Universal Bayesian Measures

Joe Suzuki
Department of Mathematics, Osaka University, Toyonaka, Osaka, 560-0043, Japan.
Email: suzuki@math.sci.osaka-u.ac.jp

***Abstract*—In the minimum description length (MDL) and Bayesian criteria, we construct description length of data $z^n = z_1 \cdots z_n$ of length n such that the length divided by n almost converges to its entropy rate as $n \to \infty$, assuming z_i is in a finite set A. In model selection, if we knew the true conditional probability $P(z^n|F)$ of $z^n \in A^n$ given each F, we would choose F such that the posterior probability $P(F|z^n)$ of F given z^n is maximized. But, in many situations, we use $Q : A^n \to [0,1]$ such that $\sum_{z^n \in A^n} Q(z^n|F) \le 1$ rather than P because only data z^n are available. In this paper, we consider an extension such that each of the attributes in data can be either discrete or continuous. The main issue is what Q is qualified to be an alternative to P in the generalized situations. We propose the condition in terms of the Radon-Nikodym derivative of P with respect to Q, and give the procedure of constructing Q in the general setting. As a result, we obtain the MDL/Bayesian criteria in a general sense.**

I. Introduction

Consider feature selection: given a set of random variables $X^{(1)}, \cdots, X^{(m)}, Y$ and n tuples of data $z^n = \{(x_i^{(1)}, \cdots, x_i^{(m)}, y_i)\}_{i=1}^n$ emitted by them, we wish to find a minimal subset $F \subset \{1, \cdots, m\}$ such that $P(Y|X^{(1)}, \cdots, X^{(m)}) = P(Y|(X^{(j)})_{j \in F})$. Let $P(z^n|F)$ be the conditional probability of z^n given F. If the prior probability $\pi(F)$ is available for each $F \in \{1, \cdots, m\}$, our goal is to find the F such that $\pi(F)P(z^n|F)$ is maximized.

However, in many practical situations, only data z^n are available. So, we need to consider an alternative Q such that $\sum_{z^n} Q(z^n|F) \le 1$ to the true probability P. If we can prepare such a Q, we only choose F such that $-\log \pi(F) - \log Q(z^n|F)$ is minimized, or equivalently, one such that $\pi(F)Q(z^n|F)$ is maximized. We refer those evaluations to the minimum description length (MDL) [8] and Bayesian criteria, respectively.

Fortunately, if each z_i in $z^n = (z_i)_{i=1}^n$ takes a value in a finite set A, then such $Q : A^n \to [0,1]$ with $\sum_{z^n} Q(z^n) \le 1$ have been developed for compression without assuming the knowledge of the true probability such as Lempel-Ziv codes, adaptive arithmetic codes [5] which are currently used extensively in internet communications. In fact, any data z^n can be compressed into at most $-\log Q(z^n)$ plus one bits, where the logarithm base is two, and it is known that the quantity $-\frac{1}{n}\log Q(z^n)$ almost surely converges to the entropy that is the lower limit of the compression ratio when we encode z^n using the knowledge of P.

The idea of replacing the true P by such a Q to obtain a Bayesian solution has been used thus far. Wray Buntine considered its application to construction of classification trees [2], Cooper and Herskovits estimated Bayesian network structures using such a Q [4], and modification to the MDL principle and its application to a Bayesian version of the Chow-Liu algorithm were considered in [12]. Since then, many authors reported applications using similar techniques thus far. However, if some attributes take continuous values, it is hard to construct such a Q in order to identify the F given z^n.

In this paper, we propose how to choose an optimal F given z^n in the sense of MDL/Bayesian criteria, without assuming that the data are either discrete or continuous. The main issue is what Q is qualified to be an alternative to P in more general settings. We will give an answer to the problem in terms of the Radon-Nikodym derivative [3] of P with respect to Q at the data z^n.

The theory developed in this paper is partially due to Boris Ryabko's density estimation [9]. The idea is to prepare many histograms nested each other: one histogram is a refinement of another; for each histogram, the quantity like $Q(z^n)$ is computed and divided by the cell volume (Lebesgue measure) of dimension n; and the final estimation is given by weighting those estimated density functions. Ryabko proved the estimation is consistent [9].

Our contribution to theory is to remove the constraint that the density function should exist for universal coding. Obviously, just because a random variable is not finite does not mean that its density function exists. For example, if the distribution function is given by

$$F_X(x) = \begin{cases} 0 & x < -1, \\ \frac{1}{2}, & -1 \le x < 0 \\ \frac{1}{2} + \frac{1}{2}\int_0^x h(t)dt, & 0 \le x \end{cases}$$

with $\displaystyle\int_0^\infty h(x)dx = 1$, then there exists no density function f_X such that $F_X(x) = \displaystyle\int_{-\infty}^x f_X(t)dt$.

As a result, the idea can be applied to many situations in estimation. In fact, this paper proposes the general notion of the MDL/Bayesian criteria even if coding is not possible. Besides the feature selection problem introduced above, we illustrate how to estimate the order of Markov ergodic sources given data sequence even if the random variables are continuous.

Section 2 introduces the notion of universal coding for finite sources. Section 3 gives two illustrative examples by which we specify the scope of the MDL/Bayesian criteria in this paper. Section 4 extends the notion of universal coding for general sources. Section 5 generalizes the MDL/Bayesian criteria using the discussion in Sections 3 and 4. Section 6 summarizes the obtained results and state future works.

II. Universal Coding

A. Coding

Let A be a finite set, and $n \in \mathbb{N} := \{1, 2, \cdots\}$. We denote the set of binary sequences of finite length by $\{0,1\}^*$, and write $|y| = m$ if $y \in \{0,1\}^m$ for $y \in \{0,1\}^*$. We define coding c and its length l_c by any mapping $A^n \to \{0,1\}^*$ and $x^n \in A^n \mapsto |c(x^n)| \in \mathbb{N}$, respectively. We say that coding $c: A^n \to \{0,1\}^*$ is *uniquely decodable* if the map $(n, x^n) \in \mathbb{N} \times A^n \mapsto c(x^n) \in \{0,1\}^n$ is one-to-one.

We say that map $l: A^n \to \mathbb{N}$ satisfies *Kraft's inequality* if $\sum_{x^n \in A^n} 2^{-l(x^n)} \leq 1$. It is known [5] that if $c: A^n \to \{0,1\}^*$ is uniquely decodable, then l_c satisfies Kraft's inequality, and that if $l: A^n \to \mathbb{N}$ satisfies Kraft's inequality, then there exists a uniquely decodable $c: A^n \to \{0,1\}^*$ such that $l = l_c$.

Let $Q^n(x^n) := 2^{-l(x^n)}$. Then, the problem of coding reduces to specifying $Q^n: A^n \to [0,1]$ such that $\sum_{x^n \in A^n} Q^n(x^n) \leq 1$.

B. Sources

We say a sequence $\{X_i\}_{-\infty}^{\infty}$ (*source*) of random variables X_i to be *finite* if the range $A := X_i(\Omega)$ of X_i is finite, and that it is *stationary* if for each $k \in \mathbb{N}$ and $i \in \mathbb{Z} := \{\cdots, -1, 0, 1, \cdots\}$, $P(X_{i-k} \cdots X_{i-1}) = P(X_{i-k+1} \cdots X_i)$. For example, for $X^n := \{X_i\}_{i=1}^n$, if X^n is i.i.d (independently and identically distributed), then $P(X^n) = \prod_{i=1}^n P(X_i)$ and if X^n depends on a finite number of latest random variables, then $P(X^n|X_{-\infty}, \cdots, X_0) = \prod_{i=1}^n P(X_i|X_{i-k} \cdots X_{i-1})$, where we say the source is *Markov* of order k if k is the minimal value satisfying the above equation.

On the other hand, if $\{X_i\}_{-\infty}^{\infty}$ is stationary, there exists the limit (*entropy*) $H := \lim_{n\to\infty} -\frac{1}{n} \sum_{x^n \in A^n} P^n(x^n) \log P^n(x^n)$.

C. Universal Coding for i.i.d. Sources

If the probability P^n is known, then $H \leq E[-l_c(X^n)]$ for any uniquely decodable c, and $l_c(x^n) := \lceil -\log P^n(x^n) \rceil$ (round-up) satisfies $E[l_c(X^n)] < H + 1$. However, if P^n is not known, those facts are not available. Instead, we can construct $Q^n: A^n \to [0,1]$ such that $\sum_{x^n \in A^n} 2^{-l(x^n)} \leq 1$, and

$$-(1/n) \log Q^n(x^n) \to H \tag{1}$$

for any stationary ergodic P^n with probability one as $n \to \infty$ (*universal coding*).

In fact, suppose P^n is i.i.d. with unknown stochastic parameters. Let $m := |A|$ and denote by $c_n[x]$ the number of occurrences of $x \in A$ in $x^n \in A^n$, where hereafter we write the cardinality of set S by $|S|$. Then, we can check that [6]

$$Q^n(x^n) := \prod_{i=1}^n \frac{c_{i-1}[x_i] + \frac{1}{2}}{i - 1 + \frac{m}{2}} = \frac{\Gamma(\frac{m}{2}) \prod_{x\in A} \Gamma(c_n[x] + \frac{1}{2})}{\Gamma(n + \frac{m}{2}) \Gamma(\frac{1}{2})^m} \tag{2}$$

satisfies (1), where Γ is the Gamma function.

On the other hand, from the Shannon-McMillan-Breiman theorem [5], we have $-\frac{1}{n} \log P^n(x^n) \to H$ for any stationary ergodic P^n with probability one as $n \to \infty$, which together with (1) means

$$\frac{1}{n} \log \frac{P(x^n)}{Q(x^n)} \to 0 \ . \tag{3}$$

III. MDL/Bayesian Criteria

A. Markov Order Identification

Suppose that the source P^n is ergodic Markov with known order k and unknown parameters. Let $S := A^k$ be the set of states of a Markov source, and $c_n[x, s]$ the number of occurrences of $(x_{i-k} \cdots x_{i-1}, x_i) = (s, x) \in S \times A$ in $x^n \in A^n$. Then, (2) is generalized as

$$Q^n(x^n|k) := \prod_{i=1}^n \frac{c_{i-1}[x_i, s_i] + \frac{1}{2}}{\{\sum_{x\in A} c_{i-1}[x, s_i]\} + \frac{m}{2}} \ . \tag{4}$$

Similarly, we can show $\frac{L(x^n)}{n} = -\frac{1}{n} \log Q^n(x^n|k) \to H$ with probability one as $n \to \infty$.

Next, suppose that the Markov order k is unknown and only actually emitted examples $x^n \in A^n$ are available. The problem is to identify the Markov order k given $x^n \in A^n$. To this end, we count $c_n[x, s]$ of occurrences of each $x \in A$ for each $s \in S$ in $x^n \in A^n$ and obtain the value of (4) for each $k = 0, 1, \cdots$.

Let $\pi(k)$ be the a prior probability of each Markov order $k = 0, 1, \cdots$. If we choose k such that $-\log \pi(k) - \log Q^n(x^n|k)$ is minimized, equivalently, such that $\pi(k) Q^n(x^n|k)$ is maximized, then we obtain a solution based on the MDL/Bayesian criteria.

B. Conditional Probabilities

Let X, Y be random variables, and $X(\Omega), Y(\Omega)$ their ranges[1] with $|Y(\Omega)| < \infty$. We specify the conditional probability $P(Y|X)$ by defining equivalent classes of $X(\Omega)$ as follows: $x \sim x' \iff P(y|x) = P(y|x'), y \in Y(\Omega)$. Let $[x]$ be the class including $x \in X(\Omega)$, and $S := \{[x] | x \in X(\Omega)\}$ the set of equivalent classes (*equivalence relation*), where we assume $|S| < \infty$.

Example 1 (feature selection): Suppose that the random variable X is expressed by a vector $X = (X^{(1)}, \cdots, X^{(m)})$ consisting of m random variables $X^{(1)}, \cdots, X^{(m)}$, and that $s \in S$ is unique given the values of random variables $\{X^{(j)}\}_{j\in F}$. Then, we say the minimal subset F satisfying this property to be a *feature set*. On the other hand, if we select F (*feature selection*), then the equivalent relation S can be decided by $S := \prod_{j\in F} X^{(j)}(\Omega)$.

The problem is to identify the equivalent relation S given n pairs of examples $\{(x_i, y_i)\}_{i=1}^n \in X(\Omega) \times Y(\Omega)$. To this end,

[1] Thoughout the paper, by $X(\Omega)$ we mean the range of random variable $X: \Omega \to \mathbb{R}$, where Ω is the sample space.

we count the number $c_n[y,s]$ of occurrences of each $y \in Y(\Omega)$ for each $s \in S$ in $\{(x_i,y_i)\}_{i=1}^n$, and obtain the value of

$$Q^n(y^n|x^n,S) := \prod_{s\in S} \frac{\Gamma(\frac{m}{2})\prod_{y\in Y(\Omega)}\Gamma(c_n[y,s]+\frac{1}{2})}{\Gamma(n+\frac{m}{2})\Gamma(\frac{1}{2})^m}$$

for each S. Let $\pi(S)$ be the prior probability of S. If we choose S such that $-\log\pi(S) - \log Q^n(y^n|x^n,S)$ is minimized, or equivalently, such that $\pi(S)Q^n(y^n|x^n,S)$ is maximized, then we obtain a solution based on the MDL/Bayesian criteria.

IV. Universal Coding in a General Sense

In this section, we construct a universal measure rather than a universal coding in order to extend the notion to general sources because coding is available only for finite sources.

A. Estimation of Density Functions

Let $\{X_i\}_{i=1}^n$ be stationary ergodic with density function f^n, which means that the source $\{X_i\}_{i=1}^n$ is not finite. Let $\{A_k\}_{k=0}^\infty$ be such that $A_0 := \{X_i(\Omega)\}$, and A_{k+1} is a refinement of A_k.

Example 2: Suppose $X_i(\Omega) = [0,1)$, and that we consider the following sequence:

$A_0 = \{[0,1)\}$

$A_1 = \{[0,1/2),[1/2,1)\}$

$A_2 = \{[0,1/4),[1/4,1/2),[1/2,3/4),[3/4,1)\}$

$\ldots$

$A_k = \{[0,2^{-k}),[2^{-k},2\cdot 2^{-k}),\cdots,[(2^k-1)2^{-k},1)\}$

$\ldots$.

Then $m = 2^k$, and if the source is i.i.d., we have $Q_k^n(s_k(x^n)) := \dfrac{\Gamma(\frac{m}{2})\prod_{a\in A_k}\Gamma(c_n[a]+\frac{1}{2})}{\Gamma(n+\frac{m}{2})\Gamma(\frac{1}{2})^m}$, where $c_n[a]$ is the number of occurrences of $a \in A_k$ in $s_k(x^n) \in A_k^n$. The Lebesgue measure is $\lambda(s_k(x)) = 2^{-k}$ for $x \in X_i(\Omega)$.

Let $s_k : X_i(\Omega) \to A_k$ be the projection, i.e. $s_k(x) = a$ if $x \in a \in A_k$. Similarly, we write $s_k(x^n) = a^n$ if $x^n \in a^n \in A_k^n$. Let $\lambda : \mathcal{B} \to \mathbb{R}$ be the Lebesgue measure, i.e. $\lambda(a) = d-c$ if $a = [c,d]$ with $c \le d$. Similarly, we write $\lambda^n(a^n) = \prod_{i=1}^n \lambda(a_i)$ for $a^n \in \mathcal{B}^n$, where $\mathcal{B}$ is the Borel set field of the entire real $\mathbb{R}$.

For each $k = 1,2,\cdots$, let P_k^n be the probability measure of $s_k(X^n)$ with alphabet A_k. Then, there exists $Q_k^n : A_k^n \to [0,1]$ such that $\sum_{a^n\in A_k^n} Q_k^n(a^n) \le 1$ and $\dfrac{1}{n}\log\dfrac{P_k^n(s_k(x^n))}{Q_k^n(s_k(x^n))} \to 0$ for any stationary ergodic f^n with probabilities as $n \to \infty$ (see (3)).

Let $g_k^n(x^n) := \dfrac{Q_k^n(s_k(x^n))}{\lambda^n(s_k(x^n))}$ and $\{\omega_k\}_{k=1}^\infty$ be such that $\sum\omega_k = 1$, $\omega_k > 0$. Suppose we estimate f^n by g^n as $g^n(x^n) := \sum_{k=1}^\infty \omega_k g_k^n(x^n)$.

Let $f_k^n(x^n) := \dfrac{P_k^n(s_k(x^n))}{\lambda^n(s_k(x^n))}$, and define the *differential entropy* by

$$h(f) := \lim_{n\to\infty}\int -f(x^n)\log f(x_n|x_1,\cdots,x_{n-1})dx^n \ . \quad (5)$$

Proposition 1 (Ryabko, 2009 [9]): With probability one as $n \to \infty$,

$$\frac{1}{n}\log\frac{f^n(x^n)}{g^n(x^n)} \to 0 \quad (6)$$

for any stationary ergodic f^n such that

$$h(f_k) \to h(f) \quad (7)$$

as $k \to \infty$.

Notice that Proposition 1 assumes the existence of a density function.

B. Conditional Probabilities

We say that a measure μ is *absolutely continuous* with respect to another measure ν and write $\mu \ll \nu$ if $\nu(A) = 0 \Longrightarrow \mu(A) = 0$ for each $A \in \mathcal{F}$, and that a measure ν is *σ-finite* if there exists $\{A_i\}$ such that $\cup_i A_i = \Omega$ and $\nu(A_i) < \infty$.

Let μ,ν be σ-finite.

Proposition 2 (Radon-Nikodym [3]): Then, $\mu \ll \nu$ if and only if there exists $\mathcal{F}$-measurable $g : \Omega \to \mathbb{R}$ such that $\mu(A) = \int_A g(\omega)d\nu(\omega), A \in \mathcal{F}$.

We write such a g by $\dfrac{d\mu}{d\nu}$.

Let X,Y be random variables. Then, $\mu(X \in D) = 0 \Longrightarrow \mu(X \in D, Y \in D') = 0$ for $D,D' \in \mathcal{B}$. From Proposition 2, for each $D' \in \mathcal{B}$, there exists $\mu(Y \in D'|\cdot) : \mathbb{R} \to \mathbb{R}$ such that $\mu(X \in D, Y \in D') = \int_D \mu(Y \in D'|x)d\mu(x)$ (the *conditional probability* of Y given X).

C. Kullback-Leibler Divergence

When $\mu \ll \nu$, we define the *Kullback-Leibler divergence* of μ with respect to ν by $D(\mu||\nu) := \int d\mu\log\dfrac{d\mu}{d\nu}$. Let $\mu(X \le x) := F_X(x)$. Then, $\mu \ll \lambda$ if and only if there exists $f_X = \dfrac{d\mu}{d\lambda} : \mathbb{R} \to \mathbb{R}$ such that $\mu(X \le x) = \int_{-\infty}^x f_X(t)d\lambda(t)$. In particular,

$$D(\mu||\lambda) = \int d\mu\log\frac{d\mu}{d\lambda} = \int_{-\infty}^{\infty} f_X(x)\log f_X(x)dx \ .$$

Let $\{X_i\}_{i=1}^n \sim \mu^n$ be stationary ergodic. Then, $\mu^{n-1}(X^{n-1} \in D^{n-1}) = 0 \Longrightarrow \mu^n(X^n \in D^n) = 0$ for $D^n \in \mathcal{B}^n$, so that there exists $\mu(X_n \in D_n|\cdot) : \mathbb{R}^{n-1} \to \mathbb{R}$ such that

$$\mu^n(X^n \in D^n) = \int_{D^{n-1}} \mu(X_n \in D_n|x^{n-1}) d\mu^{n-1}(x^{n-1})$$
(*conditional Probability* $\mu(X_n \in D_n|x^{n-1})$ of $X \in D_n$ given $x^{n-1} \in \mathbb{R}^{n-1}$).

Similarly, there exists $\nu(X_n \in D_n|\cdot) : \mathbb{R}^{n-1} \to \mathbb{R}$ as well for ν^n such that $\int_{x^n \in X^n(\Omega)} d\nu^n(x^n) \leq 1$. When $\mu^n \ll \nu^n$, write the Radon-Nikodym derivative at $x_n \in D_n$ by $\frac{d\mu}{d\nu}(x_n|x^{n-1})$, and define

$$D(\mu||\nu) := \lim_{n\to\infty} \int d\mu^n(x^n) \log \frac{d\mu}{d\nu}(x_n|x^{n-1})$$

In particular, from (5), when $\mu^n \ll \lambda^n$,

$$D(\mu||\lambda) = -h(f) \ . \qquad (8)$$

D. Universal Coding for General Sources

In this paper, we extend the constructions of Q^n and g^n achieving (3) and (6), respectively, to that of the general measure without assuming to be either discrete or continuous [13].

Let $\{X_i\}_{i=1}^n$ be stationary ergodic with probability measure μ^n, and η^n be such that $\mu^n \ll \eta^n$. Let $\frac{d\nu_k^n}{d\eta^n}(x^n) := \frac{Q_k^n(s_k(x^n))}{\eta^n(s_k(x^n))}$ and $\{\omega_i\}_{k=0}^\infty$ such that $\sum_{k=0}^\infty \omega_k = 1$, $\omega_k > 0$.

We estimate $\frac{d\mu^n}{d\eta^n}(x^n)$ by $\frac{d\nu^n}{d\eta^n}(x^n) := \sum_{k=0}^\infty \omega_k \frac{d\nu_k^n}{d\eta^n}(x^n)$. Let $\frac{d\mu_k^n}{d\eta^n}(x^n) := \frac{\mu_k^n(s_k(x^n))}{\eta^n(s_k(x^n))}$.

Theorem 1: Let $\{A_k\}_{k=1}^\infty$ such that $A_0 := \{X_i(\Omega)\}$ and A_{k+1} is a refinement of A_k, and η a σ-finite measure. There exists ν^n such that $\int_{x^n \in X^n(\Omega)} d\nu^n(x^n) \leq 1$ and with probability one

$$\frac{1}{n} \log \frac{d\mu^n}{d\nu^n}(x^n) \to 0 \qquad (9)$$

for any stationary ergodic μ^n such that $\mu^n \ll \eta^n$ and

$$D(\mu_k||\eta) \to D(\mu||\eta) \qquad (10)$$

as $k \to \infty$.
(see proof of Theorem 1 for Appendix).

Remark 1: When $\eta = \lambda$, (9) and (10) become (6) and (7) (see (8)).

Example 3: Suppose $X_i(\Omega) = \{1, 2, \cdots\}$, and that the following sequence is given:

$A_0 = \{\{1, 2, \cdots\}\}$
$A_1 = \{\{1\}, \{2, \cdots\}\}$
$A_2 = \{\{1\}, \{2\}, \{3, \cdots\}\}$
$\ldots$
$A_k = \{\{1\}, \cdots, \{k\}, \{k+1, \cdots\}\}$
$\ldots$

Then $m = k + 1$, and if the source is i.i.d., $\nu_k^n(s_k(x^n)) := \frac{\Gamma(n+\frac{m}{2})\Gamma(\frac{1}{2})^m}{\Gamma(\frac{m}{2}) \prod_{a\in A_k} \Gamma(c_n[a]+\frac{1}{2})}$, where $c_n[a]$ is the number of occurrences of $a \in A_k$ in $s_k(x^n) \in A_k^n$. If we set η as follows, then $\mu \ll \eta$: $\eta(\{j\}) := \frac{1}{j(j+1)}$ for $j = 1, 2, \cdots$, and $\eta(\{k+1, \cdots\}) = \sum_{j=k+1}^\infty \eta(\{j\}) = \frac{1}{k+1}$. Then, we can compute $\frac{d\nu_k^n}{d\eta^n} = \frac{\nu_k(s_k(x^n))}{\prod_{i=1}^n \eta(s_k(x_i))}$ to obtain $\frac{d\nu^n}{d\eta^n} = \sum_{k=1}^\infty \omega_k \frac{d\nu_k^n}{d\eta^n}$.

V. MDL/Bayesian Criteria in a General Sense

Let $\{X_i\}_{i=1}^n \sim \mu^n$ be stationary ergodic. Given n examples $x^n \in \prod_{i=1}^n X_i(\Omega)$, we estimate μ^n based on the following assumptions:

1) the true μ^n is known to be specified with some model in a countable set M and some parameters;
2) for each $m \in M$, there exists $\nu^n[m]$ such that with probability one
$$\frac{1}{n} \log \frac{d\mu^n[m]}{d\nu^n[m]}(x^n) \to 0$$
for $\mu^n[m]$ with model m and arbitrary parameters; and
3) the a prior probability $\pi[m]$ of $m \in M$ is known.

By the MDL/Bayesian criteria, we mean to apply the following decision rule:

$$\frac{\pi[m]}{\pi[m']} \cdot \frac{d\nu^n[m]}{d\nu^n[m']}(x^n) > 1 \Longleftrightarrow m \text{ is prefer to } m' \ .$$

A. Markov Order Identification for Continuous Sources

In general, by the Markov order of a stationary ergodic source $\{X_i\}_{i=-\infty}^\infty$ we mean the minimal k such that $\{X_i\}_{i=-\infty}^0$ and $\{X_i\}_{k+1}^\infty$ are conditionally independent given $\{X_i\}_{i=1}^k$.

Based on Theorem 1, for each $k = 0, 1, \cdots$, we can construct $\nu^n[k]$ such that $\int d\nu^n[k](x^n) \leq 1$, and $\frac{1}{n} \log \frac{d\mu^n[k]}{d\nu^n[k]}(x^n) \to 0$ with probability one for $\mu^n[k]$ with order k and arbitrary parameters.

Suppose we are given actually emitted n examples $x^n \in X^n(\Omega)$. Let $\pi[k]$: the a prior probability of k, and η such that $\mu[k] \ll \eta$, $k = 0, 1, \cdots$. Then, for each $k = 0, 1, \cdots$ we can construct $\frac{d\nu[k]}{d\eta}$, and choose k such that $\pi[k]\frac{d\nu[k]}{d\eta}(x^n)$ is maximized given the x^n.

B. Feature Selection Including Continuous Attributes

Consider again the feature selection problem discussed in Example 2, and give a solution to the problem raised in Introduction. Let $X = \{X^{(j)}\}_{j=1}^m$ and Y random variables with $|Y_i(\Omega)| < \infty$.

Based on Theorem 1, for each $F \subseteq \{1, \cdots, m\}$, construct $\nu_{XY}^n[F]$ and $\nu_X^n[F]$ such that $\int d\nu_{XY}^n[F](x^n, y^n) \leq 1$, $\int d\nu_X^n[F](x^n) \leq 1$ and

$$\frac{1}{n} \log \frac{d\mu_{XY}^n[F]}{d\nu_{XY}^n[F]}(x^n, y^n) \to 0 \ , \ \frac{1}{n} \log \frac{d\mu_X^n[F]}{d\nu_X^n[F]}(x^n) \to 0$$

for $\mu^n_{XY}[F]$ and $\mu_X[F]$ with feature set F and arbitrary parameters.

Suppose we are given actually emitted n examples $(x^n,y^n) \in X^n(\Omega)\times Y^n(\Omega)$ with $X^n(\Omega)=\prod_{i=1}^n\prod_{j=1}^m X_i^{(j)}(\Omega)$. Let $\pi[F]$: the a prior probability of $F\subseteq\{1,\cdots,m\}$, and η such that $\mu_X[F]\ll\eta$. Then, for $F\subseteq\{1,\cdots,m\}$, we can construct $\dfrac{d\nu_{XY}[F]}{d\eta}$ and $\dfrac{d\nu_X[F]}{d\eta}$, and choose F such that $\pi[F]\dfrac{d\nu_{XY}[F]}{d\eta}(x^n,y^n)/\dfrac{d\nu_X[F]}{d\eta}(x^n)$ is maximized given (x^n,y^n).

Example 4: Suppose $X^{(1)}(\Omega)=[0,1)$, $X^{(2)}(\Omega)=\{0,1\}$, $X^{(3)}(\Omega)=\{1,2,\cdots\}$. We estimate $F\in\{\{\},\{1\},\{2\},\{1,2\}\}$ given $\{(x_i^{(1)},x_i^{(2)},x_i^{(3)})\}_{i=1}^n\in\{X^{(1)}(\Omega)\times X^{(2)}(\Omega)\times X^{(3)}(\Omega)\}^n$. Let $\{A_k^{(1)}\}:=\{A_k\}$ be as in Example 2, let $A_0^{(2)}=\{\{0,1\}\}$, $A_1^{(2)}=A_2^{(2)}=\cdots=\{\{0\},\{1\}\}$, and $\{A_k^{(3)}\}:=\{A_k\}$ in Example 3, respectively. Also, let $\eta^{(1)}:=\eta$ in Example 2, $\eta^{(2)}(0)=\eta^{(2)}(1)=\frac{1}{2}$, and $\eta^{(3)}:=\eta$ in Example 4, respectively. Let $B:=X^{(3)}(\Omega)$.

For example, let $A:=X^{(1)}(\Omega)$. We compute the value of $\dfrac{d\nu^n}{d\eta^n}(b^n|a^n)$ for $F=\{1\}$ and $a^n\in A^n$, $b^n\in B^n$. For each $k=1,2,\cdots$, let $A_k:=A_k^{(1)}\times A_k^{(3)}$ with $m:=2^k(k+1)$, $c_n(a,b)$ the number of occurrences of $(a,b)\in A_k$ in $(a^n,b^n)\in A_k^n$, and

$$\nu_k^n(a^n,b^n):=\frac{\Gamma(\frac{m}{2})\prod_{(a,b)\in A_k}\Gamma(c_n(a,b)+\frac{1}{2})}{\Gamma(n+\frac{m}{2})\Gamma(\frac{1}{2})^m}\ .$$

Also, let $\eta(a,b):=\eta^{(1)}(a)\eta^{(3)}(b)=2^{-k}\cdot\dfrac{1}{j(j+1)}$ for $(a,b)\in A_k$ if $b=\{j\}$, and $\eta(a^n,b^n):=\prod_{i=1}^n\{\eta^{(1)}(a_i)\eta^{(3)}(b_i)\}$ for $(a^n,b^n)\in A_k^n$. We estimate $\dfrac{d\mu_k^n}{d\eta^n}$ and $\dfrac{d\mu^n}{d\eta^n}$ by $\frac{d\nu_k^n}{d\eta^n}(a^n,b^n)=\frac{\nu_k^n(a^n,b^n)}{\eta^n(a^n,b^n)}$ and $\frac{d\nu^n}{d\eta^n}(a^n,b^n)=\sum_{k=1}^\infty\omega_k\frac{d\nu_k^n}{d\eta^n}(a^n,b^n)$, respectively. On the other hand, from Example 3, we immediately obtain estimation $\dfrac{d\nu^n}{d\eta^n}(a^n)$ of $\dfrac{d\mu^n}{d\eta^n}(a^n)$ for $a^n\in A^n$. Then, we obtain the value of

$$\frac{d\nu^n}{d\eta^n}(b^n|a^n)=\left[\frac{d\nu^n}{d\eta^n}(a^n,b^n)\right]/\left[\frac{d\nu^n}{d\eta^n}(a^n)\right]$$

for $F=\{1\}$. Similarly, we obtain the values of $\dfrac{d\nu^n}{d\eta^n}$ for $F=\{2\},\{\},\{1,2\}$. Finally, we choose the maximum value among the four to choose $F\in\{\{\},\{1\},\{2\},\{1,2\}\}$.

VI. Concluding Remarks

We proposed the MDL/Bayesian criteria in a general sense, and illustrate the idea in the two examples:

- Markov order identification for continuous sources
- feature selection containing continuous attributes

We removed the constraint that each attribute should take a value in a finite set.

One could quantize continuous data to obtain the description length assuming each underlying model. However, the cell size in the partition should be optimized to obtain the correct MDL/Bayesian criteria: overestimation and underestimation may occur if the cell size is too small and too large, respectively. Intuitive speaking, our method utilizes weighting partitions to obtain more robust estimations compared with estimating only one partition.

Appendix: Proof of Theorem 1

Proof: for $k=1,2,\cdots$, $\dfrac{d\nu^n}{d\eta^n}(x^n)\geq\omega_k\dfrac{d\nu_k^n}{d\eta^n}(x^n)$, so that $\dfrac{d\mu^n}{d\nu^n}(x^n)\leq\dfrac{1}{\omega_k}\dfrac{d\mu^n}{d\nu_k^n}(x^n)$, thus $\frac{1}{n}\log\frac{d\mu^n}{d\nu^n}(x^n)$ is upper-bounded by

$$-\frac{1}{n}\log\omega_k+\frac{1}{n}\log\frac{d\mu_k^n}{d\nu^n}(x^n)+\frac{1}{n}\log\frac{d\mu^n}{d\nu^n}(x^n)-\frac{1}{n}\log\frac{d\mu_k^n}{d\nu^n}(x^n)$$

with probability one. Note that for each $k=1,2,\cdots$, from (3), we have

$$\frac{1}{n}\log\frac{d\mu_k^n}{d\nu_k^n}(x^n)=\frac{1}{n}\log\frac{P_k(s_k(x^n))}{Q_k(s_k(x^n))}\to 0\ .$$

Notice that $\int d\nu^n\leq 1$ and Proposition 3 imply $0\leq D(\mu||\nu)$ and $D(\mu||\nu)\leq D(\mu||\mu_k)$ for $k=1,2,\cdots$, respectively. Since we assume $D(\mu_k||\nu)\to D(\mu||\nu)$ $(k\to\infty)$, $D(\mu||\nu)=0$ is required, which implies Theorem 1. □

Proposition 3 (Barron, 1985 [1]): With probability one as $n\to\infty$,

$$\frac{1}{n}\log\frac{d\mu^n}{d\nu^n}(x^n)\to D(\mu||\nu)\ .$$

References

[1] A. R. Barron. “The Strong Ergodic Theorem for Densities: Generalized Shannon-McMillan-Breiman Theorem”, *Annals. Probability* 13(4):1292-1303 (1985).

[2] W.L. Buntine, “Learning Classification Trees.” *Statistics and Computing* 2: 63-73 (1991).

[3] P. Billingsley. *Probability & Measure* (1995): (3rd ed.). New York : Wiley.

[4] G.F. Cooper, Edward Herskovits. “A Bayesian Method for the Induction of Probabilistic Networks from Data.” *Machine Learning* 9: 309-347 (1992)

[5] T. M. Cover and J. A. Thomas. *Elements of Information Theory* (1995): (2nd ed.). New York : Wiley.

[6] R.E. Krichevsky and V.K. Trofimov, “The Performance of Universal Encoding”, *IEEE Trans. Inform. Theory* 27(2): 199-207 (1981).

[7] S. Kullback and R. A. Leibler. “On information and sufficiency”. *Ann. Math. Statistics* 22(1):79-86, 3 (1951).

[8] J.Rissanen, “Modeling by shortest data description”. *Automatica* 14: 465-471 (1978).

[9] B. Ryabko. “Compression-Based Methods for Nonparametric Prediction and Estimation of Some Characteristics of Time Series.” *IEEE Trans. on Inform. Theory*, 55(9):4309-4315 (2009).

[10] R. Solomonoff, ”A Formal Theory of Inductive Inference” *Information and Control*: 7(1)(2) pp 1-22, 224-254 (1964).

[11] J. Suzuki: On Strong Consistency of Model Selection in Classification. *IEEE Trans. on Inform. Theory* 52(11): 4767-4774 (2006).

[12] J. Suzuki, “A Construction of Bayesian Networks from Databases on an MDL Principle”, *The Ninth Conference on Uncertainty in Artificial Intelligence*, Washington D. C., pages 266-273, 7 (1993).

[13] J. Suzuki, “The Universal Measure for General Sources and its Application to MDL/Bayesian Criteria”, *Data Compression Conference* 2011, Snowbird, Utah (2011).