On Universal Prediction and Bayesian Confirmation

ANU

Marcus Hutter

Australian National University
Canberra, ACT, 0200, Australia
http://www.hutter1.net/

Summary

- Bayesian reasoning is consistent but incomplete.
- Solomonoff provides universal choice of model class & prior.
- I show that this solves the long-standing induction problem.

Abstract

Bayesian reasoning is a well-studied and successful framework for inductive inference, which includes hypothesis testing and confirmation, parameter estimation, sequence prediction, classification, and regression. But standard statistical guidelines for choosing the model class and prior are not always available or can fail, in particular in complex situations. Finding tailor-made solutions to every particular (new) such problem might be possible but is cumbersome and prone to disagreement or contradiction. What is desirable is a formal general theory for inductive inference, and for building general purpose intelligent machines, such a theory is not only desirable but indispensable.

Solomonoff completed the Bayesian framework by providing a rigorous, unique, formal, and universal choice for the model class and the prior. This "universal" Bayesian approach differs significantly from the classical objective as well as the subjective Bayesian philosophy. I show that Universal Bayes (UB) essentially solves the long-standing induction problem, at least from a philosophical and statistical perspective.

More specifically, I show that UB convergence rapidly and in contrast to using 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?

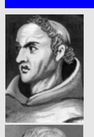
These are instances of the important problem of inductive inference or time-series forecasting or sequence prediction.

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

Goal: A single, formal, general, complete 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 Seq. Prediction & Confirmation

- Assumption: Sequence $\omega \in \mathcal{X}^\infty$ is sampled from the "true" probability measure μ, i.e. $\mu(x) := \mathrm{P}[x|\mu]$ 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 $\mathcal{M} = \{\nu_1, \nu_2, ...\}$.
  [no i.i.d./ergodic/stationary assumption]
- Hypothesis class: $\{H_\nu : \nu \in \mathcal{M}\}$ forms a mutually exclusive and complete class of hypotheses.
- Prior: $w_\nu := \mathrm{P}[H_\nu]$ is our prior belief in H_ν

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

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

Convergence and Decisions

Goal: Given seq. $x_{1:t-1} \equiv x_{<t} \equiv x_1 x_2 ... x_{t-1}$, predict 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: $\sum_{t=1}^\infty \mathbf{E}[(\sqrt{l_t^{\Lambda_\xi}} - \sqrt{l_t^{\Lambda_\mu}})^2] \leq \sum_{t=1}^\infty 2\mathbf{E}[h_t] < \infty$ implies rapid $l_t^{\Lambda_\xi} \to l_t^{\Lambda_\mu}$

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.

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

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).

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.

Universal is better than Continuous

- Problem of zero prior / confirmation of universal hypotheses:
  P[All ravens black$|n$ black ravens] $\begin{cases} \equiv 0 \text{ in Bayes-Laplace model} \\ \xrightarrow{fast} 1 \text{ for universal prior } w_\theta^U \end{cases}$
- Reparametrization and regrouping invariance: $w_\theta^U = 2^{-K(\theta)}$ always exists and is invariant w.r.t. all computable reparametrizations f. (Jeffrey prior only w.r.t. bijections, and does not always exist)
- The Problem of Old Evidence: No risk of biasing the prior towards past data, since w_θ^U is fixed and independent of $\mathcal{M}$.
- The Problem of New Theories: Updating of $\mathcal{M}$ is not necessary, since $\mathcal{M}_U$ includes already all.

- M predicts better than all other mixture predictors based on **any** (continuous or discrete) model class and prior, even in non-computable environments.

More Bounds

- Instantaneous i.i.d. bounds: For i.i.d. $\mathcal{M}$ with continuous, discrete, and universal prior, respectively:
  $\mathbf{E}[h_n] \lesssim \frac{1}{n}\ln w(\mu)^{-1}$ and $\mathbf{E}[h_n] \lesssim \frac{1}{n}\ln w_\mu^{-1} = \frac{1}{n}K(\mu)\ln 2$.
- Bounds for computable environments: Rapidly $M(x_t|x_{<t}) \to 1$ on every computable sequence $x_{1:\infty}$ (whichsoever, e.g. 1^∞ or the digits of π or e), i.e. M quickly recognizes the structure of the sequence.
- Weak instantaneous bounds: valid for all n and $x_{1:n}$ and $\bar{x}_n \neq x_n$: $2^{-K(n)} \lesssim M(\bar{x}_n|x_{<n}) \lesssim 2^{2K(x_{1:n}*)-K(n)}$
- Magic instance numbers: e.g. $M(0|1^n) \lesssim 2^{-K(n)} \to 0$, but spikes up for simple n. M is cautious at magic instance numbers n.
- Future bounds / errors to come: If our past observations $\omega_{1:n}$ contain a lot of information about μ, we make few errors in future: $\sum_{t=n+1}^\infty \mathbf{E}[h_t|\omega_{1:n}] \stackrel{+}{\leq} [K(\mu|\omega_{1:n}) + K(n)]\ln 2$

More Stuff / Critique / Problems

- Prior knowledge y can be incorporated by using "subjective" prior $w_{\nu|y}^U = 2^{-K(\nu|y)}$ 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]).

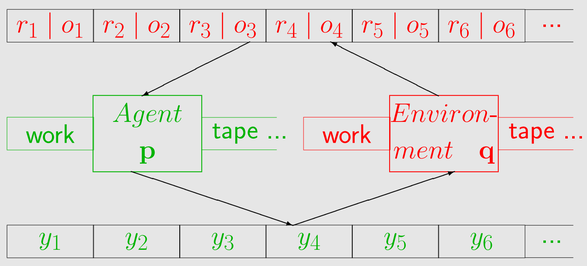
Summary

Universal Bayesian 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.

Generalization to ReActive Problems

Universal AI = Universal Induction + Sequential Decision Theory



AIXI: $y_k = \arg\max_{y_k} \sum_{x_k} ... \max_{y_m} \sum_{x_m} [r(x_k)+...+r(x_m)] M(x_{1:m}|y_{1:m})$

Claim: AIXI is the most intelligent environmental independent, i.e. universally optimal, agent possible.

Applications: Strategic Games, Function Minimization, Supervised Learning from Examples, Sequence Prediction, Classification.

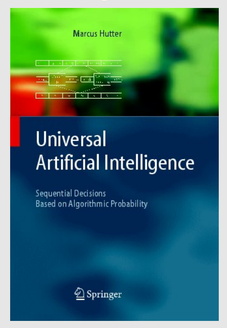
Literature

Paper1: *On Universal Prediction and Bayesian Confirmation.*
Theoretical Computer Science, 384:1 (2007) 33-48. [relevant]

Paper2: *A Philosophical Treatise of Universal Induction.*
Entropy, 13:6 (2011) 1076–1136. [gentle]

Paper3: *Universal Intelligence: A Definition of Machine Intelligence.*
Minds & Machines, 17:4 (2007) 391–444. [related]

Book intends to excite a broader AI audience about abstract Algorithmic Information Theory –and– inform theorists about exciting applications to AI.



Decision Theory = Probability + Utility Theory
Universal Induction = Ockham + Bayes + Turing
A Unified View of Artificial Intelligence