COMPSCI 650 Applied Information Theory Feb 9, 2016

Lecture 6

Instructor: Arya Mazumdar *Scribe: Names Redacted*

1 Markov Chains and Review

$X \to Y \to Z$ is called a Markov chain. In a Markov chain, later nodes depend only on previous nodes.

$$p(z|x,y) = p(z|y) \tag{1}$$
$$I(X;Z) \leq I(Y;Z) \text{ (data processing inequality)} \tag{2}$$

We have covered:

- Shannon Code
- Shannon Fano Elias Code
- Huffman Code
- Arithmetic Code

Suppose we don't know $p(x)$. For example, given 000110011001110010101001100101111 0000..., how to compress?

- Huffman code: based on data
- Arithmetic code: need to estimate $p(x)$. Training sequence is not available, however.

2 Unsupervised compression algorithm: Lempel Ziv Algorithm (LZ78/LZW78)

Lempel-Ziv is a universal coding, so we don't need to estimate the distribution. All universal codes are variants of Lempel-Ziv. It is also called a dictionary algorithm.

To do Lempel-Ziv coding, first, produce comma separated file of unique values. That is, put comma after encountering a subsequence you haven't seen before. For the example above:

1. Given 0001100110011100101010011001011110000
2. 0,001100110011100101010011001011110000
3. 0,00,1100110011100101010011001011110000
4. 0,00,1,100110011100101010011001011110000
5. 0,00,1,10,0110011100101010011001011110000
6. 0,00,1,10,01,10011100101010011001011110000
7. 0,00,1,10,01,100,11100101010011001011110000
8. 0,00,1,10,01,100,11,00101010011001011110000
9. 0,00,1,10,01,100,11,001,01010011001011110000
10. 0,00,1,10,01,100,11,001,010,10011001011110000

11. 0,00,1,10,01,100,11,001,010,1001,1001011110000

12. 0,00,1,10,01,100,11,001,010,1001,1001011110000

13. ...

14. 0,00,1,10,01,100,11,1001,010,10011,0010,111,1000,0

Note they are shorter in the beginning, then later we have longer subsequences.

We can then subset any sequence. 10011 could be encoded as the numerical index of 1001, and then the leftover binary digit 1, which would be (8,1). For example:

1. 0: (0,0)
2. 00: (1,0)
3. 1: (0,1)
4. 10: (3,0)
5. 01: (1,1)
6. 100: (4,0)
7. etc.

Finally, we convert the tuples to binary. Let $c(n)$ be the number of subsequences in the file. The number of bits the sequence has been compressed to is $c(n)(\lceil \log(c(n)) \rceil + 1)$. The bits per character is $\frac{c(n)(\lceil \log(c(n)) \rceil + 1)}{n} \rightarrow H(x)$.

2.1 Optimality

Huffman coding is optimal if the probability distribution is known.
Lempel-Ziv is closer to optimal the larger the file is.

3 Asymptotic Equipartition Property

The Asymptotic Equipartition Property is an instance of the weak law of large numbers.

3.1 Weak Law of Large Numbers

Take a random variable X.
Toss the coin several times to create X_1,X_2, etc.
If tossing is unbiased, then $E(X) = \frac{1}{2}$
$\chi = 0, 1$
$p(0) = p(1) = \frac{1}{2}$

With enough independent trials of the same event then:

$$\lim_{n \rightarrow \infty} P(|\frac{1}{n} \sum_{i=1}^{n} X_i - \frac{1}{2}| > \epsilon) = 0, \forall \epsilon > 0 \tag{3}$$

or $\bar{X} \rightarrow \mu$.

3.2 AEP Theorem

Let $X_1, \cdots X_n$ be iid, then:

$$-\frac{1}{n}\log p(x_1, ..., x_n) \to H(X) \tag{4}$$

As n approaches infinity, the probability that equation 4 is anything other than H(X) approaches zero.

Proof:

$$p(x_1, x_2...x_n) = \prod_{i=1}^{n} p(x_i) \tag{5}$$

since iid, so:

$$-\frac{1}{n}\log p(x_1, ..., x_n) = -\frac{1}{n}\log \prod_{i=1}^{n} p(x_i) \tag{6}$$

$$= -\frac{1}{n}\sum_{i=1}^{n} \log(p(x_i)) \to -E(\log(p(X_i))) \tag{7}$$

$$\approx -\sum_{x\in\chi} p(x)\log(p(x)) \tag{8}$$

$$= H(X) \tag{9}$$

$$\tag{10}$$

3.3 Typical Set

Consider the coin toss example where we have a biased coin with probability of observing tail $p(T) = 0.25$. Highly likely sequences will have 25% tails, while observing 25% heads is unlikely. A typical set is a set containing highly likely sequences. $A_{n,\epsilon}$ is a typical set of $X_1, \cdots, X_n$ iff:

$$A_{n,\epsilon} = \{(x_1....x_n) : 2^{-n(H(X)+\epsilon)} \leq p(x_1, ..., x_n) \leq 2^{-n(H(X)-\epsilon)}\} \tag{11}$$

Equivalently, if $(x_1...x_n) \in A_{n,\epsilon}$, then $H(X) - \epsilon \leq -\frac{1}{n}\log p(x_1, ..., x_n) \leq H(X) + \epsilon$. So:

$$P(A_{n,\epsilon}) \to 1 \text{ as } n \to \infty \tag{12}$$

$$2^{n(H(X)-\epsilon)} \leq |A_{n,\epsilon}| \leq 2^{n(H(X))+\epsilon)}, \epsilon \to 0 \tag{13}$$

$$|\chi^n| = 2^{n\log|\chi|} \tag{14}$$

Note:

- All sequences in a typical set have approximately equal probabilities.
- With long enough n, you're more likely to end up in the typical set.

We can assign a unique sequence for each element in the typical set, so we have $\log 2^{nH(X)} = nH(X)$ per n-length sequence $\to H(X)$ per character. Reject all other points.

3.4 Example: Stirling's approxmation for $\binom{n}{k}$

Stirling's approximation for $n!$ is:

$$n! \approx \sqrt{2\pi n}(\frac{n}{e})^n \tag{15}$$

Approximation for $\binom{n}{k}$:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \tag{16}$$

$$= \frac{\sqrt{2\pi n}(\frac{n}{e})^n}{\sqrt{2\pi k}(\frac{k}{e})^k \sqrt{2\pi(n-k)}(\frac{n-k}{e})^{n-k}} \tag{17}$$

$$= \sqrt{\frac{n}{2\pi k(n-k)}} \cdot \frac{n^n}{k^k(n-k)^{n-k}} \tag{18}$$

$$= \sqrt{\frac{n}{2\pi k(n-k)}} \cdot 2^{n\log n - k\log k - (n-k)\log(n-k)} \tag{19}$$

$$= \sqrt{\frac{n}{2\pi k(n-k)}} \cdot 2^{-n(\frac{k}{n}\log\frac{k}{n} - \frac{n-k}{n}\log\frac{n-k}{n})} \tag{20}$$

$$= \sqrt{\frac{n}{2\pi k(n-k)}} \cdot 2^{nH_2(\frac{k}{n})} \tag{21}$$

where $H_2(p) = -p\log p - (1-p)\log(1-p)$.

3.5 The Chernoff bound

Suppose $p(1) = p, p(0) = 1-p$, then:

$$P(\text{a sequence with qn 1's}) = p^{qn}(1-p)^{n-qn} \tag{22}$$

When $q = p$:

$$P(\text{a sequence with qn 1's}) = p^{pn}(1-p)^{n(1-p)} = 2^{n[p\log p + (1-p)\log(1-p)]} = 2^{-nH_2(p)} \tag{23}$$

The probability of obtaining any sequence with qn 1's is:

$$P(\text{obtaining a sequence with qn 1's}) = \binom{n}{qn} p^{qn}(1-p)^{n-qn} \tag{24}$$

$$= \frac{1}{c\sqrt{n}} 2^{nH_2(q)} \cdot 2^{nq\log p} \cdot 2^{n(1-q)\log(1-p)} \tag{25}$$

$$\approx \frac{1}{\sqrt{n}} 2^{n[-q\log q - (1-q)\log(1-q) + q\log p + (1-q)\log(1-p)]} \tag{26}$$

$$= -\frac{1}{\sqrt{n}} 2^{-n[q\log\frac{q}{p} + (1-q)\log\frac{1-q}{1-p}]} \tag{27}$$

$$= \frac{1}{\sqrt{n}} 2^{-nD(q||p)} \tag{28}$$

$$\tag{29}$$

where $D(q||p) \geq 0$.