Classical Information Theory

Notes from the lectures by prof Suhov

Trieste - june 2006

Fabio Grazioso ...

July 3, 2006

Contents

1 Lecture 1, Entropy

1.1 Random variable

Be X a random variable, and

$$\mathbb{P}(X = x_i) \tag{1}$$

be the probability that it's value is x_i. In some cases, for brevity we will write

$$\mathbb{P}(X = x_i) \equiv p(X). \tag{2}$$

1.2 Random string

Let's consider now a random variable that consists in a sequence of random variables. We can say that it's a random string, or a random array.
For example let's say that X it's the outcome of the toss of a coin, so $x_1 =$ "*tail*", and $x_2 =$ "*head*", or $x_1 = 1$, and $x_2 = 2$ for simplicity. Let's toss the coin m times, We will use the following formalism:

$$X_1^n \equiv (X_1, X_2, \dots, X_n) \tag{3}$$

We will write the possible outcomes as:

$$X_1^n = (x_1, x_2, \dots, x_n) = \begin{cases} (0, 0, 0, \dots, 0) \\ (0, 0, 0, \dots, 1) \\ \dots \\ (1, 1, 1, \dots, 1) \end{cases} \tag{4}$$

and the number of possible outcomes is 2^n (cfr. appendix (43))
There exist a particular way of ordering these outcomes, called "lexicographic", and then several other possible orderings.
If the coin is *fair*, then each
In the example of the coin, each single random variable can take only two walues.

If the coin is unfair, the probability of a string dipends on the number of 1s (and so on the number of 0s) that it contains. If we define:

$n_1 \equiv$ number of "1″ in the string

$n_0 \equiv$ number of "0″ in the string.

then the probability of a string to occour is
[...]

1.3 Markow's chains

Gambling has been the motivation for the progress of the classical probability theory. Let's now introduce a more general concept, that will be useful afterwards: the Markov's chain.
We have a Markov's chain if the result of a trial depends on the result of the previous:

$$\mathbb{P}(x_1^n = j_1^n) = \lambda_{j_1} p_{j_1 j_2} p_{j_2 j_3} \dots p_{j_{n-1} j_n}, \forall j_1^n \in I_m^n \tag{5}$$

Let's suppose to have a language, for example a text in an unknown language [...]

1.3.1 Stationary Markov chain

$$H(X) = -\sum_{j=1}^{m} p_j \ log_2 p_j \tag{6}$$

$$H(X) \geq 0 \tag{7}$$

$$h(X) = \lim_{n\to\infty} \frac{1}{n} H(X_1^n) \tag{8}$$

$$H(X) \leq log_2 m \text{ attained off } \mathbb{P}(X = x_i) = \frac{1}{m} \tag{9}$$

$$H(X_1^n) = -\sum_{x_1^n} \tag{10}$$

$$H(X) = h(X) \equiv -\sum_{j=1}^{m} p_j log_2 p_j = -\mathbb{E} \log_2 p_X(X) \tag{11}$$

$$x \to log_2 p_X(x) = log_2 \mathbb{P}(X = x) \tag{12}$$

1.4 Physical interpretation of entropy (t=27' 25")

The entropy is defined as

$$H = \sum_j p_j \ \log_2 p_j.$$

Now, how can we interpret and memorize the essence of this definition? Here I sum $p_j \log p_j$. It should remind you a mean value.
If I have a random variable, I can call it ξ, or I can call it X, I can write down the *mean value* of this random variable as:

$$\mathbb{E}[\xi] = \sum_i x_i \, \mathbb{P}(\xi = x_i)$$

where x_i are the values that the variable can take, and $\mathbb{P}(\xi = x_i)$ are the respective probabilities.
But I can also write:

$$\mathbb{E}[X] = \sum_i x_i \, \mathbb{P}(X = x_i)$$

So you can notice that the way how we denote the random variable it's not important! What is important is that the *possible values it can take* are assigned, and the *probability that they occur* are assigned.
Now, what is the random variable of which the definition of the entropy $-\sum_j \mathbb{P}(x_j) \ \log_2 \mathbb{P}(x_j)$ is the mean value? It is a quite difficult question, but it's difficult psycologically. This is (except for a sign) the expectation value of the random variable "$log_2\mathbb{P}(X)$", and **not** of $log_2 X$!!
The notation "$log_2\mathbb{P}(X)$" is selfexplanatory: $\log_2 \mathbb{P}(X)$ is the (logarithm of) *the probability* that X takes a value! What is psicologically difficult, here? We have a random variable, X, and it "produces" another random variable, that is $\mathbb{P}(X)$ wich is "the probability of X". X can take the values $\{x_1, x_2, \ldots, x_m\}$, and each of them has a probability $\{p(x_1), p(x_2), \ldots, p(x_m)\}$. The first is the set of the possible value of X, and the other is the set of the possible values of $\mathbb{P}(X)$. Of course the random variable $\mathbb{P}(X)$ is related

related, depends on the random variable X. It may seem that we have to know the value that has taken X to define $log_2\mathbb{P}(X)$, but it's not so!
We only need to know the expected probability of (all) the possible values that X can take, and this is an "a priori" property of the random variable X. In other words we define $\log_2 \mathbb{P}(X)$ just knowing the probability distribution of the random variable X, and we *don't need to know the actual outcomes* of X.
If we replace X with another variable, with the same probability distribution, the entropy it's the same. It's like we change the name of the variable in the previous example.
Since to define the mean value of a random variable we have seen that we just need the possible values it takes and their probability, we can write the mean value of $\log_2 \mathbb{P}(X)$.
(t 32'20") With this said, is now clear that the entropy is the *expected amount of information which you gain* from this random variable. Because from each value x_i you gain an amount of information $\log_2 p(x_i)$.
[...] that is: the mean value of the variable $-log_2(p_i)$

1.4.1 Other thoughts about entropy

Here I write down some thoughts developed by me and Alejandro.
Let's think we have a random variable, with different outcomes:

$$X = \begin{cases} x_1 = 1001010001001 \\ x_2 = 101 \\ x_3 = 10100101 \\ \dots \\ x_n = 10010. \end{cases}$$

The different outcomes have different lengths, and so there is needed a different amount of "means" to store or transmit them (amount of memory to store, or amount of "bandwidth" to transmit). Let's say we want to "encode" the possible outcomes, by numbering them: instead to store the actual outcome we store a numerical "label". Which is the smartest way to choose this encoding, to save the most space is possible? The labels

themselves have a different length, a different amount of memory (or bandwidth) needed, so the best is to first order the possible outcomes in a *decreasing probability order*, and number the most probable with the shortest numerical label, and the less probable with the longest numerical label!!

If $p_i = p(x_i) = p(X = x_i)$ is the probability of the outcome x_i, the "length of the label" will be something proportional to the inverse of the probability (the more is the probability, the less is the number):

$$\iota(x_i) \propto p(x_i)^{-1}$$

So, if we use this "smartest" way of storing the information, this quantity represents the amount of information "needed to store the outcome x_i". We can say it's the smallest amount of memory with which is possible to store it.

Once we have understood the reason why the information amount is directly proportional to the inverse of the probability, on the notes by Suhov (pag 2) we can also find a reason to use the logarithmic dependance.

The reason is that for two independent events, that is two outcomes of two independent random variables, we expect that the amount of information of the joint event is the sum of the amounts of information of the single events:

$$\iota(x_i \wedge y_j) = \iota(x_i) + \iota(y_j) \tag{13}$$

and since $\iota(x_i) \propto p(x_i)^{-1}$, we need a continue monotone function $\iota(x_i) = \phi[p(x_i)]$ such that

$$\phi[p(x_i \wedge y_j)] = \phi[p(x_i)] + \phi[p(y_j)]$$

but since the joint probability of independent events is the product of the probabilities of the single events, we have

$$\phi[p(x_i)p(y_j)] = \phi[p(x_i)] + \phi[p(y_j)]$$

Another property we wish is that if the event x_i has probability $p(x_i) = \frac{1}{2}$ we want $\iota(x_i) = 1$. The function that satisfy all those properties is the log. The only arbitrary choice is the base of the log, and we choose 2 becouse we think at the binary random variables:

$$\iota(x_i) = \log_2 p(x_i)^{-1}.$$

Finally, entropy is the expected value of this quantity, because we sum over all the possible outcomes, and we "weight" the sum with the probabilities:

$$\iota(x_i) = \log_2 p(x_i)^{-1}$$

1.4.2 Properties of entropy

$$0 \leq H(X) \leq \log_2 m \tag{14}$$

where m is the number of possible outcomes.

$$H(X) = 0 \Leftrightarrow X = const \tag{15}$$

$X = const$ means that the probability of one of the outcomes is 1, and the others are (of course) 0.

$$H(X) = \log_2 m \Leftrightarrow X \text{ is equiprobable} \tag{16}$$

indeed in this case we have

$$\begin{aligned} H(X) &= -\sum_{x_i} p(x_i) \log_2 p(x_i) \\ &= -\sum_{x_i} \frac{1}{m} \log_2 \frac{1}{m} \\ &= m \frac{1}{m} \log_2 m. \end{aligned}$$

1.5 Joint Entropy

Let's consider a "string", or "vector" of random variables

$$X_1^n \equiv \left\{X^{(1)}, X^{(2)}, \dots X^{(n)}\right\}.$$

The Entropy of X_1^n is

$$H(X_1^n) = -\sum_{\left\{x_i^{(1)}, x_j^{(2)}, \dots, x_k^{(n)}\right\}} p(X^{(1)}, X^{(2)}, \dots X^{(n)}) \ \log_2 p(X^{(1)}, X^{(2)}, \dots X^{(n)}). \quad (17)$$

For the simple case of a vector of just two random variables (X, Y) we have

$$H(X,Y) = -\sum_{x_i, y_j} p(X,Y) \ \log_2 p(X,Y). \quad (18)$$

Here the probability $p(X,Y) \equiv p(X = x_i, Y = y_j$ means the probability that occurs a certain couple of outcomes.
[...]

1.5.1 Properties of the joint entropy

$$H(X)\,,\, H(Y) \ \leq \ H(X,Y) \ \leq \ H(X) \ + \ H(Y) \quad (19)$$

1.6 Conditional Entropy

Let's consider now the conditional probability of a couple of variables:

$$p(X|Y). \quad (20)$$

The meaning of this probability is "the probability of the random variable X in a case when the variable Y is given". This means that we consider the situation in which we already know the outcome of the r.v. Y, since "the lecture" of it has happened, and we express the probability of the outcomes of the X r.v.. The point is that the fact that we know the outcome of Y can influence the probability distribution of X.
About this kind of probability, let's write down a couple of relations that will be useful in the following:

1. relation between joint probability and conditional probability:

$$p(X,Y) = p(X|Y)\,p(Y) \tag{21}$$

2. a strange way to write a probability:

$$p(Y) = \sum_{x_i} p(X,Y) \tag{22}$$

This stated, let's consider the following relation:

$$H(X|Y) = H(X,Y) - H(Y). \tag{23}$$

This relation is intuitive enough: if we have already known the r.v. Y, the entropy, i.e. the expected amount of information is the one of the couple (X,Y), from which we have to subtract the entropy of Y.
Is interesting to point out here that in this case the classical theory diverges from the quantum one! Indeed for the classical case this quantity is strictly positive, because of the previous relation (19), but in the classical chase we will see it is not! This is one of the interesting thing of the quantum information theory.
Let's try to expand the right term of the relation (23):

$$H(X,Y) - H(Y) \equiv -\sum_{x_i,y_j} p(X,Y)\ \log_2 p(X,Y) - \sum_{y_j} p(Y)\ \log_2 p(Y)$$

now we use the relation (22):

$$\begin{aligned}
&= -\sum_{x_i,y_j} p(X,Y)\ \log_2 p(X,Y) - \sum_{y_j} \left[\sum_{x_i} p(X,Y)\right]\ \log_2 p(Y) \\
&= -\sum_{x_i,y_j} p(X,Y)\ \log_2 p(X,Y) - \sum_{x_i,y_j} p(X,Y)\ \log_2 p(Y) \\
&= -\sum_{x_i,y_j} p(X,Y)\ [\log_2 p(X,Y) - \log_2 p(Y)]
\end{aligned}$$

and a log's property

$$= -\sum_{x_i,y_j} p(X,Y)\, \log_2 \frac{p(X,Y)}{p(Y)}$$

and the relation (21)

$$= -\sum_{x_i,y_j} p(X,Y)\, \log_2 p(X|Y). \tag{24}$$

1.6.1 Analysis of a wrong intuition

If you state that

$$H(X) = \sum_{x_i} p(X)\, \log_2 p(X)$$

and then that

$$H(X,Y) = \sum_{x_i,y_j} p(X,Y)\, \log_2 p(X,Y)$$

when you finally arrive at the conditional entropy you could be tempted to write:

$$H(X|Y) = -\sum_{x_i,y_j} p(X|Y)\, \log_2 p(X|Y). \tag{25}$$

Let's see what is this quantity, and in which way it's different form the conditional entropy.

When we write $(X|Y)$ we read it "X, known Y". This means that we know that $Y = y_j$ and we have the random variable X, but X has been conditioned in some how (it's probability distribution has changed) by the knowledge of Y.

In some sense, when we write $(X|Y)$ we should think at X as the variable, and at Y as a "parameter".

So, let's write

$$\tilde{H}_{y_j}(X) = -\sum_{x_i} p(X|Y)\, \log_2 p(X|Y). \tag{26}$$

Notice that the sum is only over x_i, and Y is a parameter. So this isn't the "erroneus quantity" $-\sum_{x_i,y_j} p(X|Y)\,\log_2 p(X|Y)$ (sum over both X and Y).
(Question: if you sum over the possible values of the variable X, does $\tilde{H}_{y_j}(X)$ still depend on X? Or not?!?)
What is this quantity $\tilde{H}_{y_j}(X)$? I still don't have a good idea, but Alejandro noticed that if you look at the expected value of it, you obtain the (correct) expression of $H(X|Y)$:

$$\begin{aligned}
\mathbb{E}[\tilde{H}_{y_j}(X)] &= \sum_{y_j} p(Y)\tilde{H}_{y_j}(X) \\
&= \sum_{y_j} p(Y) \left[\sum_{x_i} p(X|Y)\,\log_2 p(X|Y)\right] \\
&= \sum_{x_i,y_j} p(Y)p(X|Y)\,\log_2 p(X|Y) \\
&= \sum_{x_i,y_j} p(X,Y)\,\log_2 p(X|Y)
\end{aligned}$$

where, in the last passage, I have used the relation (21).
I have listened at the recording of Suhov's lecture, and he says the same thing... :-)

1.7 Mutual entropy

Let's define the *mutual* entropy by writing:

$$\begin{aligned}
H(X:Y) &\equiv H(X) + H(Y) - H(X,Y) &(27)\\
&= -\sum_{x_i,y_j} \mathbb{P}(X=x_i, Y=y_j\,\log_2 \frac{\mathbb{P}(X=x_i)\,\mathbb{P}(Y=y_j)}{\mathbb{P}(X=x_i, Y=y_j)}. &(28)
\end{aligned}$$

Since the definition is symmetrical, it's easy to see that:

$$H(X:Y) = H(Y:X). \tag{29}$$

1.8 Relative entropy

$$H(X||Y) \equiv -\sum_{x_i} \mathbb{P}(X = x_i) \log_2 \frac{\mathbb{P}(Y = x_i)}{\mathbb{P}(X = x_i)} \tag{30}$$

1.9 Entropy rate

Given a random string of lenght n: $X_1^n = \{X^1, X^2, \ldots, X^n\}$ we define

$$h \equiv \lim_{n\to\infty} \frac{1}{n} H(X_1^n). \tag{31}$$

1.9.1 particular cases

The entropy rate of a sequence of independent random variables is:

$$h = H(X^1). \tag{32}$$

The entropy rate of a stationary Markov chain is:

$$h = H(X^2|X^1). \tag{33}$$

1.10 Comments

- The concept of entropy is not so much useful in the case of a single random variable. It's use is mainly in the study of a sequence of random variables.
- following the concept of the entropy of a single random variable, are introduced similar concepts for random strings; we have
  - joint entropy
  - conditional entropy
  - mutual entropy
  - relative entropy.

  One of the most elusive concept, although one of the most useful, is the conditional entropy.

- the key to understand the properties of these quantities is relation (19).
- there are relations about conditional, mutual and relative similar to (19), that is about joint entropy (see Suhof notes).

2 Lecture 2 - Shannon's 1^{st} coding theorem

Entropy rate

$$h = \lim_{n\to\infty} \frac{1}{n} H(X_1^n) \tag{34}$$

For example, in coin tossing

$$p > \frac{1}{2} > 1 - p \tag{35}$$

then

$$x_1^n = (1, 1, \ldots, 1) \qquad (all1s) \tag{36}$$

$$\lim_{n\to\infty} \frac{1}{2} log_2 M_n(\varepsilon) = h \tag{37}$$

$$\frac{1}{n} log_2 M_n(\varepsilon) \to h \Leftrightarrow M_n(\varepsilon)\tilde{2}^{n\,h} << 2^{n\,log_2 m} m^h \tag{38}$$

$$\mathbb{X}_1^n = x_1^n \tilde{2}^{-n\,h} \tag{39}$$

M_n = number of strings that you have to select in order to have that ... is less than ε
$2^{n\,h}$ = number of strings of lenght $n\,h$
by the law of the large numbers

$$\frac{n_1}{n} \to p\,, \frac{n_0}{n} \to (1-p) \tag{40}$$

Shannon:
you have a string of lenght n. suppose that by incrising to $R^{-1}\,n$, you have a chance to beat of errors in the channel.

The maximum value of R is the channel capacity.

If this is possible, then R is called a *reliable transmission rate*.

Take $C = \sup[R : R^{-1} \quad \text{reliable}]$ then $C =$ the channel capacity

• Memoryless channels: distort every digit indipendently.

then it is clear that this channel is characterized by the channel matrix Π

$$\mathbf{\Pi} = \begin{pmatrix} (1-p) & p \\ p & (1-p) \end{pmatrix}$$

Binary Simmetric Channel

For a BSC

$$C = 1 - p \log_2 p \ - (1-p) \log_2 (1-p) \tag{41}$$

X_1^n want to increse n to R_n^{-1}

then there exixst an enchoding and a decoding such that the error probability can be made small then R is challed a reliable transmission sote.

A Combinatory Theory

A.1 Dispositions without repetitions (or simple dispositions)

A *disposition without repetitions* is a sequence of objects, took from a set of possible symbols, where is not possible to repeat many times the same object. The order matters, i.e. sequences with the same elements in a different order are different dispositions. The number of possible dispositions without repetitions, with length n and choosing in a set of k possible objects is:

$$D_k^n = P_k^n = \frac{n!}{(n-k)!} \tag{42}$$

A.1.1 proof

The number of possible choices for the first symbol is k. The number of possible choices for the second symbol is $k-1$ (since one symbol is gone and we cannot repeat it). For a sequence of n symbols the possibilities are the product of these possibilities:

$$k\,(k-1)\,\ldots,(k-n+1)$$

we can represent this quantity as a ratio of factorials: we write $k! = k(k-1)(k-2)\ldots 1$, and then we divide for the product of the "unwanted terms":

$$k\,(k-1)\,\ldots,(k-n+1) = \frac{k(k-1)(k-2)\ldots(k-n+1)(k-n)\ldots 1}{(k-n)(k-n-1)\ldots 1} = \frac{k!}{(k-n)!}.$$

QED
It's possible to show that permutations are special cases of simple dispositions.

A.2 Dispositions with repetitions

A *disposition with repetitions* is a sequence of objects, took from a set of possible symbols, where is possible to repeat many times the same object. The order matters, i.e. sequences with the same elements in a different order are different dispositions. The number of possible dispositions with repetitions, with length n and choosing in a set of k possible objects is:

$$DR_k^n = k^n \tag{43}$$

A.2.1 proof

The possibilities for the choice of the first element are k, the possibilities for the choice of the second element are k, and so on. For a sequence of n symbols the possibilities are the product of these possibilities. QED

A.3 Permutations

A.4 Combinations without repetitions (or simple combinations)

A combination is a sequence of symbols taken out from a set of possible, where the order doesn't matters, i.e. sequences with the same elements in a different order are considered the same combination. A simple one is when is not allowed to repeat a symbol. The number of possible simple combinations, with length n and choosing in a set of k possible objects is:

$$C_k^n = \frac{n!}{k\,(n-k)!} \equiv \left(\begin{array}{c} n \\ k \end{array} \right) \tag{44}$$

A.4.1 proof

A.5 Combinations with repetitions

A combination is a sequence of symbols taken out from a set of possible, where the order doesn't matters, i.e. sequences with the same elements in a different order are considered the same combination. In this case is allowed to repeat a symbol. The number of possible combinations with repetitions, with length n and choosing in a set of k possible objects is:

$$CR_n^k = \frac{(n+k-1)!}{k!\,(n-1)!} \equiv \left(\begin{array}{c} n+k-1 \\ k \end{array} \right) \tag{45}$$

A.5.1 proof