# Info theory (TN) notes

Kirsty McNaught

2017

# 1 Lots of juicy terms

## 1.1 Entropy $H(S)$

"How many bits do I need to record the value of s?"

$$H[S] = \sum_i P(s) \log_2 P(s)$$

### 1.1.1 Some examples

- Uniform distribution, 2 discrete values: **1 bit**
- Uniform distribution, 16 discrete values: **4 bits**
- 16 discrete values, but most of probability density on 8 values: **just over 2 bits**
- A sequence of M entries, each one of N equiprobable values: $\boldsymbol{\log_2 M^N = N \log_2 M}$ **bits**

### 1.1.2 A derivation

We can derive the above definition for the entropy by considering how many bits would be required to encode a sequence of M entries, each one of N possible values, but which are not equiprobable. The likelihood of a given sequence is:

$$P(S_1, S_2, ..., S_3) = \prod_m p_m^{n_m} \tag{1}$$

where $n_m$ is the number of times the m-th value occurred. Now we use the asymptotic equipartition property, which says that the only set of sequences with non-zero probability in the limit of large $M$ is those for which all $n_m = p_m N$, i.e. the non-surprising, or *typical* sequences. We consider the probability of all other (surprising) sequences to be infinitesimally small. All typical sequences have equal likelihood, so we are back in the easy situation of a uniform distribution. The number of bits required to encode these sequences is:

$$\begin{aligned}
N &= \log_2 \left[ \frac{1}{\prod_m p_m^{p_m N}} \right] \\
&= -\log_2 \prod_m p_m^{p_m N} \\
&= -N \sum_m p_m \log_2 p_m \\
&= NH[S]
\end{aligned}$$

## 1.2 Conditional entropy $H(S \mid R)$

"How many bits do I need to record the value of s, given that I already know r?"

Given a particular value of r, the conditional entropy $H(S \mid r)$ is given by

$$H(S \mid r) = -\sum_s P(s \mid r) \log_2 P(s \mid r)$$

The overall conditional entropy must be averaged over all possible values of $r$, i.e.

$$\begin{aligned}
H(S \mid R) &= \sum_r H(S \mid r) \\
&= -\sum_r P(r) \sum_s P(s \mid r) \log_2 P(s \mid r) \\
&= -\sum_{s,r} [P(r)P(s \mid r)] \log_2 P(s \mid r) \\
&= -\sum_{s,r} P(s,r) \log_2 P(s \mid r)
\end{aligned}$$

## 1.3 Mutual information $I(S;R)$

"How much information is gained about S if I am told R"

This can be interpreted as the reduction of entropy about S caused by finding out about R, i.e.

$$I(S;R) = H(S) - H(S \mid R)$$

We can also derive a direct definition:

$$\begin{aligned}
I(S;R) &= H(S) - H(S \mid R) \\
&= -\sum_s P(s) \log_2 P(s) + \sum_{s,r} P(s,r) \log_2 P(s \mid r) \\
&= -\sum_{s,r} P(s,r) \log_2 P(s) + \sum_{s,r} P(s,r) \log_2 P(s \mid r) \\
&= \sum_{s,r} P(s,r) \log_2 \frac{P(s \mid r)}{P(s)} \\
&= \sum_{s,r} P(s,r) \log_2 \frac{P(s,r)}{P(s)P(r)}
\end{aligned}$$

## 1.4 Cross entropy

"How many bits do I need to encode a value using the p.d.f. of Q instead of (true) p.d.f. P?"

$$H_{\mathrm{x}}(P,Q) = -\sum_s P(s) \log_2 Q(s)$$

## 1.5 KL divergence

"How many excess bits does it cost me to encode using p.d.f. of Q instead of (true) p.d.f. P?"

$$\begin{aligned}
KL[P;Q] &= H_{\mathrm{x}}(P,Q) - H[P] \\
&= -\sum_s P(s) \log_2 Q(s) + \sum_s P(s) \log_2 P(s) \\
&= -\sum_s P(s) \log_2 \frac{P(s)}{Q(s)}
\end{aligned}$$

**Mutual information** can be described as the **KL divergence** between the joint distribution $P(S, R)$ and the marginals $P(S)P(R)$, i.e. the cost of encoding using the marginals if you don't know the joint.

$$I[S; R] = \sum_{s,r} P(s, r) \log_2 \frac{P(s, r)}{P(s)P(r)} = KL[P(S, R) \mid\mid P(S)P(R)]$$

## 1.6 Relationships between these quantities

With two variables, we can draw a Venn diagram that summarises the relationship between each quantity.