CSE 5523: Lecture Notes 6
Information Theory

Contents

6.1 A Formal definition of information [Shannon, 1948]

We can formalize the contribution of learning as **information**.

In this sense, information about a distribution makes it more predictable.

For example, if you think a promotion is [.50, .50], then learn you got it, your distribution is [0, 1].

This is a **bit** of information: the difference between no knowledge and certainty of a Bernoulli trial.

They are the bits you'd use to optimally encode probability-weighted outcomes of a distribution.

For example, with a distribution [0.5, 0.25, 0.125, 0.125] the optimal encoding is ***not*** this:

event	freq.	code	cost
A	500	00	$500 \times 2 =$ 1000 bits
B	250	01	$250 \times 2 =$ 500 bits
C	125	10	$125 \times 2 =$ 250 bits
D	125	11	$125 \times 2 =$ 250 bits
	1000		**2000 bits**

but rather ***this***, with variable length tokens, inversely proportional to the log of the probability:

event	freq.	code	cost
A	500	0	$500 \times 1 =$ 500 bits
B	250	10	$250 \times 2 =$ 500 bits
C	125	110	$125 \times 3 =$ 375 bits
D	125	111	$125 \times 3 =$ 375 bits
	1000		**1750 bits**

To encode 1000 outcomes, you use only 1750 bits instead of 2000!

If an event ***always*** happens, you give it zero bits – the receiver ***already knows*** the outcome.

If an event ***never*** happens, you don't give it a code – you can't send it, but you ***won't need to***.

Formally, the information (in bits) of an event is the negative log of its probability:

$$\mathsf{I}_{p_1,p_2,\ldots}(x) = -\log_2 \mathsf{P}_{p_1,p_2,\ldots}(x)$$

If an event has probability 1, it has information 0 (use most efficient code imaginable: nothing!).

If an event has probability 0, it has information ∞ (use least efficient code imaginable: everything!).

This is called **self-information** or **surprisal**. It's the information ***of the event, given a distribution***.

6.2 Entropy

The expected information of a distribution is then:

$$\begin{aligned}
\mathsf{H}(X) = \mathsf{H}(\mathsf{P}_{p_1,p_2,\ldots}(X)) &= \mathsf{E}_{x\sim\mathsf{P}_{p_1,p_2,\ldots}(X)}\, \mathsf{I}_{p_1,p_2,\ldots}(x) \\
&= \mathsf{E}_{x\sim\mathsf{P}_{p_1,p_2,\ldots}(X)}(-\log_2 \mathsf{P}_{p_1,p_2,\ldots}(x)) && \text{definition of self-information} \\
&= -\sum_{x\in X} \mathsf{P}_{p_1,p_2,\ldots}(x) \log_2 \mathsf{P}_{p_1,p_2,\ldots}(x) && \text{definition of expected value}
\end{aligned}$$

This is also called the **entropy** (from Greek 'entropia' roughly meaning 'disorder' or 'chaos').

Indeed, expecting lots of information indicates chaos; expecting no information indicates order.

(Why abbreviate entropy as H? It's a capital Greek eta η, pronounced 'eh', as in 'eh'ntropy.)

And here's the entropy of our promotion distribution, before and after finding out:

$$\begin{aligned}
&\mathsf{H}([.5, .5]) && = .5 \cdot 1 + .5 \cdot 1 && = 1 \\
&\mathsf{H}([0, 1]) && = 0 \cdot \infty + 1 \cdot 0 && = 0
\end{aligned}$$

6.3 Cross entropy and Kullback-Leibler (KL) divergence

In defining loss functions for parameters, it's useful to quantify how wrong a distribution Q is.

First, using distribution Q on data distributed according to P has the following information:

$$\begin{aligned}
\mathsf{H}(P, Q) &= -\mathsf{E}_{x\sim P(X)} \log_2 Q(x) \\
&= -\sum_{x\in X} P(x) \log_2 Q(x) && \text{definition of expected value}
\end{aligned}$$

(Here we assume P and Q share the same event space, but are not in the same probability space.)

This is called **cross entropy**.

Using a different distribution is ***always worse*** (optimality of maximum likelihood estimation).

The loss in expected information from using Q instead of P on data distributed according to P is:

$$\mathsf{D}_{\mathrm{KL}}(P \,\|\, Q) = \mathsf{H}(P, Q) - \mathsf{H}(P)$$

$$= \Big(-\mathsf{E}_{x\sim P(X)} \log_2 Q(x) \Big) - \Big(-\mathsf{E}_{x\sim P(X)} \log_2 P(x) \Big) \qquad \text{definition of (cross) entropy}$$

$$= \Big(-\sum_{x\in X} P(x) \log_2 Q(x) \Big) - \Big(-\sum_{x\in X} P(x) \log_2 P(x) \Big) \qquad \text{definition of expected value}$$

$$= -\sum_{x\in X} P(x) \Big(\log_2 Q(x) - \log_2 P(x) \Big) \qquad \text{distributive axiom}$$

$$= -\sum_{x\in X} P(x) \log_2 \frac{Q(x)}{P(x)} \qquad \text{addition of logs}$$

This is called **Kullback-Leibler (KL) divergence** or **relative entropy**.

It's zero (log of one) when the distributions match, and positive when they don't.

6.4 Conditional entropy and mutual information

Sometimes it's valuable to see how much information two variables share.

First, loss in expected information from using $\mathsf{P}(X)$ instead of $\mathsf{P}(X, Y)$ on distribution $\mathsf{P}(X, Y)$ is:

$$\mathsf{H}(Y \mid X) = \mathsf{D}_{\mathrm{KL}}(\mathsf{P}(X, Y) \,\|\, \mathsf{P}(X))$$

$$= \Big(-\mathsf{E}_{x,y\sim\mathsf{P}(X,Y)} \log_2 \mathsf{P}(x) \Big) - \Big(-\mathsf{E}_{x,y\sim\mathsf{P}(X,Y)} \log_2 \mathsf{P}(x, y) \Big) \qquad \text{def. of KL divergence}$$

$$= \Big(-\sum_{x,y\in X\times Y} \mathsf{P}(x, y) \log_2 \mathsf{P}(x) \Big) - \Big(-\sum_{x,y\in X\times Y} \mathsf{P}(x, y) \log_2 \mathsf{P}(x, y) \Big) \qquad \text{def. of expected value}$$

$$= -\sum_{x,y\in X\times Y} \mathsf{P}(x, y) \Big(\log_2 \mathsf{P}(x) - \log_2 \mathsf{P}(x, y) \Big) \qquad \text{distributive axiom}$$

$$= -\sum_{x,y\in X\times Y} \mathsf{P}(x, y) \log_2 \frac{\mathsf{P}(x)}{\mathsf{P}(x, y)} \qquad \text{addition of logs}$$

$$= \sum_{x,y\in X\times Y} \mathsf{P}(x, y) \log_2 \frac{\mathsf{P}(x, y)}{\mathsf{P}(x)} \qquad \text{log of inverse}$$

$$= \sum_{x,y\in X\times Y} \mathsf{P}(x, y) \log_2 \mathsf{P}(y \mid x) \qquad \text{addition of logs}$$

(Here we assume $\mathsf{P}(X, Y)$ and $\mathsf{P}(X)$ are in the same probability space.)

This is called **conditional entropy**.

When $\mathsf{P}(X)$ is predictive of $\mathsf{P}(X, Y)$ (e.g. X and Y are correlated), this loss is small; otherwise big.

Loss in expected information from using $\mathsf{P}(X)\cdot\mathsf{P}(Y)$ instead of $\mathsf{P}(X, Y)$ on distribution $\mathsf{P}(X, Y)$ is:

$$\mathsf{I}(X; Y) = \mathsf{D}_{\mathrm{KL}}(\mathsf{P}(X, Y) \,\|\, \mathsf{P}(X)\cdot\mathsf{P}(Y))$$

$$= \Big(-\mathsf{E}_{x,y\sim\mathsf{P}(X,Y)} \log_2 \mathsf{P}(x)\cdot\mathsf{P}(y) \Big) - \Big(-\mathsf{E}_{x,y\sim\mathsf{P}(X,Y)} \log_2 \mathsf{P}(x, y) \Big) \qquad \text{def. of KL divergence}$$

$$
\begin{aligned}
&= \Big(-\sum_{x,y \in X\times Y} \mathsf{P}(x,y) \log_2 \mathsf{P}(x)\cdot\mathsf{P}(y) \Big) - \Big(-\sum_{x,y \in X\times Y} \mathsf{P}(x,y) \log_2 \mathsf{P}(x,y) \Big) && \text{def. of expected value} \\
&= -\sum_{x,y \in X\times Y} \mathsf{P}(x,y) \Big(\log_2 \mathsf{P}(x)\cdot\mathsf{P}(y) - \log_2 \mathsf{P}(x,y) \Big) && \text{distributive axiom} \\
&= -\sum_{x,y \in X\times Y} \mathsf{P}(x,y) \log_2 \frac{\mathsf{P}(x)\cdot\mathsf{P}(y)}{\mathsf{P}(x,y)} && \text{addition of logs} \\
&= -\sum_{x,y \in X\times Y} \mathsf{P}(x,y) \Big(\log_2 \mathsf{P}(x) + \log_2 \frac{\mathsf{P}(y)}{\mathsf{P}(x,y)} \Big) && \text{addition of logs} \\
&= \Big(-\sum_{x,y \in X\times Y} \mathsf{P}(x,y) \log_2 \mathsf{P}(x) \Big) + \Big(-\sum_{x,y \in X\times Y} \mathsf{P}(x,y) \log_2 \frac{\mathsf{P}(y)}{\mathsf{P}(x,y)} \Big) && \text{distributive axiom} \\
&= \Big(-\sum_{x,y \in X\times Y} \mathsf{P}(x,y) \log_2 \mathsf{P}(x) \Big) - \Big(-\sum_{x,y \in X\times Y} \mathsf{P}(x,y) \log_2 \frac{\mathsf{P}(x,y)}{\mathsf{P}(y)} \Big) && \text{log of inverse} \\
&= \Big(-\sum_{x \in X} \mathsf{P}(x) \log_2 \mathsf{P}(x) \Big) - \Big(-\sum_{x,y \in X\times Y} \mathsf{P}(x,y) \log_2 \frac{\mathsf{P}(x,y)}{\mathsf{P}(y)} \Big) && \text{marginalization} \\
&= \mathsf{H}(X) - \mathsf{H}(X \mid Y) && \text{def. of (conditional) entropy}
\end{aligned}
$$

This is called **mutual information**. Unlike conditional entropy, it is symmetric.

When X and Y are independent, it's low; otherwise it's high.

(Note this can differ from conditional entropy, e.g. if X is more fine-grained than Y.)

References

[Shannon, 1948] Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 623–656.