Source coding

Shashank Vatedka

Recap

- Three problems
  - Data compression
  - Reliable communication - error correction
  - Classification
- Concept of information as measure of randomness - links with compression
- Probability recap

Compression: Goals

- Purpose of data compression: save space.
- Convert a *source* file having n bytes to a *compressed* file having $k < n$ bytes.

Compression: Goals

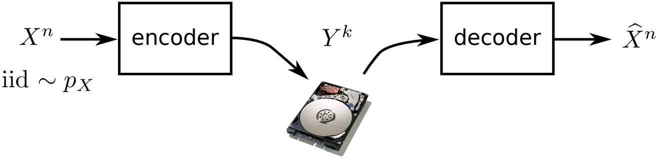
- Purpose of data compression: save space.
- Convert a *source* file having n bytes to a *compressed* file having $k < n$ bytes.
- Model for source file X^n: randomly generated according to *known* source distribution $X^n \sim p_{X^n}$.

Compression: Goals

- Purpose of data compression: save space.
- Convert a *source* file having n bytes to a *compressed* file having $k < n$ bytes.
- Model for source file X^n: randomly generated according to *known* source distribution $X^n \sim p_{X^n}$.
- This course: X^n is iid $\sim p_X$.
- More realistic model: Markov source. But ideas similar for this case as well.

Model

Memoryless source: $X^n \sim \mathrm{iid}(p_X)$.
p_X known



Fixed vs variable length compression

- Fixed-length: k fixed beforehand.
  - Nonzero probability of error
- Variable-length: k different for different x^n.
  - Zero probability of error
  - Additional requirements: usually prefix-free

Encoding and decoding rules

- What is an encoder?
- What is a decoder?
- Computational complexity?

Fixed length compression

- At most 2^k message sequences can be reconstructed with zero error.
- Set of recoverable sequences: $\mathcal{S}$

Fixed length compression

- At most 2^k message sequences can be reconstructed with zero error.
- Set of recoverable sequences: $\mathcal{S}$
- Error occurs whenever $X^n \notin \mathcal{S}$.

$$P_e = \Pr[X^n \notin \mathcal{S}].$$

Reformulation of problem

- Goal: Find optimal $\mathcal{S}$ which minimizes P_e
- Question: What is this optimal $\mathcal{S}$?
- Expression for optimal P_e?

Source coding theorem

Theorem (Shannon, 1948)

For iid source $\sim p_X$*, there exist fixed length source codes for which* $\lim_{n\to\infty} P_e = 0$ *and*

$$\lim_{n\to\infty} \frac{k}{n} = \underbrace{-\sum_{x\in\mathcal{X}} p_X(x) \log_2 p_X(x)}_{H(p_X)}$$

General sources

Theorem

For any stationary and ergodic source, there exist fixed length source codes for which $\lim_{n\to\infty} P_e = 0$ *and*

$$\lim_{n\to\infty} \frac{k}{n} = -\lim_{n\to\infty} \underbrace{\frac{H(X^n)}{n}}_{\text{entropy rate}}$$

Entropy rate

- For iid sources,
$$\frac{H(X^n)}{n} = H(X)$$

Entropy rate

- For iid sources,
$$\frac{H(X^n)}{n} = H(X)$$
- For first-order Markov sources (with stationary initial distribution),
$$\lim_{n\to\infty} \frac{H(X^n)}{n} = -\sum_{x_1, x_2 \in \mathcal{X}} \pi(x_1) p_{X_2|X_1}(x_2|x_1) \log_2 p_{X_2|X_1}(x_2|x_1).$$

Comments

- $H(X)$: abuse of notation
- We actually "redefine" $x \log_2(x)$ such that $x \log_2 x = 0$ for $x = 0$. Indeed, $\lim_{x \to 0} x \log x = 0$.
- $H(X)$ captures the amount of randomness of a source. The data compression problem can be thought of as one of formulating a sequence of yes/no questions to arrive at X^n.

Proof for Bernoulli sources

Use a suboptimal $\mathcal{S}$.