Source coding

Last modified: 2026-11-01 19:38:09 (PST)

Codes

Definition 0.1 (Binary code)**.** Let X be a discrete random variable. A **binary code** for X assigns each value x a *codeword*, a finite string of binary digits, of length $\ell(x)$.

Definition 0.2 (Uniquely decodable code)**.** A binary code is **uniquely decodable** if every finite sequence of codewords, written one after another, can be split back into codewords in only one way.

Definition 0.3 (Expected length)**.** The **expected length** of a binary code for X, where X has probability mass function $p(x) \stackrel{\text{def}}{=} \mathbb{P}(X = x)$, is

$$L \stackrel{\text{def}}{=} \mathbb{E}[\ell(X)] = \sum_x p(x)\,\ell(x).$$

Remark 0.1 (Short codewords for common values). A code with a small expected length gives frequent values short codewords and rare values long ones, since each value's length counts in L in proportion to its probability.

The source coding theorem

Theorem 0.1 (Source coding theorem)**.** *Let X be a discrete random variable that takes finitely many values, at least two of them with positive probability, and let $H_2(X)$ be its entropy with base-2 logarithms.*

1. *Every uniquely decodable binary code for X has expected length $L \geq H_2(X)$.*

2. *Some uniquely decodable binary code for X has expected length $L < H_2(X) + 1$.*

3. *Coding blocks of n independent copies of X at a time, some uniquely decodable binary code spends, on average, fewer than $H_2(X) + 1/n$ binary digits per copy.*

Remark 0.2 (Where the theorem comes from). Shannon (1948) proved that entropy is the limit on how far the output of a source can be compressed. Cover and Thomas (2006, sec. 5.3 to 5.5) prove each part of Theorem 0.1. For part 2, they give each value x with $p(x) > 0$ the length $\ell(x) = \lceil -\log_2 p(x) \rceil$, its surprisal in bits rounded up; the Kraft inequality (Cover and Thomas 2006, sec. 5.2) guarantees that a code with these lengths exists.

Coding with the wrong distribution

Theorem 0.2 (The cost of coding with the wrong distribution)**.** *Let X take finitely many values, with probability mass function p, and let q be a probability mass function, possibly different from p, with $q(x) > 0$ wherever $p(x) > 0$. Suppose a code gives each value x with $p(x) > 0$ the length $\ell(x) = \lceil -\log_2 q(x) \rceil$. Then its expected length L under p satisfies*

$$H_2(p, q) \le L < H_2(p, q) + 1,$$

where $H_2(p, q)$ is the cross-entropy of q relative to p, with base-2 logarithms.

Proof

Proof. Every sum here runs over the values x with $p(x) > 0$. Rounding up adds less than one binary digit, so for each such x

$$-\log_2 q(x) \le \ell(x) < -\log_2 q(x) + 1.$$

Multiply each part by $p(x) > 0$ and sum. Since $\sum_x p(x) = 1$,

$$-\sum_x p(x) \log_2 q(x) \le \sum_x p(x)\, \ell(x) < -\sum_x p(x) \log_2 q(x) + 1.$$

The middle sum is L, and the outer sums are $H_2(p, q)$. □

Remark 0.3 (Cross-entropy is the cost of the wrong code). Write $H_2(p)$ for the entropy of a random variable with mass function p, and $D_2(p \,\|\, q)$ for the Kullback–Leibler divergence of q from p, both with base-2 logarithms. Cross-entropy splits into those two parts:

$$H_2(p, q) = H_2(p) + D_2(p \,\|\, q).$$

By Theorem 0.1, no uniquely decodable code averages fewer than $H_2(p)$ bits. By Theorem 0.2, a code built for q costs about $D_2(p \,\|\, q)$ bits per value more than that. The divergence is never negative, and it is 0 only when $q = p$ (Cover and Thomas 2006, sec. 2.6), so the excess vanishes only when the code was built for the true distribution. Cover and Thomas (2006, sec. 5.4) state Theorem 0.2 in this form, with $H_2(p) + D_2(p \,\|\, q)$ in place of $H_2(p, q)$.

Remark 0.4 (Why models are fit by minimizing cross-entropy). Fitting a classifier by minimizing cross-entropy chooses the model's distribution q so that a code built for q would spend as few bits as possible on the observed labels. By Remark 0.3, that pushes q toward the distribution the labels came from.

For a single labeled example, let e be the distribution that puts all its mass on the observed label. Then $H_2(e) = 0$ (as for a coin that always lands heads), so the example's cross-entropy loss $H_2(e, q)$ equals $D_2(e \,\|\, q)$: every bit the loss reports is a wasted one.

References

Cover, Thomas M., and Joy A. Thomas. 2006. *Elements of Information Theory*. 2nd ed. Wiley-Interscience. https://doi.org/10.1002/047174882X.

Shannon, C. E. 1948. "A Mathematical Theory of Communication." *Bell System Technical Journal* 27 (3): 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x.