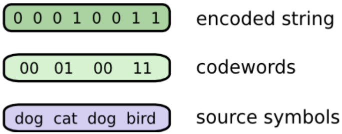
Introduction to Machine Learning

Information Theory
Entropy and Optimal Code Length



Learning goals

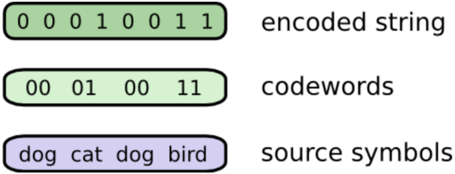
- Know that source coding is about encoding messages efficiently
- Know how to compute the average length of a code
- Know that the entropy of the source distribution is the lower bound for the average code length

SOURCE CODING

- There is an interesting connection between entropy and a subfield of information theory known as **source coding**.
- Abstractly, a source is any system or process that generates messages or information.
- A code is simply a way to represent the message so that it can be stored or transmitted over a communication channel (such as radio or fiber-optic cables).
- For example, one could use binary strings (0's and 1's) to encode messages.
- Because it may be expensive to transmit or store information, an important problem addressed by source coding is efficient coding schemes of minimal average length.

SOURCE CODING / 2

- Formally, given a discrete alphabet/dictionary X of message symbols, a **binary code** is a mapping from symbols in X to a set of codewords of binary strings.
- For example, if our dictionary only consists of the words "dog", "cat", "fish" and "bird", each word can be encoded as a binary string of length 2 : "dog" $\rightarrow$ **00**, "cat" $\rightarrow$ **01**, "fish" $\rightarrow$ **10** and "bird" $\rightarrow$ **11**.
- For this code, a binary string can be decoded by replacing each successive pair of digits with the associated word.

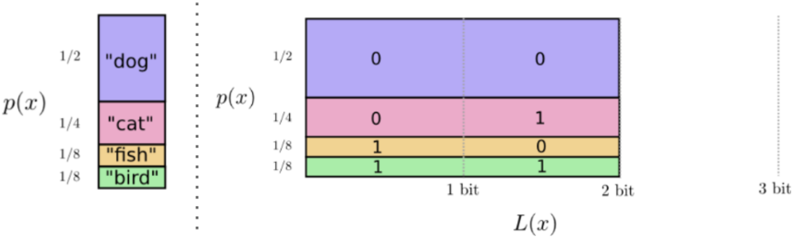


Credit: Chris Olah

Chris Olah (2015): Visual Information Theory. *http://colah.github.io/posts/2015-09-Visual-Information/*

SOURCE CODING / 3

- Encoded messages are emitted by a source which can be modeled as a probability distribution over the message symbols in the dictionary.
- Let X be a random variable that represents a symbol from our data source and let $p(x) = \mathbb{P}(X = x)$, for symbol x in our dictionary.

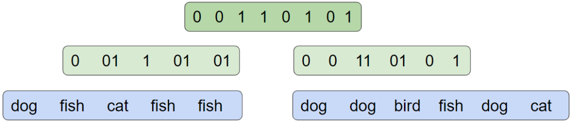


Credit: Chris Olah

- Length $L(x)$ is simply number of bits in corresponding codeword. Here all codewords have length 2.
- Area of rectangles on the right reflect contributions to $\mathbb{E}[L(X)]$

SOURCE CODING / 4

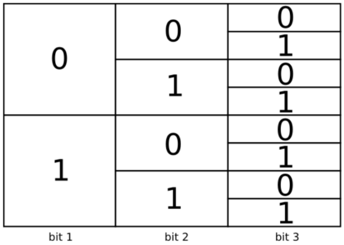
- Maybe we can create better average-length coding schemes with **variable-length** codes by assigning shorter codes to more likely messages and longer one to less likely messages.
- However, this can be problematic because we want the receiver to be able to unambiguously decode the encoded string.
- Let us say the words in our dictionary are encoded in this way: "dog" $\rightarrow$ **0**, "cat" $\rightarrow$ **1**, "fish" $\rightarrow$ **01** and "bird" $\rightarrow$ **11**.
- In this case, the string 00110101 can be decoded in multiple ways.



- One way to make variable-length messages unambiguous is by ensuring that no codeword is a prefix (initial segment) of any other codeword. Such a code is known as a **prefix code**.

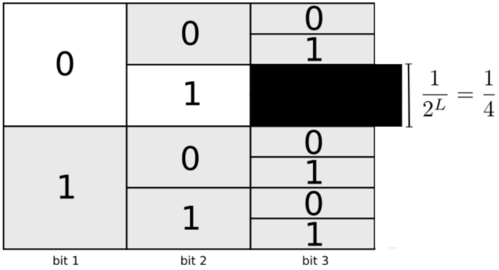
SOURCE CODING / 5

- In general, the number of possible codewords grows exponentially in length L.
- For binary codes, there are two possible words of length one, four possible words of length two and 2^L possible words of length L.



- In total, there are $(2^{L+1} - 2)$ codewords of length $\leq$ L.

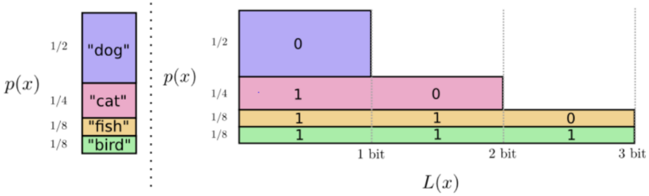
SOURCE CODING / 6



- Here, if the codeword **01** is assigned to a symbol, then **010** and **011** cannot be assigned to any other symbol because that would break the prefix property.
- If a codeword of length L is assigned to a symbol, then $\frac{1}{2^L}$ of the possible codewords of length $> L$ must be discarded.
- If some symbols are assigned short codewords, due to the prefix property, many marginally longer codewords cannot be assigned to other symbols.

SOURCE CODING / 7

- An example of prefix code:
  "dog" $\rightarrow$ **0**, "cat" $\rightarrow$ **10**, "fish" $\rightarrow$ **110** and "bird" $\rightarrow$ **111**.



- Here, the expected code length is :

$$
\begin{aligned}
\mathbb{E}[L(X)] &= \frac{1}{2} \cdot 1 + \frac{1}{4} \cdot 2 + \frac{1}{8} \cdot 3 + \frac{1}{8} \cdot 3 \\
&= -\frac{1}{2} \cdot \log_2 \left(\frac{1}{2}\right) - \frac{1}{4} \cdot \log_2 \left(\frac{1}{4}\right) - \frac{1}{8} \cdot \log_2 \left(\frac{1}{8}\right) - \frac{1}{8} \cdot \log_2 \left(\frac{1}{8}\right) \\
&= H(X) = 1.75 \text{ bits. } (< 2 \text{ bits})
\end{aligned}
$$

SOURCE CODING / 8

- Actually, this coding scheme is the most efficient way to store and transmit these messages. It is simply not possible to do better!
- In fact, Shannon's **source coding theorem** (or **noiseless coding theorem**) tells us that the optimal trade-off is made when the code length of a symbol with probability p is $\log(1/p)$.
- In other words, the entropy of the source distribution is the theoretical lower bound on the average code length.
- If it is any lower, some information will be distorted or lost.
- In practice, algorithms such as Huffman Coding can be used to find variable-length codes that are close (in terms of expected length) to the theoretical limit.