M-Data Compression

Erick Oduniyi (eeoduniyi@gmail.com)

M-Data Compression

Erick Oduniyi (erickoduniyi@ku.edu)
Department of Electrical Engineering & Computer Science, University of Kansas
1450 Jayhawk Blvd, Lawrence, KS 66045, USA

1. Discussion

The problem of data compression is to choose code-word lengths that minimize the average number of bits per symbol, in addition to the codes are instantaneous. The four algorithms accomplish this, however they each prescribe their respective sets of pros and cons. In general, the *Huffman* coding algorithm provides a simple and fast coding scheme, where the algorithm complexity grows exponentially with the block length, but requires a probability distribution. Further, it is difficult to use Huffman coding for extended sources with large alphabets or when the symbol probabilities vary with time. *Shannon-Fano-Elias* coding is a type of arithmetic coding where complexity grows linearly with the block size, but also requires source of symbol statistics. Finally, the *LZ78* scheme utilizes a dictionary to store patterns and transmit the buffer offset and length. These universal coders do not need to know the symbol probability distribution and is instead estimated in a single pass of the data.

2. Continuing the Discussion

Dynamic Markov Compression (DMC) is a lossless data compression algorithm developed by Gordon Cormack and Nigel Horspool. DMC predicts and codes one bit at a time. The predicted bit is then coded using arithmetic coding. To help introduce notions of capacity, compression, and Bayesian dynamics, we review concepts from ***Probabilistic Machine Learning.***

2.0.0.1. Definition Let λ be a discrete-time Markov chain composed of a collection of random variables $\boldsymbol{X} = \{X_t, t \in \mathbf{N_0}\}$ on a probability space (Ω, P). Now, assume the states S of the Markov chain are unobservable at time t (s_t is hidden). However, the Markov chain produces observable symbols π with a certain probability, which is an additional stochastic process.

2.0.0.2. Definition A subset of a sample space is called an ***event***.

2.0.0.3. Example Suppose person Z is interested in calculating the probability of flipping two coins and exactly one of them landing on heads (H). In order to do this, Z first needs to define the sample space $S = \{HH, HT, TH, TT\}$ and the specific event E where exactly one of the coins is heads $= \{HT, TH\}$. Now Z can calculate the probability of event E occuring by dividing the absolute number of elements in E by the absolute number of elements in S:

$$P(E) = \frac{|E|}{|S|} = \frac{2}{4}$$

2.0.0.4. Definition A ***random variable*** is a measurable function that takes in values within some arbitrary set S: $x = \{\Omega \in S\}$. Random variables can be *discrete* and *continuous*, however, this module restricts our discussion to ***real-valued discrete random variables***, where their values are in some infinite or finite countable subset of $\mathbb{R}$.

2.0.0.5. Example Person Z is interested in using a random variable x to describe the humidity of Lawrence, Kansas on a day in November. First Z finds out that the humidity set H contains values anywhere between 48% - 92% So, he tells his cat that the event where the humidity on November 21st is less than 90% is given by:

$$Humidity = \{x < 90\} = \{\Omega \in H : x(\Omega) < 90\}$$

2.0.0.6. Definition Within a data set k, the value computed by summing all data points (values) and then dividing by the total number of data points N is called the ***mean*** or *arithmetic average*. Practically, the mean describes the concetrantion or location of data. ***Note:*** *The* ***expected value*** *is often used interchangeably with mean when random variables are under consideration. Simply, It's the mean of a random variable x.*

$$mean(k) = \frac{\sum_{i=1}^{N} k_n}{N} \tag{1}$$

2.0.0.7. Definition The ***sample variance*** of a data set k is the "estimation" of the mean *squared deviation*.

$$sample\ variance(k) = \frac{\sum_{i=1}^{n} (k_n - mean(k))^2}{n-1} \tag{2}$$

Here n denotes the *sample size* and $n-1$ is used to simulate an *unbiased estimator* of the standard deviation. ***Note:*** *often N is used to denote "population" size; total number of members within a defined group, while n describes the "sample" size, a portion of the population.*

2.0.0.8. Definition 6. The ***standard deviation*** s of a data set k is the measure of how far away the data points are from the mean. More precisely, it measures the spread of a distribution and is the square root of sample variance.

$$s(k) = \sqrt{sample\ variance(k)} \tag{3}$$

2.0.0.9. Definition Two or more events are statistically ***independent*** if information about one event cannot affect the other event.

2.0.0.10. Definition The ***conditional probability*** is the probability of event A occuring given event B has happened:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{4}$$

Where $P(A \cap B)$ denotes that both events A and B occur.

2.0.0.11. Definition If S is a finite or countably infinite sample space and P denotes the probability of some event happening within S, then the pair (S, P) is called the ***probability space***.

2.0.0.12. Definition The ***probability mass function*** (PMF) of a data set k is a function that gives the probability of a discrete random variable x being equal to some value X within k.

$$f(x) = P(x = X) \tag{5}$$

PMF has some associated properties. f(x) is always greater than zero: $f(x) > 0$ and the sum of the PMF must equal to 1: $\sum f(x) = 1$

2.0.0.13. Example Suppose person Z has some urn that contains 7 red balls and 5 green balls. Z now chooses three balls at random from the urn. Let x denote the number of red balls Z chooses. The PMF for Z choosing exactly 1 red ball is given by:

$$f(x) = P(x = 1) = \frac{(7C_1)(5C_2)}{12C_3} = \frac{7}{22}$$

Here the combinatorics notation nC_k is used to describe: $\frac{n!}{k!(n-k)!}$

2.0.0.14. Definition The ***cumulative distribution function*** (CDF) of a data set k gives the probability that a random variable x is less than or equal to some value X within k:

$$F(x) = P(x \leq X) \tag{6}$$

For a discrete distribution the CDF is calculated by:

$$F(x) = \sum_{n=-\infty}^{x} f(x) \tag{7}$$

2.1. Stochastic Processes

The application of stochastic processes can be found in an array of fields like computer systems, control systems, wireless communication, operations research, and finance. The section of the module will outline the formal definition of stochastic process, then attempt to further narrow its focus on Markov chains and a few examples that are utilized within the above mentioned fields.

2.1.0.1. Definition A ***stochastic process*** is a collection of random variables defined on some probability space T that is composed of a sample space S, a probability measure P, some σ-field (set of all subsets) F, and random variables **X** = $\{X_t, t \in S\}$ on the same probability space:

$$T = (S, P, F) \tag{8}$$

There are two general classes of stochastic process: If there are only finite or infinite countably components of the process, then they are called *discrete-time process*, otherwise they are refereed to as *continuous-time process* often denoted by $\{X_t : t \geq 0\}$. Again, this module will restrict the conversation to discrete-time processes.

3. Markov Chains

Now we consider a subclass of stochastic processes called ***discrete-time Markov chains*** (DTMCs). In order to construct DTMCs is necessary to outline three simplifying assumptions. The first assumption or restriction is that the process is a discrete-time process. Secondly, the processes must have discrete (countable) state space, and finally, DTMCs must be a process that satisfies the *Markov property*.

3.0.0.1. Definition A stochastic process $T = (S, P, F)$ on a state space $S = \{1, 2, 3...\}$ where for all $t \geq 0, X_t \in S$ and for all t and all states $i_0, i_1, ..., i_{t-1}, i_t$ fulfills the Markov property:

$$P(X_{t+1} = s_{t+1} | X_t = s_t, ..., X_n = s_n) \tag{9}$$
$$= P(X_{t+1} = s_{t+1} | X_t = s_t, X_{t-1} = s_{t-1}, ..., X_0 = s_0) \tag{10}$$

is called a ***discrete-time Markov chain***.

3.1. Markov Chain Representations

Markov chains are typically represented in two forms. The first representation are *state diagrams* that outline the Markov chain's states as nodes and links as the probabilities from transitioning from one state to another. State diagrams are very appealing when a concrete visualization of systems composed of many interacting units is needed. The second main representation are *stochastic matrices* (sometimes referred to as a transition matrix) which simply describe the Markov chains *transition probabilities* and are often more computationally useful. Here, the module will go into more detail about the respective representations of Markov chains.

3.1.1. State Diagrams

Markov chains are often represented by ***state diagrams***, this provides a clear graphical representation of the systems states and transition probabilities.

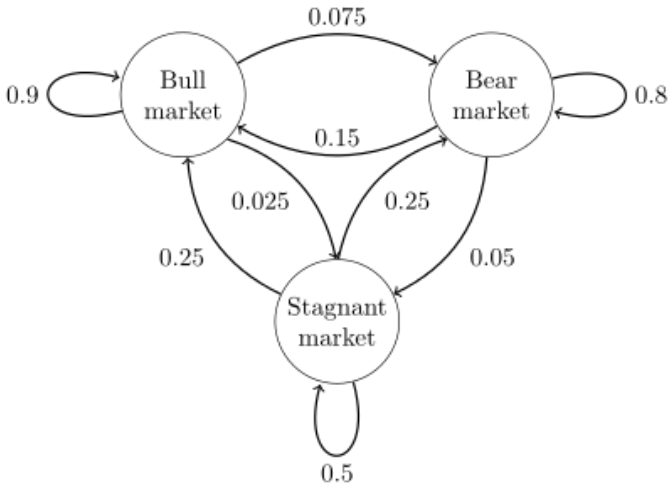

Fig. 1 – Graphical representation of a DTMC with the Markov chains modeling financial markets

3.1.2. Stochastic Matrix

3.1.2.1. Formal Definition A *n x n* matrix *R* that has elements which are all non-negative and where each row sums to one is called a ***stochastic matrix***.

3.1.3. Transition Probabilities

3.1.3.1. Definition The elements of the stochastic matrix *R* are called the ***transition probabilities***.

$$R = \begin{bmatrix} 0.9 & 0.025 & 0.075 \\ 0.25 & 0.5 & 0.25 \\ 0.15 & 0.05 & 0.8 \end{bmatrix}$$

Here the stochastic matrix is given to illustrate Figure 1, where element $R(1,1)$ describes the probability of the bull market staying in the bull market.