A Conjecture on Strongly Consistent Learning

Joe Suzuki

Osaka University

July 7, 2009

Outline

1. Stochastic Learning with Model Selection
2. Learning CP
3. Learning ARMA
4. Conjecture
5. Summary

Probability Space $(\Omega, \mathcal{F}, \mu)$

Ω: the entire set

$\mathcal{F}$: the set of events over Ω

$\mu : \mathcal{F} \to [0, 1]$: a probability measure ($\mu(\Omega) = 1$)

Stochastic Learning

$X : \Omega \to \mathbb{R}$: random variable
μ_X: the measure w.r.t. X
$x_1, \cdots, x_n \in X(\Omega)$: n samples

Induction/Inference

- $\mu_X \longmapsto x_1, \cdots, x_n$ (Random Number Generation)
- $x_1, \cdots, x_n \longmapsto \mu_X$ (Stochastic Learning)

Two Problems with Model Selection

Conditional Probability for Y given X

$\mu_{Y|X}$
Identify the equivalence relation in X from samples.

$$x \sim x' \iff \mu(Y=y|X=x) = \mu(Y=y|X=x') \text{ for } y \in Y(\Omega)$$

ARMA for $\{X_n\}_{n=-\infty}^{\infty}$

$X_n + \sum_{j=1}^{k} \lambda_j X_{n-j} \sim \mathcal{N}(0, \sigma^2)$ with $\{\lambda\}_{j=1}^{k}$ $(0 \le k < \infty)$
Identify the true order k from samples.

This paper

compares CP and ARMA to find how close they are.

CP

$A \in \mathcal{F}$
$\mathcal{G} \subseteq \mathcal{F}$

CP of A given $\mathcal{G}$ (Radon-Nykodim)

$\exists\mathcal{G}$-measurable $g : \Omega \to \mathbb{R}$ s.t $\mu(A \cap G) = \int_G g(\omega)\mu(d\omega)$ for $G \in \mathcal{G}$.

CP $\mu_{Y|X}$

$A := (Y = y)$, $y \in Y(\Omega)$
$\mathcal{G}$: generated by subsets of X.

Assumption

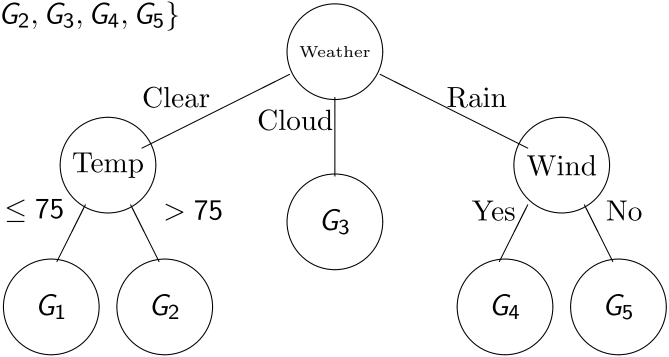
$|Y(\Omega)| < \infty$

Applications for CP

Stochastic Decision Trees, Stochastic Horn Clauseses $Y|X$, etc.

Find $\mathcal{G}$ from samples.

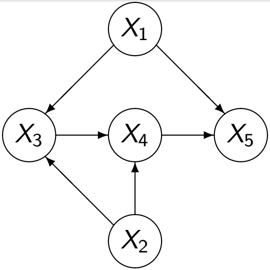
$\mathcal{G} = \{G_1, G_2, G_3, G_4, G_5\}$



Applications for CP (cont'd)

Finite Bayesian Networks $X_i|X_j, j \in \pi(i)$

Find $\pi(i) \subseteq \{1, \cdots, i-1\}$ from samples.



Learning CP

$z_i := (x_i, y_i) \in Z(\Omega) := X(\Omega) \times Y(\Omega)$
From $z^n := (z_1, \cdots, z_n) \in Z^n(\Omega)$, find a minimal $\mathcal{G}$.
$\mathcal{G}^*$: true
$\hat{\mathcal{G}}_n$: estimated

Assumption

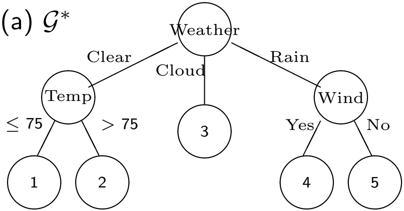
$|\mathcal{G}^*| < \infty$

Two Types of Errors

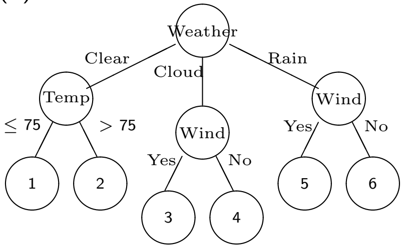
- $\mathcal{G}^* \subset \hat{\mathcal{G}}_n$ (Over Estimation)
- $\mathcal{G}^* \not\subseteq \hat{\mathcal{G}}_n$ (Under Estimation)

Example: Quinlan's Q4.5

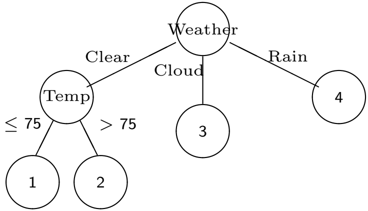
(a) $\mathcal{G}^*$



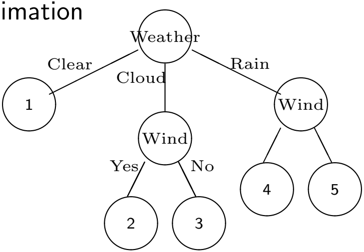
(b) Overestimation



(c) Underestimation



(d) Underestimation



Information Criteria for CP

Given $z^n \in Z(\Omega)$, find $\mathcal{G}$ minimizing

$$I(\mathcal{G}, z^n) := H(\mathcal{G}, z^n) + \frac{k(\mathcal{G})}{2} d_n$$

$H(\mathcal{G}, z^n)$: Empirical Entropy (Fitness of z^n to $\mathcal{G}$)
$k(\mathcal{G})$: the # of Parameters (Simplicity of $\mathcal{G}$)
$d_n \geq 0$: $\frac{d_n}{n} \to 0$

$d_n = \log n$ BIC/MDL
$d_n = 2$ AIC

Consistency

Consistency ($\hat{\mathcal{G}}_n \longrightarrow \mathcal{G}^*$ $(n \to \infty)$)

Weak Consistency Probability Convergence ($O(1) < d_n < o(n)$)

Strong Consistency Almost Surely Convergence (MDL/BIC etc.)

AIC ($d_n = 2$) is not consistent because $\{d_n\}$ is too small !

Problem

What is the minimum $\{d_n\}$ for Strong Consistency ?

Answer (Suzuki, 2006)

$d_n = 2 \log \log n$

(the Law of Iterated Logarithms)

Error Probability for CP

$\mathcal{G}^* \subset \mathcal{G}$ (Over Estimation)

$$\begin{aligned} & \mu\{\omega \in \Omega | I(\mathcal{G}, Z^n(\omega)) < I(\mathcal{G}^*, Z^n(\omega))\} \\ = & \mu\{\omega \in \Omega | \chi^2_{K(\mathcal{G})-K(\mathcal{G}^*)}(\omega) > (K(\mathcal{G}) - K(\mathcal{G}^*))d_n\} \end{aligned}$$

$\chi^2_{K(\mathcal{G})-K(\mathcal{G}^*)} \sim \chi^2$ of freedom $K(\mathcal{G}) - K(\mathcal{G}^*)$

$\mathcal{G}^* \not\subseteq \mathcal{G}$ (Under Estimation)

$$\mu\{\omega \in \Omega | I(\mathcal{G}, Z^n(\omega)) < I(\mathcal{G}^*, Z^n(\omega))$$

diminishes exponentially with n

Applications for ARMA

Time Series $X_i|X_{i-k},\cdots,X_{i-1}$

$$X_i+\sum_{j=1}^{k}\lambda_j X_{i-j}\sim\mathcal{N}(0,\sigma^2)$$

Find k from samples.

Gaussian Bayesian Networks $X_i|X_j, j\in\pi(i)$

$$X_i+\sum_{j\in\pi(i)}\lambda_{j,i}X_j\sim\mathcal{N}(0,\sigma_i^2)$$

Find $\pi(i)\subseteq\{1,\cdots,i-1\}$ from samples.

Learning ARMA

$k \geq 0$
$\{\lambda_j\}_{j=1}^k$: $\lambda_i \in \mathbb{R}$
$\sigma^2 \in \mathbb{R}_{>0}$

$$\{X_i\}_{i=-\infty}^{\infty}: \; X_i + \sum_{j=1}^{k} \lambda_j X_{i-j} = \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

Given $x^n := (x_1, \cdots, x_n) \in X_1(\Omega) \times \cdots \times X_n(\Omega)$, estimate

k: known $\{\lambda_j\}_{j=1}^k$, σ^2

k: Unknown k as well as $\{\lambda_j\}_{j=1}^k$, σ^2

Yule-Walker

If k is known, solve $\{\hat{\lambda}_{j,k}\}_{j=1}^{k}$ and $\hat{\sigma}_k^2$ w.r.t.

$$\bar{x} := \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$c_j := \frac{1}{n}\sum_{i=1}^{n-j}(x_i - \bar{x})(x_{i+j} - \bar{x}) \ , \ j = 0, \cdots, k$$

$$\begin{bmatrix} -1 & c_1 & c_2 & \cdots & c_k \\ 0 & c_0 & c_1 & \cdots & c_{k-1} \\ 0 & c_1 & c_0 & \cdots & c_{k-2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & c_{k-1} & c_{k-2} & \cdots & c_0 \end{bmatrix} \begin{bmatrix} \hat{\sigma}_k^2 \\ \hat{\lambda}_{1,k} \\ \hat{\lambda}_{2,k} \\ \vdots \\ \hat{\lambda}_{k,k} \end{bmatrix} = \begin{bmatrix} -c_0 \\ -c_1 \\ -c_2 \\ \vdots \\ -c_k \end{bmatrix}$$

Information Criteria for ARMA

If k is unknown, given $x^n \in X^n(\Omega)$, find k minimiziing

$$I(k, x^n) := \frac{1}{2}\log\hat{\sigma}_k^2 + \frac{k}{2}d_n$$

$\hat{\sigma}_k^2$: obtain using Yule-Walker

$d_n \geq 0$: $\frac{d_n}{n} \to 0$

$d_n = \log n$ BIC/MDL

$d_n = 2$ AIC

$d_n = 2\log\log n$ Hannan-Quinn (1979)

$\Longrightarrow$ the minimum $\{d_n\}$ satisfying Strong Consistency

Suzuki (2006) was inspired by Hannan-Quinn (1979) !

Error Probability for ARMA

k_*: true Order

$k_* > k$ (Under Estimation)

$$\mu\{\omega \in \Omega | I(k, X^n(\omega)) < I(k_*, X^n(\omega))\}$$

diminishes exponentially with n

Error Probability for ARMA (cont'd)

Conjecture: $k^* < k$ (Over Estimation)

$$\begin{aligned} & \mu\{\omega \in \Omega | I(k, X^n(\omega)) < I(k_*, X^n(\omega))\} \\ = \ & \mu\{\omega \in \Omega | \chi^2_{k-k_*}(\omega) > (k-k_*)d_n\} \end{aligned}$$

$\chi^2_{k-k_*} \sim \chi^2$ of freedom $k - k_*$

In fact, ...

For $k > k_*$ and large n (use $\hat{\sigma}_k^2 = (1-\hat{\lambda}_{k,k}^2)\sigma_{k-1}^2$),

$$\frac{1}{2}\log\hat{\sigma}_k^2 + \frac{k}{2}d_n < \frac{k_*}{2}\log\sigma_{k_*}^2 + \frac{k_*}{2}d_n$$

$$\iff \sum_{j=k_*+1}^{k}\log\frac{\hat{\sigma}_j^2}{\hat{\sigma}_{j-1}^2} > (k-k_*)d_n$$

$$\iff \sum_{j=k_*+1}^{k}\log(1-\hat{\lambda}_{k,k}^2) > d_n$$

$$\iff \sum_{j=k_*+1}^{k}\hat{\lambda}_{k,k}^2 > (k-k_*)d_n$$

It is very likely that $\sum_{j=k_*+1}^{k}\hat{\lambda}_{k,k}^2 \sim \chi_{k-k_*}^2$ although $\hat{\lambda}_{k,k} \not\sim \mathcal{N}(0,1)$.

Summary

CP and ARMA are similar

1. Results in ARMA can be applied to CP
2. Results in Gausssian BN can be applied to finite BN.

The conjecture is likely enough ?

Answer: Perhaps. Several evidences.

$\{Z_i\}_{i=1}^n$: independent
$Z_i \sim \mathcal{N}(0,1)$

$$rankA = n-k, A^2 = A \Longrightarrow {}^t[Z_1, \cdots, Z_n]A[Z_1, \cdots, Z_n] \sim \chi^2_{n-k}$$