On the Foundations of Universal Sequence Prediction

Marcus Hutter

Istituto Dalle Molle di Studi sull'Intelligenza Artificiale
IDSIA, Galleria 2, CH-6928 Manno-Lugano, Switzerland
marcus@idsia.ch, http://www.idsia.ch/~marcus

TAMC, 15 - 20 May 2006

Contents

Abstract

Solomonoff completed the Bayesian framework by providing a rigorous, unique, formal, and universal choice for the model class and the prior. I will discuss in breadth how and in which sense universal (non-i.i.d.) sequence prediction solves various (philosophical) problems of traditional Bayesian sequence prediction. I show that Solomonoff's model possesses many desirable properties: Fast convergence, and in contrast to most classical continuous prior densities has no zero p(oste)rior problem, i.e. can confirm universal hypotheses, is reparametrization and regrouping invariant, and avoids the old-evidence and updating problem. It even performs well (actually better) in non-computable environments.

Induction Examples

Sequence prediction: Predict weather/stock-quote/... tomorrow, based on past sequence. Continue IQ test sequence like 1,4,9,16,?

Classification: Predict whether email is spam.
Classification can be reduced to sequence prediction.

Hypothesis testing/identification: Does treatment X cure cancer?
Do observations of white swans confirm that all ravens are black?

These are instances of the important problem of inductive inference or time-series forecasting or sequence prediction.

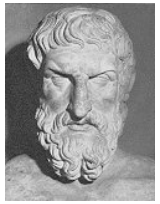
Problem: Finding prediction rules for every particular (new) problem is possible but cumbersome and prone to disagreement or contradiction.

Goal: Formal general theory for prediction.

Beyond induction: active/reward learning, fct. optimization, game theory.

Foundations of Universal Induction

Ockhams' razor (simplicity) principle
Entities should not be multiplied beyond necessity.



Epicurus' principle of multiple explanations
If more than one theory is consistent with the observations, keep all theories.

Bayes' rule for conditional probabilities
Given the prior belief/probability one can predict all future probabilities.

Turing's universal machine
Everything computable by a human using a fixed procedure can also be computed by a (universal) Turing machine.

Kolmogorov's complexity
The complexity or information content of an object is the length of its shortest description on a universal Turing machine.

Solomonoff's universal prior=Ockham+Epicurus+Bayes+Turing
Solves the question of how to choose the prior if nothing is known.
$\Rightarrow$ universal induction, formal Occam, AIT,MML,MDL,SRM,...

Bayesian Sequence Prediction and Confirmation

- Assumption: Sequence $\omega \in \mathcal{X}^\infty$ is sampled from the "true" probability measure μ, i.e. $\mu(x) := \mathbf{P}[x|\mu]$ is the μ-probability that ω starts with $x \in \mathcal{X}^n$.
- Model class: We assume that μ is unknown but known to belong to a countable class of environments=models=measures $\mathcal{M} = \{\nu_1, \nu_2, ...\}$.
- Hypothesis class: $\{H_\nu : \nu \in \mathcal{M}\}$ forms a mutually exclusive and complete class of hypotheses.
- Prior: $w_\nu := \mathbf{P}[H_\nu]$ is our prior belief in H_ν

$\Rightarrow$ Evidence: $\xi(x) := \mathbf{P}[x] = \sum_{\nu\in\mathcal{M}} \mathbf{P}[x|H_\nu]\mathbf{P}[H_\nu] = \sum_\nu w_\nu \nu(x)$ must be our (prior) belief in x.

$\Rightarrow$ Posterior: $w_\nu(x) := \mathbf{P}[H_\nu|x] = \frac{\mathbf{P}[x|H_\nu]\mathbf{P}[H_\nu]}{\mathbf{P}[x]}$ is our posterior belief in ν (Bayes' rule).

Convergence and Decisions

Goal: Given sequence $x_1 x_2 ... x_{t-1}$, predict its likely continuation x_t.

Expectation w.r.t. μ: $\mathbf{E}[f(\omega_{1:n})] := \sum_{x \in \mathcal{X}^n} \mu(x) f(x)$

KL-divergence: $D_n(\mu||\xi) := \mathbf{E}[\ln \frac{\mu(\omega_{1:n})}{\xi(\omega_{1:n})}] \leq \ln w_\mu^{-1} \; \forall n$

Hellinger distance: $h_t(\omega_{<t}) := \sum_{a \in \mathcal{X}} (\sqrt{\xi(a|\omega_{<t})} - \sqrt{\mu(a|\omega_{<t})})^2$

Rapid convergence: $\boxed{\sum_{t=1}^{\infty} \mathbf{E}[h_t(\omega_{<t})] \leq D_\infty \leq \ln w_\mu^{-1} < \infty}$ implies $\xi(x_t|\omega_{<t}) \to \mu(x_t|\omega_{<t})$, i.e. ξ is a good substitute for unknown μ.

Bayesian decisions: Bayes-optimal predictor Λ_ξ suffers instantaneous loss $l_t^{\Lambda_\xi} \in [0,1]$ at t only slightly larger than the μ-optimal predictor Λ_μ.

Pareto-optimality of Λ_ξ: Every predictor with loss smaller than Λ_ξ in some environment $\mu \in \mathcal{M}$ must be worse in another environment.

Generalization: Continuous Classes $\mathcal{M}$

In statistical parameter estimation one often has a continuous hypothesis class (e.g. a Bernoulli(θ) process with unknown $\theta \in [0,1]$).

$$\mathcal{M} := \{\nu_\theta : \theta \in I\!R^d\}, \quad \xi(x) := \int_{I\!R^d} d\theta\, w(\theta)\, \nu_\theta(x), \quad \int_{I\!R^d} d\theta\, w(\theta) = 1$$

Under weak regularity conditions [CB90,H'03]:

Theorem: $D_n(\mu||\xi) \;\leq\; \ln w(\mu)^{-1} + \frac{d}{2} \ln \frac{n}{2\pi} + O(1)$

where $O(1)$ depends on the local curvature (parametric complexity) of $\ln \nu_\theta$, and is independent n for many reasonable classes, including all stationary (k^{th}-order) finite-state Markov processes ($k=0$ is i.i.d.).

$D_n \propto \log(n) = o(n)$ still implies excellent prediction and decision for most n.

How to Choose the Prior?

- Subjective: quantifying personal prior belief (not further discussed)
- Objective: based on rational principles (agreed on by everyone)
- Indifference or symmetry principle: Choose $w_\nu = \frac{1}{|\mathcal{M}|}$ for finite $\mathcal{M}$.
- Jeffreys or Bernardo's prior: Analogue for compact parametric spaces $\mathcal{M}$.
- Problem: The principles typically provide good objective priors for small discrete or compact spaces, but not for "large" model classes like countably infinite, non-compact, and non-parametric $\mathcal{M}$.
- Solution: Occam favors simplicity $\Rightarrow$ Assign high (low) prior to simple (complex) hypotheses.
- Problem: Quantitative and universal measure of simplicity/complexity.

Kolmogorov Complexity K(x)

K. of string x is the length of the shortest (prefix) program producing x:

$$K(x) := \min_p\{l(p) : U(p) = x\}, \quad U = \text{universal TM}$$

For non-string objects o (like numbers and functions) we define $K(o) := K(\langle o \rangle)$, where $\langle o \rangle \in \mathcal{X}^*$ is some standard code for o.

+ Simple strings like $000...0$ have small K,
  irregular (e.g. random) strings have large K.
- The definition is nearly independent of the choice of U.
+ K satisfies most properties an information measure should satisfy.
+ K shares many properties with Shannon entropy but is superior.
− $K(x)$ is not computable, but only semi-computable from above.

Fazit: K is an excellent universal complexity measure, suitable for quantifying Occam's razor.

The Universal Prior

- Quantify the complexity of an environment ν or hypothesis H_ν by its Kolmogorov complexity $K(\nu)$.
- Universal prior: $w_\nu = \boxed{w_\nu^U := 2^{-K(\nu)}}$ is a decreasing function in the model's complexity, and sums to (less than) one.

$\Rightarrow$ $D_n \leq K(\mu)\ln 2$, i.e. the number of ε-deviations of ξ from μ or l^{Λ_ξ} from l^{Λ_μ} is proportional to the complexity of the environment.

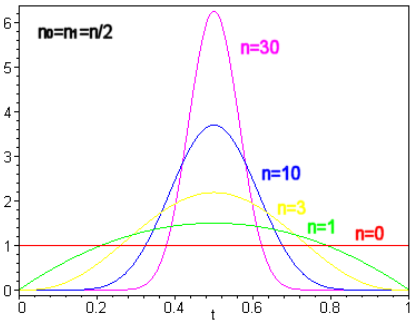
- No other semi-computable prior leads to better prediction (bounds).
- For continuous $\mathcal{M}$, we can assign a (proper) universal prior (not density) $w_\theta^U = 2^{-K(\theta)} > 0$ for computable θ, and 0 for uncomp. θ.
- This effectively reduces $\mathcal{M}$ to a discrete class $\{\nu_\theta \in \mathcal{M} : w_\theta^U > 0\}$ which is typically dense in $\mathcal{M}$.
- This prior has many advantages over the classical prior (densities).

Example: Bayes' and Laplace's Rule

Let $x \in \mathcal{X}^n = \{0,1\}^n$ be generated by a coin with bias $\theta \in [0,1]$

i.e. $\nu_\theta(x) = \mathbf{P}[x|H_\theta] = \theta^{n_1}(1-\theta)^{n_0}$, $n_1 = x_1 + ... + x_n = n - n_0$.

Bayes (1763) assumed a uniform prior density $w(\theta) = 1$.



The evidence is $\xi(x) = \int_0^1 \nu_\theta(x) w(\theta)\, d\theta = \frac{n_1! n_0!}{(n+1)!}$

Bayes: The posterior density $w(\theta|x) = \nu_\theta(x) w(\theta) / \xi(x)$
is strongly peaked around the frequency estimate $\hat{\theta} = \frac{n_1}{n}$ for large n.

Laplace (1812) asked for the pred. prob. $\xi(x_{n+1} = 1|x) = \frac{\xi(x1)}{\xi(x)} = \frac{n_1+1}{n+2}$

Laplace believed that the sun had risen for 5000 years = 1'826'213 days,
so he concluded that the probability of doomsday tomorrow is $\frac{1}{1826215}$.

The Problem of Zero Prior

= the problem of confirmation of universal hypotheses

Problem: If the prior is zero, then the posterior is necessarily also zero.

Example: Consider the hypothesis $H = H_1$ that all balls in some urn or all ravens are black (=1) or that the sun rises every day.

Starting with a prior density as $w(\theta) = 1$ implies that prior $\mathbf{P}[H_\theta] = 0$ for all θ, hence posterior $P[H_\theta|1..1] = 0$, hence H never gets confirmed.

3 non-solutions: define $H = \{\omega = 1^\infty\}$ | use finite population | abandon strict/logical/all-quantified/universal hypotheses in favor of soft hyp.

Solution: Assign non-zero prior to $\theta = 1 \quad \Rightarrow \quad \mathbf{P}[H|1^n] \to 1$.

Generalization: Assign non-zero prior to all "special" θ, like $\frac{1}{2}$ and $\frac{1}{6}$, which may naturally appear in a hypothesis, like "is the coin or die fair".

Universal solution: Assign non-zero prior to all comp. θ, e.g. $w_\theta^U = 2^{-K(\theta)}$

Reparametrization Invariance

- New parametrization e.g. $\psi = \sqrt{\theta}$, then the ψ-density $w'(\psi) = 2\sqrt{\theta}\, w(\theta)$ is no longer uniform if $w(\theta) = 1$ is uniform
  $\Rightarrow$ indifference principle is not reparametrization invariant (RIP).

- Jeffrey's and Bernardo's principle satisfy RIP w.r.t. differentiable bijective transformations $\psi = f^{-1}(\theta)$.

- The universal prior $w_\theta^U = 2^{-K(\theta)}$ also satisfies RIP w.r.t. simple computable f. (within a multiplicative constant)

Regrouping Invariance

- Non-bijective transformations:
  E.g. grouping ball colors into categories black/non-black.
- No classical principle is regrouping invariant.
- Regrouping invariance is regarded as a very important and desirable property. [Walley's (1996) solution: sets of priors]
- The universal prior $w_\theta^U = 2^{-K(\theta)}$ is invariant under regrouping, and more generally under all simple [computable with complexity $O(1)$] even non-bijective transformations. (within a multiplicative constant)
- Note: Reparametrization and regrouping invariance hold for arbitrary classes and are not limited to the i.i.d. case.

Universal Choice of Class $\mathcal{M}$

- The larger $\mathcal{M}$ the less restrictive is the assumption $\mu \in \mathcal{M}$.
- The class $\mathcal{M}_U$ of all (semi)computable (semi)measures, although only countable, is pretty large, since it includes all valid physics theories. Further, ξ_U is semi-computable [ZL70].
- Solomonoff's universal prior $M(x) :=$ probability that the output of a universal TM U with random input starts with x.
- Formally: $\boxed{M(x) := \sum_{p\,:\,U(p)=x*} 2^{-l(p)}}$ where the sum is over all (minimal) programs p for which U outputs a string starting with x.
- M may be regarded as a $2^{-l(p)}$-weighted mixture over all deterministic environments ν_p. ($\nu_p(x) = 1$ if $U(p) = x*$ and 0 else)
- $M(x)$ coincides with $\xi_U(x)$ within an irrelevant multiplicative constant.

The Problem of Old Evidence / New Theories

- What if some evidence $E \widehat{=} x$ (e.g. Mercury's perihelion advance) is known well-before the correct hypothesis/theory/model $H \widehat{=} \mu$ (Einstein's general relativity theory) is found?
- How shall H be added to the Bayesian machinery a posteriori?
- What should the "prior" of H be?
- Should it be the belief in H in a hypothetical counterfactual world in which E is not known?
- Can old evidence E confirm H?
- After all, H could simply be constructed/biased/fitted towards "explaining" E.

Solution of the Old-Evidence Problem

- The universal class $\mathcal{M}_U$ and universal prior w_ν^U formally solves this problem.
- The universal prior of H is $2^{-K(H)}$ independent of $\mathcal{M}$ and of whether E is known or not.
- Updating $\mathcal{M}$ is unproblematic, and even not necessary when starting with $\mathcal{M}_U$, since it includes all hypothesis (including yet unknown or unnamed ones) a priori.

Universal is Better than Continuous $\mathcal{M}$

- Although $\nu_\theta()$ and w_θ are incomp. for cont. classes $\mathcal{M}$ for most θ, $\xi()$ is typically computable. (exactly as for Laplace or numerically)

$\Rightarrow$ $\boxed{D_n(\mu||M) \stackrel{+}{\leq} D_n(\mu||\xi)+K(\xi)\ln 2 \text{ for all } \mu}$

- That is, M is superior to all computable mixture predictors ξ based on any (continuous or discrete) model class $\mathcal{M}$ and weight $w(\theta)$, save an additive constant $K(\xi)\ln 2 = O(1)$, even if environment μ is not computable.

- While $D_n(\mu||\xi) \sim \frac{d}{2}\ln n$ for all $\mu \in \mathcal{M}$, $D_n(\mu||M) \leq K(\mu)\ln 2$ is even finite for computable μ.

Fazit: $\boxed{\text{Solomonoff prediction works also in non-computable environments}}$

More Stuff / Critique / Problems

- Prior knowledge y can be incorporated by using "subjective" prior $w^U_{\nu|y} = 2^{-K(\nu|y)}$ or by prefixing observation x by y.
- Additive/multiplicative constant fudges and U-dependence is often (but not always) harmless.
- Incomputability: K and M can serve as "gold standards" which practitioners should aim at, but have to be (crudely) approximated in practice (MDL [Ris89], MML [Wal05], LZW [LZ76], CTW [WSTT95], NCD [CV05]).

Summary

Universal Solomonoff prediction solves/avoids/meliorates many problems of (Bayesian) induction. We discussed:

+ general total bounds for generic class, prior, and loss,
+ i.i.d./universal-specific instantaneous and future bounds,
+ the D_n bound for continuous classes,
+ indifference/symmetry principles,
+ the problem of zero p(oste)rior & confirm. of universal hypotheses,
+ reparametrization and regrouping invariance,
+ the problem of old evidence and updating,
+ that M works even in non-computable environments,
+ how to incorporate prior knowledge,
− the prediction of short sequences,
− the constant fudges in all results and the U-dependence,
− M's incomputability and crude practical approximations.

Literature

- M. Hutter, *Optimality of Universal Bayesian Prediction for General Loss and Alphabet.* Journal of Machine Learning Research 4 (2003) 971–1000. http://arxiv.org/abs/cs.LG/0311014
- M. Hutter, *Convergence and Loss Bounds for Bayesian Sequence Prediction.* IEEE Transactions on Information Theory, 49:8 (2003) 2061–2067. http://arxiv.org/abs/cs.LG/0301014
- M. Hutter and An. Muchnik, *Universal Convergence of Semimeasures on Individual Random Sequences.* Proc. 15th International Conf. on Algorithmic Learning Theory (ALT-2004) 234–248. http://arxiv.org/abs/cs.LG/0407057
- M. Hutter, *On the Foundations of Universal Sequence Prediction.* Proc. 3rd Annual Conference on Theory and Applications of Models of Computation (TAMC-2006), Beijing. http://www.idsia.ch/idsiareport/IDSIA-03-06.pdf
- M. Hutter, *Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability.* EATCS, Springer, 300 pages, 2005. http://www.idsia.ch/~marcus/ai/uaibook.htm