Optimality of Universal Bayesian
Prediction for General Loss and

Marcus Hutter

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

2000 – 2002

Universal Induction = Ockham + Epicurus + B

$$\frac{\text{Loss(Universal Prediction Scheme)}}{\text{Loss(Any other Prediction Scheme)}} \leq 1 + o($$

Table of Contents

Problem Setup

- Every induction problem can be phrased as a sequence pred
- Classification is a special case of sequence prediction. (With some tricks the other direction is also true)
- I'm interested in maximizing profit (minimizing loss). I'm not (primarily) interested in finding a (true/predictive/c
- Separating noise from data is *not* necessary in this setting!

My Position to Occam

- Most of us belief in or at least use the axioms of logic, proo natural numbers when doing science, without questioning th
- We should/must add Occam's razor in some quantified form because it is the foundation of machine learning and science
- There is (yet) no mathematical proof of Occam's razor, and an independent axiom, but there is lots of evidence that this

On the Foundations of Machine L

- Example: Algorithm/complexity theory: The goal is to find problems and to show lower bounds on their computation ti rigorously defined: algorithm, Turing machine, problem class ...
- Most disciplines start with an informal way of attacking a su get more and more formalized often to a point where they a Examples: set theory, logical reasoning, proof theory, probab infinitesimal calculus, quantum field theory, ...
- Machine learning: Tries to build and understand systems wh data, to make good prediction, which are able to generalize, Many terms only vaguely defined or there are many alternat

Occam to the Rescue

- Is it possible to give machine learning a rigorous mathemati framework/definition?
- Yes! Use Occam's razor, quantified in terms of Kolmogorov combine it with Bayes, and possibly sequential decision theo
- There is at the moment no alternative suggestion of how to learning rigorously.

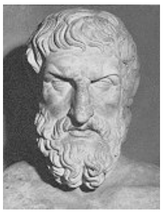
My view of (future) Machine lea

- Application = Solve learning tasks by approximating Kolmo (MML, MDL, SRM, and much more specific ones, like SVM
- Theory = Proof theorems, especially on convergence and ap
- Non-standard ML = Modify "Occam's axiom" with the goa better.

Induction = Predicting the Fu

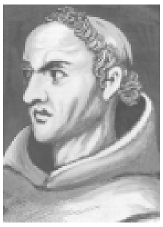
Extrapolate past observations to the future

but how can we know something about the fu



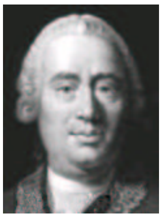
Epicurus' principle of multiple explanations

If more than one theory is consistent with the observa



Ockhams' razor (simplicity) principle

Entities should not be multiplied beyond necessity.



Hume's negation of Induction The only form of indu

tion as the conclusion is already logically contained in

Bayes' rule for conditional probabilities

Given the prior believe/probability one can predict a

Solomonoff's universal prior

Solves the question of how to choose the prior if no

Strings and Conditional Probab

Strings: $x = x_1 x_2 ... x_n$ with $x_t \in \mathcal{X}$ and $x_{1:m} := x_1 x_2 ... x_{m-1} x_m$

$\rho(x_1...x_n)$ is the probability that an (infinite) sequence starts wi

Heavy use of Bayes' rule in the following forms:

$$\rho(x_n|x_{<n}) = \rho(x_{1:n})/\rho(x_{<n}),$$

$$\rho(x_1...x_n) = \rho(x_1)\cdot\rho(x_2|x_1)\cdot...\cdot\rho(x_n|x_1...x_n$$

If the true prior probability $\mu(x_1...x_n)$ is known, then the optima minimize the μ- expected loss.

Interpretation of Probabiliti

Frequentist: Probabilities come from experiments.
Objectivist: Probabilities are real aspects of the world.
Subjectivist: Probabilities describe ones believe.

Probability of Sunrise Tomorr

What is the probability that the sun will rise tomorrow? It is $\mu($
lifetime of the sun in days.

$1 =$ sun raised. $0 =$ sun will not raise.

- The probability is undefined, because there has never been a tested the existence of the sum *tomorrow* (reference class p
- The probability is 1, because in all experiments that have be days) the sun raised.
- The probability is $1 - \epsilon$, where ϵ is the proportion of stars in explode in a supernova per day.
- The probability is $(d + 1)/(d + 2)$ (Laplace estimate by assu process with uniformly distributed raising prior probability p
- The probability can be derived from the type, age, size and sun, even though we never have observed another star with

Solomonoff solved the problem of unknown prior μ by introducin probability distribution ξ related to Algorithmic Information The

Kolmogorov Complexity

The Kolmogorov Complexity of a string x is the length of the sh
producing x.

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

The definition is "nearly" independent of the choice of U

$$|K_U(x) - K_{U'}(x)| < c_{UU'}, \quad K_U(x) \stackrel{\pm}{=} K_{U'}$$

$\stackrel{\pm}{=}$ indicates equality up to a constant $c_{UU'}$ independent of x.

K satisfies most properties an information measure should satisf
$K(xy) \stackrel{+}{\leq} K(x) + K(y)$.

$K(x)$ is not computable, but only co-enumerable (semi-computa

Universal Probability Distribut

The universal semimeasure is the probability that output of U st
input is provided with fair coin flips

$$\xi(x) \;=\; \sum_{\mu_i\in\mathcal{M}} w_{\mu_i}\cdot\mu_i(x) \;\stackrel{\times}{=}\; \sum_{p\,:\,U(p)=x*} 2^{-l(p)}, \quad e.g. \quad w$$

[Solomonoff 64]

Universality property of ξ: ξ dominates every computable probal

$$\xi(x) \;\geq\; w_{\mu_i}\!\cdot\!\mu_i(x) \quad \forall\mu_i\in\mathcal{M}$$

Furthermore, the μ expected squared distance sum between ξ a
computable μ

$$\sum_{t=1}^{\infty}\sum_{x_{1:t}}\mu(x_{<t})(\xi(x_t|x_{<t})-\mu(x_t|x_{<t}))^2 \;\leq\; \ln 2$$

[Solomonoff 78] (for binary alphabet)

$\Rightarrow \xi(x_n|x_{<n}) \stackrel{n\to\infty}{\longrightarrow} \mu(x_n|x_{<n})$ with μ probability 1 $\quad\Longrightarrow\quad \xi$ i

Convergence Theorem

The universal conditional probability $\xi(x_t|x_{<t})$ of the next symb
related to the true conditional probability $\mu(x_t|x_{<t})$ in the follov

$$i) \quad \sum_{t=1}^{n} \mathbf{E}\Big[\sum_{x_t}\Big(\mu(x_t|x_{<t})-\xi(x_t|x_{<t})\Big)^2\Big] \;\equiv\; S_n \;\leq\; D_n \;\leq$$

$$ii) \quad \sum_{x_t}\Big(\mu(x_t|x_{<t})-\xi(x_t|x_{<t})\Big)^2 \;\equiv\; s_t(x_{<t}) \;\leq\; d_t(x_{<t}) \to$$

$$iii) \quad \xi(x'_t|x_{<t}) \to \mu(x'_t|x_{<t}) \quad \text{for } t\to\infty \text{ with } \mu \text{ probability } 1$$

$$iv) \quad \sum_{t=1}^{n} \mathbf{E}\left[\left(\sqrt{\frac{\xi(x_t|x_{<t})}{\mu(x_t|x_{<t})}}-1\right)^2\right] \;\leq\; D_n \;\leq\; \ln w_\mu^{-1} \;<\; \infty$$

$$v) \quad \frac{\xi(x_t|x_{<t})}{\mu(x_t|x_{<t})} \to 1 \quad \text{for } t\to\infty \text{ with } \mu \text{ probability } 1$$

where $d_t = \sum_{x_n}\mu(x_n|x_{<n})\ln\frac{\mu(x_n|x_{<n})}{\xi(x_n|x_{<n})}$ and $D_n = \sum_{x_{1:n}}\mu(x_{1:}$
relative entropies, and w_μ is the weight of μ in ξ.

Universal Sequence Predicti

A prediction is very often the basis for some decision. The decisi
which itself leads to some reward or loss. Let $\ell_{x_t y_t} \in [0,1]$ be th
taking action $y_t \in \mathcal{Y}$ and $x_t \in \mathcal{X}$ is the t^{th} symbol of the sequenc
decision $\mathcal{Y} = \{$umbrella, sunglasses$\}$ based on weather forecasts

Loss	sunny	rainy
umbrella	0.3	0.1
sunglasses	0.0	1.0

The goal is to minimize the μ-expected loss. More generally we
prediction scheme

$$y_t^{\Lambda_\rho} \;:=\; \arg\min_{y_t \in \mathcal{Y}} \sum_{x_t} \rho(x_t|x_{<t}) \ell_{x_t y_t}$$

which minimizes the ρ-expected loss. The actual μ-expected lo
the t^{th} symbol and the total μ-expected loss in the first n predi

$$l_{t\Lambda_\rho}(x_{<t}) \;:=\; \sum_{x_t} \mu(x_t|x_{<t}) \ell_{x_t y_t^{\Lambda_\rho}} \quad , \quad L_n^{\Lambda_\rho} \;:=\; \sum_{t=1}^{n} \sum_{x_{<t}}$$

Loss Bounds (Main Theoren

$L_n^{\Lambda_\mu}$ made by the informed scheme Λ_μ,
$L_n^{\Lambda_\xi}$ made by the universal scheme Λ_ξ,
L_n^{Λ} made by any (causal) prediction scheme Λ.

$i)$ $L_n^{\Lambda_\mu} \leq L_n^{\Lambda}$ for any (causal) prediction scheme Λ.

$ii)$ $0 \leq L_n^{\Lambda_\xi} - L_n^{\Lambda_\mu} \leq 2D_n + 2\sqrt{L_n^{\Lambda_\mu} D_n}$

$iii)$ if $L_{\infty\Lambda_\mu}$ is finite, then $L_{\infty\Lambda_\xi}$ is finite

$iv)$ $L_n^{\Lambda_\xi}/L_n^{\Lambda_\mu} = 1 + O((L_n^{\Lambda_\mu})^{-1/2}) \stackrel{L_n^{\Lambda_\mu}\to\infty}{\longrightarrow} 1$

$v)$ $\sum_{t=1}^n \mathbf{E}[(l_{t\Lambda_\xi}(x_{<t}) - l_{t\Lambda_\mu}(x_{<t}))^2] \leq 2D_n \leq 2\ln w_\mu^{-}$

$vi)$ $0 \leq l_{t\Lambda_\xi}(x_{<t}) - l_{t\Lambda_\mu}(x_{<t}) \leq \begin{cases} \sqrt{2d_t(x_{<t})} \\ 2d_t(x_{<t}) + 2\sqrt{l_{t\Lambda_\mu}(x_{<t})} \end{cases}$

where $D_n := \sum_{x_{1:n}} \mu(x_{1:n}) \ln \frac{\mu(x_{1:n})}{\xi(x_{1:n})} \leq \ln \frac{1}{w_\mu} = \ln 2 \cdot K(\mu$

Remark: The bound is valid for any loss function $\in [0,1]$ with n
i.i.d., Markovian, stationary, ergodic, ...) on the structure of the

Example Application

A dealer has two dice, one with 2 white and 4 black faces, the o
2 black faces. He chooses a die according to some deterministic
we bet $s = \$3$ on white or black and receive $r = \$5$ for every corr

If we know μ, i.e. the die the dealer chooses, we should predict
sides and win money. Expected Profit (= –Loss): $P_{n\Lambda_\mu}/n \;=\; \frac{1}{3}$

If we don't know μ we can use Solomonoff prediction scheme Λ_ξ
the same profit:

$$P_{n\Lambda_\xi}/P_{n\Lambda_\mu}, \;=\; 1 - O(n^{-1/2})$$

Bound on Winning Time

Estimate of the number of rounds before reaching the winning z

$P_{n\Lambda_\xi} > 0$ if $L_n^{\Lambda_\xi} < 0$ if $n \;>\; 330 \ln 2 \cdot K(\mu) + O(1)$

Λ_ξ is asymptotically optimal with rapid convergence.

General Bound for Winning T

For every (passive) game of chance for which there exists a winn make money by using Λ_ξ even if you don't know the underlying process/algorithm.

Λ_ξ finds and exploits every regularity.

The time n needed to reach the winning zone is

$$n \;\leq \left(\frac{2p_\Delta}{\bar{p}_{n\Lambda_\mu}}\right)^2 \cdot \ln\frac{1}{w_\mu}, \quad \bar{p}_{n\Lambda_\mu} := \frac{1}{n}\sum_{t=1}^{n} p_{t\Lambda_\mu}, \quad p_\Delta =$$

Generalization: Continuous Probability

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

$$\mathcal{M} := \{\mu_\theta : \theta \in I\!R^d\}, \qquad \xi(x_{1:n}) := \int_{I\!R^d} d\theta\, w(\theta) \cdot \mu_\theta(x_{1:n}),$$

The only property of ξ needed was $\xi(x_{1:n}) \geq w_{\mu_i} \cdot \mu_i(x_{1:n})$ whic
dropping the sum over μ_i. Here, restrict the integral over $I\!R^d$ to
θ. For sufficiently smooth μ_θ and $w(\theta)$ we expect

$$\xi(x_{1:n}) \gtrsim |N_{\delta_n}| \cdot w(\theta) \cdot \mu_\theta(x_{1:n}) \quad \Longrightarrow \quad D_n \lesssim \ln \frac{1}{w_\mu}$$

The average Fisher information $\bar{\jmath}_n$ measures the curvature (par
$\ln \mu_\theta$. Under some weak regularity conditions on $\bar{\jmath}_n$ one can sho

$$D_n := \sum_{x_{1:n}} \mu(x_{1:n}) \ln \tfrac{\mu(x_{1:n})}{\xi(x_{1:n})} \leq \ln \tfrac{1}{w_\mu} + \tfrac{d}{2} \ln \tfrac{n}{2\pi} + \tfrac{1}{2}$$

i.e. D_n grows only logarithmically with n.

Optimality of the Universal Pre

- There are $\mathcal{M}$ and $\mu \in \mathcal{M}$ and weights w_μ for which the los

- The universal prior ξ is pareto-optimal, in the sense that the $\mathcal{F}(\mu_a, \rho) \leq \mathcal{F}(\mu_a, \xi)$ for all $\mu_a \in \mathcal{M}$ and strict inequality fo where $\mathcal{F}$ is the instantaneous or total squared distance s_t, S d_t, D_n , or error e_t, E_n, or loss l_t, L_n.

- ξ is elastic pareto-optimal in the sense that by accepting a s decrease in some environments one can only achieve a sligh in other environments.

- Within the set of enumerable weight functions with short pr weights $w_\nu = 2^{-K(\nu)}$ lead to the smallest performance bou (to $\ln w_\mu^{-1}$) constant in all enumerable environments.

$\boxed{\text{Does all this justify Occam's razor ?}}$

Larger & Smaller Environmental C

- all finitely computable probability measures
  ($\xi \notin \mathcal{M}$ in no sense computable)
- all enumerable (approximable from below) semi-measures [S
  ($\xi \in \mathcal{M}$ enumerable)
- all cumulatively enumerable semi-measures [Schmidhuber 01
  (distribution enumerable and $\in \mathcal{M}$)
- all approximable (asymptotically computable) measures [Sch
  ($\xi \notin \mathcal{M}$ in no sense computable)
- Speed prior related to Levin complexity and Levin search [Sc
  (which distributions are dominated?)
- finite-state automata instead of general Turing machines [Fe
  to Lempel-Ziv data compression ($\xi \notin \mathcal{M}$)

Generalization: The Universal AIξ

Universal AI = Universal Induction + Sequential Dec

Replace μ^{AI} in decision theory model AIμ by an appropriate ger

$$\xi(y\!\!\!x_{1:t}) \;:=\; \sum_{q:q(y_{1:t})=x_{1:t}} 2^{-l(q)}$$

$$y_t \;=\; \arg\max_{y_t}\sum_{x_t}\max_{y_{t+1}}\sum_{x_{t+1}}\;...\;\max_{y_m}\sum_{x_m}$$
$$(r(x_t)+\,...\,+r(x_m))\cdot\xi(x_{t:m}|y_{t:m})$$

Claim: AIξ is the most intelligent environmental independent, i.e
agent possible.

Applications: Strategic Games, Function Minimization, Supervis
Examples, Sequence Prediction, Classification.

[Proceedings of ECML-2001] and [http://www.hutt

Further Generalizations:

- Time and history dependent loss function in general interval
- Infinite (countable and uncountable) action/decision space.
- Partial Sequence Prediction.
- Independent Experiments & Classification.

Outlook:

- Infinite (prediction) alphabet $\mathcal{X}$.
- Delayed and Probabilistic Sequence Prediction.
- Unification with (Lossbounds for) aggregating strategies.
- Determine suitable performance measures for universal AIξ
- Study learning aspect of Λ_ξ and AIξ.
- Information theoretic interpretation of winning time.
- Implementation and application of Λ_ξ for specific finite $\mathcal{M}$.
- Downscale theory and results to MDL approach.

Conclusions

- Solomonoff's prediction scheme, which is related to Kolomo
  formally solves the general problem of induction.
- We proved convergence and loss-bounds for Solomonoff pre
  is well suited, even for difficult prediction problems.
- We proved several optimality properties for Solomonoff pred
- We made no structural assumptions on the probability distri
- The bounds are valid for any bounded loss function.
- We proved a bound on the time to win in games of chances
- Discrete and continuous probability classes have been consic
- Generalizations to active agents with reinforcement feedbacl
- At least all this is a lot of evidence that Occam's razor is a

See [http://www.idsia.ch/~marcus] for deta