Lecture 2: Probability and Entropy

ENGR 76 lecture notes — April 4, 2024

Ayfer Ozgur, Stanford University

In this lecture, we revisit the basic notions of probability and define some quantities like entropy that will form the foundation for the upcoming lectures.

1 Probability

For events $A, B, \ldots$, we will write $P(A)$ to denote the probability of event A happening, and $P(A, B) = P(A \cap B)$ to denote the probability of both A and B happening. The basic intuition for probability is given by the *law of large numbers*, which shows that if you repeat an experiment independently multiple times, the fraction of times A happens is close to its probability $P(A)$.

Independence of events: As the name suggests, two events are independent when knowing that one of them happened doesn't change the probability of the other happening. More formally, we say events A and B are independent if $P(A, B) = P(A)P(B)$. Intuitively, if A has no influence on the occurance of B, then B will happen in a fraction $P(B)$ of those times when A has already happenned, which in turn happens a fraction $P(A)$ of the time. Hence, the fraction of times A and B happen together, i.e. $P(A, B)$, is the product of these two fractions.

Random variables

At a basic level, X is a *random variable* (r.v.) if the value it takes is random. You might have seen more formal definitions in other classes, but this level of understanding suffices for us.

X is a *discrete random variable* if the set of possible value that it can take, denoted by $\mathcal{X}$ is discrete. We will focus on discrete r.v.s for now.

Some examples of discrete sets are $\mathcal{X} = \{0, 1\}$ (in which case X is a binary r.v.), $\mathcal{X} = \{1, 2, \ldots, m\}$ (all integers from 1 to m), and $\mathcal{X} = \{1, 2, \ldots\}$ (set of all positive integers, infinite but discrete). An example of a continuous (not discrete) set is the real interval from 0 to 1, i.e., the set $[0, 1]$.

It is easy to characterize discrete r.v.s by specifying $P(X = x)$ for $x \in \mathcal{X}$, i.e., by specifying the probability of each value in the alphabet $\mathcal{X}$.

Independence of random variables: Similar to independence for events, two r.v.s are independent if knowing the value of one doesn't tell us anything about the other. More formally, two r.v.s X and Y (taking values in $\mathcal{X}$ and $\mathcal{Y}$, respectively) are independent if the events $\{X = x\}$ and $\{Y = y\}$ are independent for every $x \in \mathcal{X}$ and every $y \in \mathcal{Y}$. Equivalently, X and Y are independent if $P(X = x, Y = y) = P(X = x)P(Y = y)$ for every $x \in \mathcal{X}$ and every $y \in \mathcal{Y}$.

Expectation of random variables: The expectation of a random variable X is defined as

$$\mathbb{E}[X] \triangleq \sum_{x \in \mathcal{X}} xP(X = x)$$

If we draw multiple random variables according to the distribution of X, then we expect the value x to occur a fraction $P(X = x)$ times based on the law of large numbers. Thus, if take the average of the random values, we expect it to be close to the expectation of X, which is simply a weighted average of the values X takes, with the weights equal to the respective probabilities.

Expectation of functions of random variables: For a random variable X and a function f, we can consider $Y = f(X)$ which is simply another random variable obtained by the applying the function f to X. The expectation of Y can in fact be directly computed based on the distribution of X as follows:

$$\mathbb{E}[Y] = \sum_{y \in \mathcal{Y}} yP(Y = y) = \sum_{x \in \mathcal{X}} f(x)P(X = x)$$

2 Surprise and Entropy

Let's think about how *informative* an event is. If A is a very likely event, and I learn that A happened, that doesn't tell me much since this was already expected. But if A is a very unlikely event, and I learn that A happened, that is very informative. In a sense, we can say that less likely events are more informative because they are more surprising.

With this intuition, let's attempt to find a *surprise* function S which captures the amount of surprise in an event A. First we look at some properties such a function should satisfy.

- S is a function of the probability of the event, i.e., we expect two events with the same probability to have the same surprise. Hence we can write the function as $S(p)$ for $0 < p \leq 1$ or as $S(P(A))$.
- $S(p)$ should decrease with increasing p since more likely events are less surprising.
- $S(p)$ is a continuous function of p, since we don't expect the surprise value to suddenly jump for an event of probability, say, 0.2 as compared to an event of probability 0.2001.
- Consider two independent events A and B. Then we can consider $S(P(A, B))$ which is the surprise value for both A and B happening. We might expect this to be the sum of the individual surprise values, i.e., $S(P(A, B)) = S(P(A)P(B)) = S(P(A)) + S(P(B))$. While it's not immediately obvious why we have the sum and not some other operation, we will see later that this property ends up being very important.

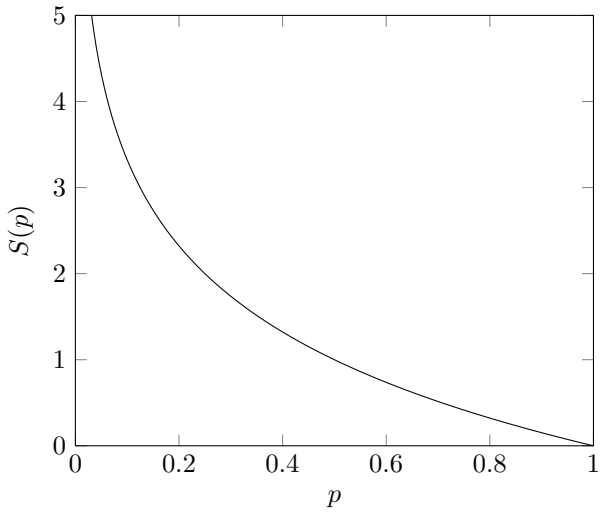
More concisely, we are looking for a function $S(p)$ for $0 < p \leq 1$, which satisfies:

1. $S(p)$ is a decreasing function of p.
2. $S(p)$ is a continuous function of p.
3. For $p_1, p_2 \in (0, 1]$, $S(p_1 p_2) = S(p_1) + S(p_2)$.

Surprisingly, these properties are enough to define a *unique* function S.

Theorem. *The only function S satisfying properties 1, 2 and 3 above is $S(p) = \log 1/p$.*

Note that the function S is defined up to a constant, or equivalently, the base of the logarithm can be chosen arbitrarily. The figure below shows a plot of the function with base 2. We see that the value of $S(p)$ grows arbitrarily large as p goes closer to zero. An elementary proof of the theorem is provided in Section 4.



Entropy

We will use the notion of the surprise function to define the Shannon entropy as the average surprise of the random variable X. The surprise associated with the the event $X = x$ is given by

$$S(P(X = x)) = \log \frac{1}{P(X = x)}.$$

Then, the **Shannon Entropy**, or simply the **entropy**, of X is

$$H(X) \triangleq \sum_{x \in \mathcal{X}} P(X = x) S(P(X = x)) = \sum_{x \in \mathcal{X}} P(X = x) \log \frac{1}{P(X = x)}$$

The Shannon Entropy of X can be interpreted as the average amount of surprise on learning the value of X. Note that the entropy only depends on the probability values and not on the actual alphabet $\mathcal{X}$, i.e., a change in the labels does not impact the entropy.

3 Entropy, joint entropy and some properties

For a discrete random variable X with alphabet $\mathcal{X}$, recall that its entropy $H(X)$ is defined as

$$H(X) \triangleq \sum_{x \in \mathcal{X}} P(X = x) \log \frac{1}{P(X = x)}$$

Unless otherwise specified, we will use log base 2, and hence the entropy will be measured in bits ("binary digits").

(0) $H(X) \geq 0$. This holds because the probabilities $P(X = x) \leq 1$, and hence $\log \frac{1}{P(X=x)} \geq 0$. Equality holds iff ("if and only if") X is deterministic, i.e., $P(X = x) = 1$ for some x and 0 otherwise. Intuitively, if X is deterministic, there is no surprise associated with it, and hence the entropy is zero.

(1) **Example:** Let $\mathcal{X} = \{1, \ldots, M\}$ and $P(X = i) = \frac{1}{M}$ for all $1 \leq i \leq M$. In words, X is uniformly distributed over $\{1, \ldots, M\}$ and each value in the set is equally likely.

$$\begin{aligned} H(X) &= \sum_{i=1}^{M} P(X = i) \log \frac{1}{P(X = i)} \\ &= \sum_{i=1}^{M} \frac{1}{M} \log M \\ &= \log M \end{aligned}$$

This already suggests that there is a link between the entropy and representation of information in bits. Note that k bits can represent 2^k distinct values, and in this example $k = \log M$ bits are capable of representing a random variable uniformly distributed over $2^k = M$ values. We also note that as M grows, the entropy also grows which is consistent with our intuitive understanding of entropy measuring the amount of surprise in the random variable.

Fact: For any r.v. X with alphabet $\mathcal{X} = \{1, \ldots, M\}$, $H(X) \leq \log M$. This means that among all r.v.s over alphabet $\mathcal{X}$, the uniformly distributed random variable has the highest entropy.

Joint entropy: For any pair of r.v.s X and Y (with alphabets $\mathcal{X}$ and $\mathcal{Y}$, respectively), we define their joint entropy as

$$H(X, Y) \triangleq \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P(X = x, Y = y) \log \frac{1}{P(X = x, Y = y)}$$

(2) **Example:** Consider
$\mathcal{X} = \{1, \ldots, M\}$, $P(X = i) = \frac{1}{M}$, $1 \leq i \leq M$
$\mathcal{Y} = \{1, \ldots, N\}$, $P(Y = j) = \frac{1}{N}$, $1 \leq j \leq N$
X and Y are independent, i.e., $P(X = i, Y = j) = P(X = i)P(Y = j) = \frac{1}{MN}$ for $1 \leq i \leq M, 1 \leq j \leq N$

So the pair (X, Y) takes values in an $M \times N$ grid of points with each point having the same probability of $\frac{1}{MN}$. Now, using the result from example 1 above, we get

$$\begin{aligned} H(X,Y) &= \log MN \\ &= \log M + \log N \\ &= H(X) + H(Y) \end{aligned}$$

where the last equality again makes use of example 1 since X and Y are uniformly distributed over their respective alphabets.

Fact: For any pair of independent r.v.s X and Y, $H(X,Y) = H(X) + H(Y)$. More generally, $H(X,Y) \leq H(X) + H(Y)$ with equality iff X and Y are independent.

We will learn more about this when we encounter conditional entropy in a later lecture.

Next week, we will see how the entropy is closely tied with representation of information and provides fundamental limits for the same.

4 Appendix: Proof that $S(p) = \log 1/p$

Consider a function $S(p)$ for $0 < p \leq 1$, which satisfies:

1. $S(p)$ is a decreasing function of p.
2. $S(p)$ is a continuous function of p.
3. For $p_1, p_2 \in (0, 1]$, $S(p_1 p_2) = S(p_1) + S(p_2)$.

Theorem. *The only function S satisfying properties 1, 2 and 3 above is $S(p) = \log 1/p$.*

Proof: Our approach is to first prove this for $p = 1/n$, then extend it to rational values of p and finally show it for all $p \in (0, 1]$. To that end, define $f(n) = S(1/n)$. By property 1, we have $m < n$ implies $f(m) < f(n)$.

By property 3, we have

$$f(nm) = f(n) + f(m)$$

By applying this repeatedly (using induction), we can show that

$$f(n^k) = kf(n) \tag{1}$$

Also, note that $f(1) = 0$ since $1 \times 1 = 1$. So we fix $n > 1$. For any integer $r > 0$, we can find k such that

$$n^k \leq 2^r < n^{k+1} \tag{2}$$

which implies

$$f(n^k) \leq f(2^r) < f(n^{k+1})$$

Now applying equation (1) above to each of the terms, we get

$$kf(n) \leq rf(2) < (k+1)f(n)$$

We can simplify this to

$$\frac{k}{r} \leq \frac{f(2)}{f(n)} < \frac{k+1}{r} \tag{3}$$

Also, from (2),

$$k \log n \leq r \log 2 < (k+1) \log n$$

or

$$\frac{k}{r} \leq \frac{\log 2}{\log n} < \frac{k+1}{r} \tag{4}$$

Noting the similarities in equations (3) and (4), we combine them to get

$$\left| \frac{\log 2}{\log n} - \frac{f(2)}{f(n)} \right| < \frac{1}{r}$$

Since r can be arbitrarily large, we obtain

$$\frac{\log 2}{\log n} = \frac{f(2)}{f(n)}$$

Thus, for some constant $c > 0$, we obtain

$$f(n) = c \log n$$

We have now shown the result for p of the form $1/n$. Next we consider $p = m/n$ where m and n are positive integers. Then we can use property 3 to get

$$S\left(\frac{1}{n}\right) = S\left(\frac{m}{n} \times \frac{1}{m}\right) = S\left(\frac{m}{n}\right) + S\left(\frac{1}{m}\right)$$

This directly gives us

$$S\left(\frac{m}{n}\right) = S\left(\frac{1}{n}\right) - S\left(\frac{1}{m}\right) = c \log \frac{m}{n}$$

In other words, $S(p) = c \log 1/p$ for all rational $0 < p \leq 1$. But this means $S(p) = c \log 1/p$ for all $0 < p \leq 1$, using the continuity of S (property 2) and the denseness of rationals over the real line. This completes the proof. Note that the constant c can be absorbed in the logarithm by changing the base. □