INF264 - Exercise 3

Pierre Gillot   Natacha Galmiche[*]

Week 36 - 2021

In the following exercise, we will refresh our memory on the concept of entropy from information theory, then we will construct a toy-example decision tree on a small dataset.

1 Basic properties of entropy

Let X a random discrete variable taking values in $\{x_0, \dots, x_{N-1}\}$. Denote by p_X the probability distribution of X:

$$X \sim p_X = \big\{P(x_0), \dots, P(x_{N-1})\big\} = \{p_0, \dots, p_{N-1}\}. \tag{1}$$

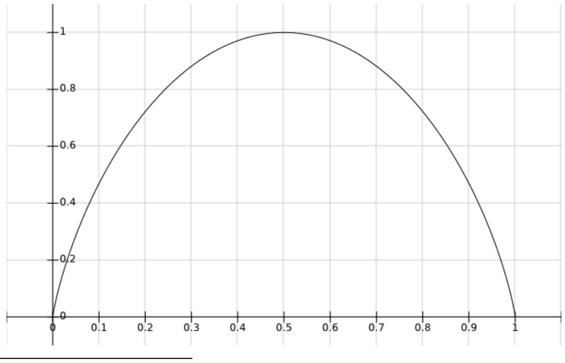
The definition of the entropy of X is:

$$H(X) = -\sum_{n=0}^{N-1} p_n \log_2(p_n). \tag{2}$$

For the sake of simplicity and visualization purpose, let us consider the special case where $N = 2$, i.e X follows a Bernoulli distribution. We can write $p_0 = p$ and $p_1 = 1 - p_0 = 1 - p$, thus the entropy of $X \sim \mathcal{B}_p$ is simply:

$$H(\mathcal{B}_p) = -p \log_2(p) - (1-p) \log_2(1-p). \tag{3}$$

We can therefore plot the function $p \mapsto H(\mathcal{B}_p)$ in order to visualize what is the entropy in the Bernoulli case:



[*] Not a TA for this course this year.

This function increases on $[0, \frac{1}{2}]$, reaches its maximum at $p = \frac{1}{2}$ and decreases on $[\frac{1}{2}, 1]$. This means entropy increases when the distribution of X converges to the uniform distribution, which models complete disorder.

More generally for any N, it is possible to show that the entropy of X is maximal if and only if X follows a uniform distribution, that is if and only if $p_0 = \cdots = p_{N-1} = \frac{1}{N}$. Moreover, the maximal entropy (modelling complete disorder) attained by the uniform distribution is equal to $\log_2(N)$. **The latter statement is important for the next section !**

1. Verify that if X follows a uniform distribution, then $H(X) = \log_2(N)$. **Hint:** Start with equation (2) and replace the p_n by their value in the special case where X follows a uniform distribution.

2 Entropy and information in cards

Consider a traditional 52-card deck, with 4 colors (hearts, diamonds, clubs and spades) and 13 ranks (from highest to lowest: ace, king, queen, jack and numbered ranks from 10 to 2), where each rank comes in every color. Suppose the deck was meticulously shuffled and is placed face-down. We are interested in the entropy of the random variable T that models the top card's identity (rank and color).

2. What is the probability distribution of T ? Deduce the value of its entropy.

Consider now the random variable A that models the event E_A: "Top card is an ace". We want to measure the impact of the information carried by A on the entropy of T.

3. What is $P(E_A)$? Deduce the probability distribution of A.

4. What are the probability distributions of $T|A{=}True$ and $T|A{=}False$? Deduce the value of their respective entropy $H(T|A{=}True)$ and $H(T|A{=}False)$.

5. The entropy of T given information A is defined as the mean over the values a taken by A of $H(T|A{=}a)$, that is in our case:

$$H(T|A) = P(A{=}True)H(T|A{=}True) + P(A{=}False)H(T|A{=}False) \tag{4}$$

   Imagine a friend of yours draws the top card, looks at it without revealing anything and tells you that you can ask any "Yes or No" question about its identity and he will answer (truthfully) to it. Information A in this context can be understood as you asking the question "Is the top card an ace ?". Explain in your own words the physical interpretation of equation (4) in the case of our variables T and A.

6. Compare the entropies $H(T)$ and $H(T|A)$. How did the addition of the information carried by A affect the entropy of T ?

7. The quantity $IG(T|A) = H(T) - H(T|A)$ is called the information gain of T given information A. For A to be a valuable source of information about T, should $IG(T|A)$ be small or large ?

8. Is it more informative to first ask the question "Is the top card an ace ?" or the question "Is the top card a spades ?" when trying to identify the top card ? **Hint:** You should create a random variable S that models the event E_S: "Top card is a spades" and compute the information gain of T given information S given by $IG(T|S) = H(T) - H(T|S)$.

3 Binary decision trees

Consider the following training dataset:

X_0	X_1	X_2	Y
0	1	0	A
0	0	0	B
1	0	0	B
0	1	1	A
0	0	0	B
1	1	1	B
1	0	1	B
0	0	1	A
0	1	0	A
1	1	1	B

9. Construct a decision tree on the training dataset using the entropy as a metric: iteratively split on the feature X_i among those remaining such that the information gain is maximized by X_i. Make your reasoning and computations apparent. To gain some time, you can use the following approximations:

$$\begin{aligned} &-\tfrac{4}{5}\log_2\left(\tfrac{4}{5}\right) - \tfrac{1}{5}\log_2\left(\tfrac{1}{5}\right) \simeq 0.722 \\ &-\tfrac{2}{3}\log_2\left(\tfrac{2}{3}\right) - \tfrac{1}{3}\log_2\left(\tfrac{1}{3}\right) \simeq 0.918 \\ &-\tfrac{3}{5}\log_2\left(\tfrac{3}{5}\right) - \tfrac{2}{5}\log_2\left(\tfrac{2}{5}\right) \simeq 0.971 \end{aligned} \tag{5}$$

Now consider the following "pruning" dataset:

X_0	X_1	X_2	Y
0	0	1	B
0	0	0	A
0	0	1	A
0	0	1	B
0	1	0	A

10. Use this pruning dataset to perform reduced-error post-pruning of your decision tree: in a recursive bottom-up fashion, compare the error rate on the pruning set between the decision tree at its present state and the modified decision tree where the subtree rooting at the currently investigated node is replaced by the dominant label in the training set at this node. If the error rate is reduced with the modification, the decision tree becomes the modified decision tree.