frac{P(T_1 \cap T_2|H)P(H)}{P(T_1 \cap T_2|H)P(H) + P(T_1 \cap T_2|\bar{H})P(\bar{H})} \\
&= \frac{P(T_1|H)P(T_2|H)P(H)}{P(T_1|H)P(T_2|H)P(H) + P(T_1|\bar{H})P(T_2|\bar{H})P(\bar{H})} \\
&\approx 0.693
\end{aligned}$$

23rd January 2020

Simple Random Variables

Suppose we have identified a sample space S and a probability measure $P(E)$ on (measurable subsets) $E \subseteq S$. A random variable is a mapping from the sample space to the real numbers, such that a random variable $X : S \to \mathbb{R}$. Each element $s \in S$ is assigned a numerical value $X(s)$ (not always unique). We denote the outcome of the random experiment as s^*, the corresponding unknown outcome of the random variable $X(s^*)$ will be referred to as X.

The probability measure P defined on S induces a **probability distribution function**, P_X, on the random variable $X \in \mathbb{R}$. For each $x \in \mathbb{R}$, let $S_x \subseteq S$ be the set containing the elements of S which are mapped by X to numbers no greater than x, precisely $S_x = X^{-1}((-\infty, x])$.

$$P_X(X \leq x) = P(S_x)$$

We define the image of S under X as the range of the random variable X;

$$\text{range}(X) \equiv X(S) = \{x \in \mathbb{R} \mid \exists s \in S.\ [X(s) = x]\}$$

Consider this applied to the experiment of a fair coin toss, with $S = \{H, T\}$, probability measure $P(\{H\}) = P(\{T\}) = \frac{1}{2}$, and a random variable $X : \{H, T\} \to \mathbb{R}$ (such that $X(T) = 0$ and $X(H) = 1$);

$$X^{-1}((-\infty, x]) = \begin{cases} \varnothing & x < 0 \\ \{T\} & 0 < x < 1 \\ \{H, T\} & x \geq 1 \end{cases}$$

$$\begin{aligned}
P_X(X \leq x) &= \begin{cases} P(\varnothing) & x < 0 \\ P(\{T\}) & 0 < x < 1 \\ P(\{H, T\}) & x \geq 1 \end{cases} \\
&= \begin{cases} 0 & x < 0 \\ \frac{1}{2} & 0 < x < 1 \\ 1 & x \geq 1 \end{cases}
\end{aligned}$$

The **cumulative distribution function** of a random variable X, $F_X(x)$ is the probability that X takes a value less than or equal to x;

$$F_X(x) = P_X(X \leq x)$$

To verify a function $F_X(x)$ is a valid cdf, we need to verify the following properties;

- $0 \leq F_X(x) \leq 1$, $\forall x \in \mathbb{R}$
- $\forall x_1, x_2 \in \mathbb{R}.\ [x_1 \leq x_2 \Rightarrow F_X(x_1) \leq F_X(x_2)]$ monotonicity
- $F_X(-\infty) = 0$, and $F_X(\infty) = 1$

Note that for finite intervals $(a, b] \subseteq \mathbb{R}$; $P_X(a < X \leq b) = F_X(b) - F_X(a)$. Unless there is ambiguity, we can generally omit the subscript of P_X, to just write P - thus we just consider the random variable from the start, letting the range of X be the sample space.

We define a random variable as simple if it can only take a finite number of possible values. Suppose X is simple, and can take m values $\mathcal{X} = \{x_1, x_2, \ldots, x_m\}$, ordered $x_1 < x_2 < \cdots < x_m$. Each $s \in S$ is mapped to one of these values by X. The sample space S can then be partitioned into m disjoint subsets, $\{E_1, E_2, \ldots, E_m\}$, such that $s \in E_i \Leftrightarrow X(s) = x_i$. Therefore we have $P_X(X = x_i) = P(E_i)$, and $P_X(X = x_i) = F_X(x_i) - F_X(x_{i-1})$, with $x_0 = -\infty$.

A random variable is simply a numeric relabelling of our underlying sample space.

Discrete Random Variables

A random variable is discrete if it can take only a **countable** number of possible values (the range is countable). Therefore a simple random variable is a special case of a discrete random variable. Similar to above, we can partition S into a countable collection of disjoint subsets. For a discrete random variable X, F_X is a monotonic increasing step function with jumps at points in $\mathcal{X} = \{x_1, x_2, \ldots\}$, where $x_1 < x_2 < \ldots$, continuous on the right.

For a discrete random variable X and $x \in \mathbb{R}$, we define the **probability mass function**, $p_X(x)$ or just $p(x)$ as;

$$p_X(x) = P_X(X = x)$$

Given that X can take the values $\mathcal{X} = \{x_1, x_2, \ldots\}$, then the following must hold;

- $0 \leq p_X(x) \leq 1$, $\forall x \in \mathbb{R}$
- $\sum_{x \in \mathcal{X}} p_X(x) = 1$

Either the probability mass function (pmf) or the cumulative distribution function (cdf) of a random variable fully characterises its distribution, as we can work one out from the other;

- $p(x_i) = F(x_i) - F(x_{i-1})$
- $F(X_i) = \sum_{j=1}^{i} p(x_j)$

Link to Statistics

Consider the set of data $(x_1, x_2, \ldots, x_n)$ as n realisations of a random variable X. The frequency counts in the histogram for that set of data can be seen as an estimate for the probability mass function. Similarly, a cumulative histogram is an estimate of the cumulative distribution function.

Expectation

We define the **expectation** (also written as $E(X)$ or μ_X) of a discrete random variable X as

$$E_X(X) = \sum_x x p_X(x)$$

This gives a weighted average of the possible values, with the weights coming from the probability of a particular outcome. Occasionally referred to as the mean of the distribution.

The expectation of a function of a random variable is denoted $E\{g(X)\}$, where $g : \mathbb{R} \to \mathbb{R}$. We notice that $g(X)(s) = (g \circ X)(s)$ is also a random variable, therefore the expectation is;

$$E_X\{g(X)\} = \sum_x g(x) p_X(x)$$

Note that for a linear function $g(X) = aX + b$, where $a, b \in \mathbb{R}$, we have $E_X(aX + b) = aE_X(X) + b$. Similarly, for two linear functions g, h, $E_X(g(x)+h(x)) = E_X(g(x))+E_X(h(x))$. Therefore expectation is a linear operator.

The variance is the expectation of X, with $g(X) = (X - E(X))^2$. This is denoted $\mathrm{Var}_X(X)$, or sometimes σ_X^2.

$$\mathrm{Var}_X(X) = E_X((X - E_X(X))^2) = E(X^2) - E(X)^2$$

The variance of a linear function of a random variable is as follows;

$$\mathrm{Var}(aX + b) = a^2 \mathrm{Var}(X)$$

The standard deviation of a random variable, $\mathrm{sd}_X(X)$ (also σ_X) is the square root of the variance.

$$\mathrm{sd}_X(X) = \sqrt{\mathrm{Var}_X(X)}$$

The skewness γ_1 of a discrete random variable X is defined;

$$\gamma_1 = \frac{E_X((X - E_X(X))^3)}{\mathrm{sd}_X(X)^3} = \frac{E_X((X - \mu)^3)}{\sigma^3}$$

The part in violet is when $\mu = E(X), \sigma = \mathrm{sd}(X)$.

Example Questions

1. If X is a random variable taking the integer value scored with a single roll of a fair die, what is

   (a) the expected value

$$\begin{aligned} E(X) &= \sum_{x=1}^{6} x p(x) \\ &= 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} \\ &= \frac{21}{6} \qquad (= 3.5) \end{aligned}$$

   (b) the variance

$$\begin{aligned} \mathrm{Var}(X) &= \sum_{x=1}^{6} x^2 p(x) - 3.5^2 \\ &= 1^2 \cdot \frac{1}{6} + 2^2 \cdot \frac{1}{6} + 3^2 \cdot \frac{1}{6} + 4^2 \cdot \frac{1}{6} + 5^2 \cdot \frac{1}{6} + 6^2 \cdot \frac{1}{6} - \left(\frac{7}{2}\right)^2 \\ &= \frac{35}{12} \end{aligned}$$

2. A student gets X marks answering a single multiple choice question with four options, where 3 marks are awarded for a correct answer, and -1 for a wrong answer - what is

   (a) the expected value

$$\begin{aligned} E(X) &= 3 \cdot P(\text{correct}) + -1 \cdot P(\text{incorrect}) \\ &= 3 \cdot \frac{1}{4} + -1 \cdot \frac{3}{4} \\ &= 0 \end{aligned}$$

   (b) the standard deviation

$$\begin{aligned} E(X^2) &= 3^2 \cdot P(\text{correct}) + (-1)^2 \cdot P(\text{incorrect}) \\ &= 9 \cdot \frac{1}{4} + 1 \cdot \frac{3}{4} \\ &= 4 \qquad \Rightarrow \\ \text{sd}(X) &= \sqrt{3} \end{aligned}$$