AICC II
Prof. Bixio Rimoldi

Benjamin Bovey

Semester of Spring 2019

19th February 2019

As opposed to the first AICC course, where we were mostly presented with tools, we will now see more applications of these tools for communication and computation. Mainly, we will see 3 applications in the first part of the semester:

- **Source coding** (compressing information)
- **Cryptography** (authentication / privacy / integrity of information)
- **Channel Coding** (dealing with noise and loss of information / protecting the information from natural damages)

What these three have in common is the idea of storing and communicating information. The notion of entropy, which will come up quite often, will also be important.

Basic probability review

Special case first: finite sample space Ω and uniform distribution. $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$. Events: $E \subseteq \Omega$. Then:

$$P(E) = \frac{|E|}{|\Omega|} \qquad \text{(uniform distribution)} \tag{1}$$

Conditional probability

Let E, F be two events. Then, the probability that event E occurs knowing that F has occured:

$$P(E|F) = \frac{|E \cap F|}{|F|} \tag{2}$$

Intuitively, you restrict the sample space to F only, because you *know* that F has happened: this translates to the division by the cardinality of F instead of the cardinality of Ω, like we did previously. The intersection follows from the

fact that the sample space is restricted to F: if there exists elements that are in E but not in F, they are now outside of the sample space, which means that these elements CANNOT "occur" in conjunction with F. Therefore, we take the intersection of E and F to assure that these elements are not taken into account in the computation.

Law of total probability

Let E and F be two events in Ω, and let F^C denote the complement of F. Then:

$$P(E) = P(E|F)P(F) + P(E|F^C)P(F^C) \tag{3}$$

This follows quite directly from the fact that $E = (E \cap F) \cup (E \cap F^C)$, so $P(E) = P(E \cap F) + P(E \cap F^C)$...

Application: divide and conquer You can sometimes create a partition of your sample space, in a way that allows you to better apply the numbers you are given (p.27-28).

Bayes' Rule

Bayes' rule allows you to compute $p(F|E)$, given that you know $p(E|F), p(E)$ and $p(F)$:

$$p(F|E) = \frac{p(E|F)p(F)}{p(E)} \tag{4}$$

This is useful, in real scenarios, when either one of $p(E|F)$ and $p(F|E)$ is easily observable, but the other isn't.

Random variables

A *random variable* X is a function $X : \Omega \to \mathbb{R}$. It is attached a *probability distribution function* $p_X(x)$, which represents the probability that X will take on the value x, that is, that the following event occurs:

$$E = \{\omega \in \Omega : X(\omega) = x\} \tag{5}$$

Hence,

$$p_X(x) = p(E) = \sum_{\omega \in E} p(\omega) \tag{6}$$

The set of all possible values of X is sometimes called the *alphabet* of X, written with more curly letters like $\mathcal{A}$.

Two random variables

Let $X : \Omega \to \mathbb{R}$ and $Y : \Omega \to \mathbb{R}$ be two random variables.
The probability of the event $E = \{X = x\} \cap \{Y = y\} (= \{X = x \wedge Y = y\})$ is

$$p_{X,Y}(x, y) = \sum_{\omega \in E} p(\omega) \tag{7}$$

We can compute p_X (or p_Y, similarly) from $p_{X,Y}$:

$$p_X(x) = \sum_y p_{X,Y}(x, y) \tag{8}$$

In one sense, we "fix in place" the value of x and "scroll through" all possible values of y, and add their probabilities up. Here, p_X is called the **marginal distribution** of $p_{X,Y}(x, y)$ with respect to x.

21st February 2019

Expected value

The **expected value**, or **mean** of a random variable $X : \Omega \to \mathbb{R}$, can be computed as

$$E[X] = \sum_{x \in \mathcal{A}(X)} x p_X(x) \quad \text{(requires } p_X\text{)}, \tag{9}$$

or as

$$E[X] = \sum_{\omega \in \Omega} X(\Omega) p(\omega). \tag{10}$$

One could say that the first way is "calculating over the codomain", and the second way is "calculating over the domain" (of X).

The expected value is a linear operation. Let $X_1, X_2, \ldots, X_n$ be random variables from Ω to $\mathbb{R}$, and let $\lambda_1, \lambda_2, \ldots, \lambda_n$ be real numbers. Then

$$E \underbrace{\Big[\sum_{i=1}^{n} X_i \lambda_i \Big]}_{\text{random variable}} = \sum_{i=1}^{n} \lambda_i E[X_i] \tag{11}$$

Extending notions from events to random variables

The notion of independent events extends to random variables. Recall that two events E and F are independent iff. $p(E|F) = p(E)$, which is equivalent to saying that $p(E \cap F) = p(E)p(F)$.

Similarly, two random variables $X, Y : \Omega \to \mathbb{R}$ are **independent** iff.

$$p_{X,Y}(x, y) = p_X(x) p_Y(y) \tag{12}$$

The notion of conditional probability also extends to random variables:

$$p(X = x|Y = y) = \frac{p(X = x \wedge Y = y)}{p(Y = y)} \tag{13}$$

The following statements are equivalent:

$$p_{X,Y}(x, y) = p_X(x) \tag{14}$$
$$p_{X|Y}(x|y) = p_X(x) \tag{15}$$
$$p_{Y|X}(y|x) = p_Y(y) \tag{16}$$

USEFUL TRICK if you are asked to check the independence of X and Y, you don't have to check the equality of (15) or (16). You just have to find the expression for the left-hand side function, and see if it depends on the other variable ($\implies$ they are dependent), or if it is just a function of one variable ($\implies$ they are independent).

CONSEQUENCE OF THE INDEPENDENCE OF RANDOM VARIABLES in general, $E[XY] \neq E[X]E[Y]$. However, when X and Y are *independent*, we have that

$$E[XY] = E[X]E[Y] \tag{17}$$

Source & Entropy

How do we define a source? This is a question that took some time to find an answer. We can think of a source as a black box, whose center of interest isn't really its mechanism, but rather what comes out of it, that is, some information. Since we are considering sources from a computer science point of view, we will be interested in sources that shite out sequences of numbers.
Entropy comes into the frame when we realize that a symbol that can be *predicted* provides no information. For example, if the sequence of numbers coming out of the source is (1, 1, 5, 5, 3, 3, 19, 19, 1, ...), we quickly realize that we do not need to store the second number of each pair, as it brings no new information to the table.
An important observation that we can make (that was initially made by Hartley in 1929), is the fact that this link between information and entropy is very much (or can very much be) linked to random variables, since a random variable gives you a number that you can not *predict* until you actually *do* the experiment, and see what comes out of it (hence, in fact, the name of *random* variable). In this case, doing the experiment brings you a new, unpredicted and unpredictable piece of information. A source can then be viewed as outputting a sequence of random variables.
Let us now consider a source outputting a sequence of random variables $S_1, S_2, S_3, \ldots$.
Another fundamental question we may ask ourselves is: how much information is actually stored in a symbol? A partial answer was given by Hartley, that is, that this must depend on the alphabet of the random variable. The bigger the alphabet, the more information it must carry (more possiblities $\iff$ more

entropy?). An alphabet $\mathcal{A}$ of size $|\mathcal{A}| = n$ symbols should carry $n\times$ information of one symbol, which means that the information grows linearly in the size of the alphabet.
With basic combinatorics, we can see that there are $|\mathcal{A}|^n$ possible sequences $(s_1, s_2, \ldots, s_n)$. Therefore, the amount of information carried by S_i is $\log|\mathcal{A}|$.
EXAMPLE 1 Let's say we have a jukebox with 64 songs in a restaurant. Every time a client makes a choice, the music that the client has requested is played. If you wanted to do statistics on what song was played how many times... TO BE COMPLETED WHEN I UNDERSTAND THIS EXAMPLE, the end result was that each new client that came in and played a song gives you $\log_2(64) = 6$ new bits of information.
EXAMPLE 2 London good days vs bad days: $(s_1, s_2, \ldots, s_{365}) = (0, 1, 1, 0, 1, \ldots, 0, 1, 1)$. There is a lot of entropy, it is unpredictable. It would therefore be hard to optimize the storage of this information better.
San Diego: $(\overbrace{0, 0, 0, \ldots, 0}^{\text{24 zeros}}, 1, \overbrace{0, 0, 0, \ldots, 0, 0}^{\text{340 zeros}})$. This is a very predictable sequence, which would allow us to just store $(24, 1, 340)$ rather than storing all 365 bits.

Entropy redefined by Shannon

Shannon, in 1948, gave a new formula for the amount of information carried by the random variable $S \in \mathcal{A}$. He said that this amount *is* in fact the entropy itself, and gave this formula for the entropy $H(S)$:

$$H(S) = -\sum_{s\in\text{supp}(P_S)} p_S(s) \cdot \log_b\big(p_S(s)\big) \tag{18}$$

COMMENTS

- $s \in \text{supp}(S)$ is needed because $\log(0)$ is not defined
- if we are using the convention $0 \cdot \log(0) = 0$, then we can simplify the notation by writing $H(S) = -\sum_{s\in\mathcal{A}} p_S(s) \log_b\big(p_S(s)\big)$
- when $b = 2$ (default) then the unit is the bit. $H(S) = H_2(S)$
- if we rewrite the formula as $H(S) = \sum\limits_{s\in\mathcal{A}} p_S(s) \underbrace{\Big(-\log_b\big(p_S(s)\big)\Big)}_{\text{rand. var. } X} = E[X]$, because this is the expression of the expected value of a random variable X.

EXAMPLE Let the random variable $S \in \mathcal{A}$ be uniformly distributed. Then

$$H(S) = -\sum_{s\in\text{supp}(P_S)} \underbrace{p_S(s)}_{\frac{1}{|\mathcal{A}|}} \log_b\big(\underbrace{p_S(s)}_{\frac{1}{|\mathcal{A}|}}\big)$$

$$=$$

So, Hartley and Shannon agree when the random variable has a uniform distribution.

Entropy extends to any number of random variables