10-801: Advanced Topics in Graphcal Models 10-801, Spring 2007

Information Theory

Lecturer: Eric P. Xing *Scribes: Fan Guo*

1 Data Compression in Information Theory: Basics

To pass some information x through a channel, the sender encodes x and sends the code $C(x)$ through the channel, then the receiver receives $C(x)$ from the channel (assumed to be noiseless) and decodes $C(x)$ to get back x. Coding and decoding can be formulated as a data compression problem.

1.1 Some Definition

Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$,

A *random variable* $X(\omega)$ is a function from Ω to $\mathbb{R}$. The range of $X(\omega)$ is denoted by $\mathfrak{R}_X$.[1] If $\mathfrak{R}$ is countable, X is a *discrete random variable.*

The *probability mass function* $p_X(x)$ is a function from $\mathfrak{R}$ to $[0, 1]$, such that $p(x) = \mathbb{P}(\{\omega | X(\omega) = x\})$.

The *entropy* of a pmf $p(x)$ is given by $H_p = -\sum_x p(x) \log_2 p(x) = \mathbb{E}_p \left[\log_2 \frac{1}{p(x)}\right]$.

The *mutual information* between two random variables X and Y which have a well-defined joint distribution $p(x, y)$ is given by $MI(X, Y) = \sum_{x,y} p(x, y) \log \frac{p(x,y)}{p(x)p(y)}$. The mutual information $MI(X, Y)$ is non-negative; $MI(X, Y) = 0$ if and only if X and Y are independent.

Given an finite alphabet $\mathcal{D}$ whose size $|\mathcal{D}| = D$.

A *source code* $C_X(x)$ is a mapping from to $\mathcal{D}^* = \bigcup_{k=0}^{\infty} \mathcal{D}^k$,

The *length* of the code $C(x)$ is defined to be the non-negative number $l(x)$ such that $C(x) \in \mathcal{D}^{l(x)}$.

The *expected length* $L(C)$ of a source code C with probability mass function $p(x)$ is given by $L(C) = \mathbb{E}[l(x)] = \sum_{x \in \mathfrak{R}} p(x) l(x)$.

1.2 Kraft Inequality

(p.82 in [?]) *For any prefix code over an alphabet of size D, the codeword lengths $l_1, l_2, \ldots, l_m$ must satisfy the inequality*

$$\sum_{i=1}^{m} D^{-l_i} \leq 1. \tag{1}$$

[1] When there is no confusion, we may drop the subscripts and function arguments for clearer notation.

Conversely, given a set of codeword lengths that satisfy this inequality, there exists an instantaneous code with these word lengths.

The proof of the theorem is detailed in [**?**], which also gives a generalized version which allow the set of codewords to be countably infinite:

$$\sum_{i=1}^{\infty} D^{-l_i} \leq 1. \tag{2}$$

Eq. **??** is called the *extended Kraft inequality.*

1.3 Lower Bound on Expected Length of Code

(p.86 in [**?**]) *The expected length of any prefix D-ary code C for a discrete random variable X is greater or equal to the entropy $H_D(X)$, i.e.*

$$L(C) \geq H(X)/\log_2 D \equiv H_D(X), \tag{3}$$

with equality iff for any $x \in \mathfrak{R}$, $D^{-l(x)} = p(x)$.

Proof:

$$\begin{aligned} L(C) - H(X)/\log_2 D &= \sum_x p(x)l(x) + \sum_x p(x)\log_2 p(x)/\log_2 D \\ &= \sum_x p(x)\log_D D^{l(x)} + \sum_x p(x)\log_D p(x) \\ &= -\sum_x p(x)\log_D \frac{D^{-l(x)}}{p(x)} = -\mathbb{E}_p\left[\log_D \frac{D^{-l(x)}}{p(x)}\right] \\ &\geq -\log \mathbb{E}_p\left[\frac{D^{-l(x)}}{p(x)}\right] = -\log\left(\sum_{x\in\mathfrak{R}} D^{-l(x)}\right) \qquad \text{(Jensen's inequality)} \\ &\geq 0. \qquad \text{(extended Kraft inequality)} \end{aligned} \tag{4}$$

Both the inequality signs become equality when $p(x) = D^{-l(x)}$ for any $x \in \mathfrak{R}$.

References

[Cover and Thomas, 1991] T.M. Cover and J.A. Thomas (1991) *Elements of Information Theory*, Wiley Interscience.