COMP4620/8620: Advanced Topics in AI
Foundations of Artificial Intelligence

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

ANU

7 BAYESIAN SEQUENCE PREDICTION

- The Bayes-Mixture Distribution
- Relative Entropy and Bound
- Predictive Convergence
- Sequential Decisions and Loss Bounds
- Generalization: Continuous Probability Classes
- Summary

Bayesian Sequence Prediction: Abstract

We define the Bayes mixture distribution and show that the posterior converges rapidly to the true posterior by exploiting some bounds on the relative entropy. Finally we show that the mixture predictor is also optimal in a decision-theoretic sense w.r.t. any bounded loss function.

Notation: Strings & Probabilities

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

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

Conditional probability:

$$\rho_n := \rho(x_n | x_{<n}) = \rho(x_{1:n}) / \rho(x_{<n}),$$
$$\rho(x_1 ... x_n) = \rho(x_1) \cdot \rho(x_2 | x_1) \cdot ... \cdot \rho(x_n | x_1 ... x_{n-1}).$$

True data generating distribution: μ

The Bayes-Mixture Distribution ξ

- Assumption: The true (objective) environment μ is unknown.
- Bayesian approach: Replace true probability distribution μ by a Bayes-mixture ξ.
- Assumption: We know that the true environment μ is contained in some known countable (in)finite set $\mathcal{M}$ of environments.

Definition 7.1 (Bayes-mixture ξ)

$$\xi(x_{1:m}) := \sum_{\nu\in\mathcal{M}} w_\nu \nu(x_{1:m}) \quad \text{with} \quad \sum_{\nu\in\mathcal{M}} w_\nu = 1, \quad w_\nu > 0 \;\forall \nu$$

- The weights w_ν may be interpreted as the prior degree of belief that the true environment is ν, or $k^\nu = \ln w_\nu^{-1}$ as a complexity penalty (prefix code length) of environment ν.
- Then $\xi(x_{1:m})$ could be interpreted as the prior subjective belief probability in observing $x_{1:m}$.

A Universal Choice of ξ and $\mathcal{M}$

- We have to assume the existence of some structure on the environment to avoid the No-Free-Lunch Theorems [Wolpert 96].
- We can only unravel effective structures which are describable by (semi)computable probability distributions.
- So we may include *all* (semi)computable (semi)distributions in $\mathcal{M}$.
- Occam's razor and Epicurus' principle of multiple explanations tell us to assign high prior belief to simple environments.
- Using Kolmogorov's universal complexity measure $K(\nu)$ for environments ν one should set $w_\nu = 2^{-K(\nu)}$, where $K(\nu)$ is the length of the shortest program on a universal TM computing ν.
- The resulting mixture ξ is Solomonoff's (1964) universal prior.
- In the following we consider generic $\mathcal{M}$ and w_ν.

Relative Entropy

Relative entropy: $D(\mathbf{p}||\mathbf{q}) := \sum_i p_i \ln \frac{p_i}{q_i}$

Properties: $D(\mathbf{p}||\mathbf{q}) \geq 0$ and $D(\mathbf{p}||\mathbf{q}) = 0 \Leftrightarrow \mathbf{p} = \mathbf{q}$

Instantaneous relative entropy: $d_t(x_{<t}) := \sum_{x_t \in \mathcal{X}} \mu(x_t|x_{<t}) \ln \frac{\mu(x_t|x_{<t})}{\xi(x_t|x_{<t})}$

Theorem 7.2 (Total relative entropy) $D_n := \sum_{t=1}^{n} \mathbb{E}[d_t] \leq \ln w_\mu^{-1}$

$\mathbb{E}[f]$ =Expectation of f w.r.t. the *true* distribution μ, e.g.

If $f : \mathcal{X}^n \to \mathbb{R}$, then $\mathbb{E}[f] := \sum_{x_{1:n}} \mu(x_{1:n}) f(x_{1:n})$.

Proof based on dominance or universality: $\xi(x) \geq w_\mu \mu(x)$.

Proof of the Entropy Bound

$$D_n \equiv \sum_{t=1}^{n}\sum_{x_{<t}} \mu(x_{<t})\cdot d_t(x_{<t}) \overset{(a)}{=} \sum_{t=1}^{n}\sum_{x_{1:t}} \mu(x_{1:t}) \ln \frac{\mu(x_t|x_{<t})}{\xi(x_t|x_{<t})} =$$

$$\overset{(b)}{=} \sum_{x_{1:n}} \mu(x_{1:n}) \ln \prod_{t=1}^{n} \frac{\mu(x_t|x_{<t})}{\xi(x_t|x_{<t})} \overset{(c)}{=} \sum_{x_{1:n}} \mu(x_{1:n}) \ln \frac{\mu(x_{1:n})}{\xi(x_{1:n})} \overset{(d)}{\leq} \ln w_\mu^{-1}$$

(a) Insert def. of d_t and used chain rule $\mu(x_{<t})\cdot\mu(x_t|x_{<t})=\mu(x_{1:t})$.

(b) $\sum_{x_{1:t}} \mu(x_{1:t}) = \sum_{x_{1:n}} \mu(x_{1:n})$ and argument of log is independent of $x_{t+1:n}$. The t sum can now be exchanged with the $x_{1:n}$ sum and transforms to a product inside the logarithm.

(c) Use chain rule again for μ and ξ.

(d) Use dominance $\xi(x) \geq w_\mu \mu(x)$.

Predictive Convergence

Theorem 7.3 (Predictive convergence)

$$\xi(x_t|x_{<t}) \to \mu(x_t|x_{<t}) \text{ rapid w.p.1 for } t \to \infty$$

Proof: $D_\infty \equiv \sum_{t=1}^{\infty} \mathbb{E}[d_t] \leq \ln w_\mu^{-1}$ and $d_t \geq 0$

$$\Longrightarrow \quad d_t \stackrel{t\to\infty}{\longrightarrow} 0 \quad \Longleftrightarrow \quad \xi_t \to \mu_t.$$

Fazit: ξ is excellent universal predictor if unknown μ belongs to $\mathcal{M}$.

How to choose $\mathcal{M}$ and w_μ? Both as large as possible?! More later.

Sequential Decisions

A prediction is very often the basis for some decision. The decision results in an action, which itself leads to some reward or loss.

Let $\text{Loss}(x_t, y_t) \in [0,1]$ be the received loss when taking action $y_t \in \mathcal{Y}$ and $x_t \in \mathcal{X}$ is the t^{th} symbol of the sequence.

For instance, decision $\mathcal{Y} = \{\text{umbrella, sunglasses}\}$ based on weather forecasts $\mathcal{X} = \{\text{sunny, rainy}\}$.

Loss	sunny	rainy
umbrella	0.1	0.3
sunglasses	0.0	1.0

The goal is to minimize the μ-expected loss. More generally we define the Λ_ρ prediction scheme, which minimizes the ρ-expected loss:

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

Loss Bounds

- Definition: μ-expected loss when Λ_ρ predicts the t^{th} symbol:
  $\text{Loss}_t(\Lambda_\rho)(x_{<t}) := \sum_{x_t} \mu(x_t|x_{<t})\text{Loss}(x_t, y_t^{\Lambda_\rho})$
- $\text{Loss}_t(\Lambda_{\mu/\xi})$ made by the informed/universal scheme $\Lambda_{\mu/\xi}$.
  $\text{Loss}_t(\Lambda_\mu) \leq \text{Loss}_t(\Lambda)\ \forall t, \Lambda$.
- Theorem: $0 \leq \text{Loss}_t(\Lambda_\xi) - \text{Loss}_t(\Lambda_\mu) \leq \sum_{x_t} |\xi_t - \mu_t| \leq \sqrt{2d_t} \stackrel{w.p.1}{\longrightarrow} 0$
- Total $\text{Loss}_{1:n}(\Lambda_\rho) := \sum_{t=1}^{n} \mathbb{E}[\text{Loss}_t(\Lambda_\rho)]$.
- Theorem: $\sqrt{\text{Loss}_{1:n}(\Lambda_\xi)} - \sqrt{\text{Loss}_{1:n}(\Lambda_\mu)} \leq \sqrt{2D_n} \leq \sqrt{2\ln w_\mu^{-1}}$
- Corollary: If $\text{Loss}_{1:\infty}(\Lambda_\mu)$ is finite, then $\text{Loss}_{1:\infty}(\Lambda_\xi)$ is finite, and $\text{Loss}_{1:n}(\Lambda_\xi)/\text{Loss}_{1:\infty}(\Lambda_\mu) \to 1$ if $\text{Loss}_{1:\infty}(\Lambda_\mu) \to \infty$.
- Remark: Holds for any loss function $\in [0,1]$ with no assumptions (like i.i.d., Markovian, stationary, ergodic, ...) on $\mu \in \mathcal{M}$.

Proof of Instantaneous Loss Bounds

Abbreviations: $\mathcal{X}=\{1,...,N\}$, $N=|\mathcal{X}|$, $i=x_t$, $y_i=\mu(x_t|x_{<t})$, $z_i=\xi(x_t|x_{<t})$, $m=y_t^{\Lambda_\mu}$, $s=y_t^{\Lambda_\xi}$, $\ell_{xy}=\mathsf{Loss}(x,y)$.

This and definition of $y_t^{\Lambda_\mu}$ and $y_t^{\Lambda_\xi}$ and $\sum_i z_i\ell_{is}\leq\sum_i z_i\ell_{ij}\;\forall j$ implies

$$\mathsf{Loss}_t(\Lambda_\xi)-\mathsf{Loss}_t(\Lambda_\mu)\equiv\sum_i y_i\ell_{is}-\sum_i y_i\ell_{im}\overset{(a)}{\leq}\sum_i(y_i-z_i)(\ell_{is}-\ell_{im})$$
$$\leq\sum_i|y_i-z_i|\cdot|\ell_{is}-\ell_{im}|\overset{(b)}{\leq}\sum_i|y_i-z_i|\overset{(c)}{\leq}\sqrt{\sum_i y_i\ln\frac{y_i}{z_i}}\equiv\sqrt{2d_t(x_{<t})}$$

(a) We added $\sum_i z_i(\ell_{im}-\ell_{is})\geq 0$.

(b) $|\ell_{is}-\ell_{im}|\leq 1$ since $\ell\in[0,1]$.

(c) Pinsker's inequality (elementary, but not trivial)

Optimality of the Universal Predictor

- There are $\mathcal{M}$ and $\mu \in \mathcal{M}$ and weights w_μ for which the loss bounds are tight.
- The universal prior ξ is pareto-optimal, in the sense that there is no ρ with $\mathcal{F}(\nu,\rho) \leq \mathcal{F}(\nu,\xi)$ for all $\nu \in \mathcal{M}$ and strict inequality for at least one ν, where $\mathcal{F}$ is the instantaneous or total squared distance s_t, S_n, or entropy distance d_t, D_n, or general Loss_t, $\text{Loss}_{1:n}$.
- ξ is balanced pareto-optimal in the sense that by accepting a slight performance decrease in some environments one can only achieve a slight performance increase in other environments.
- Within the set of enumerable weight functions with short program, the universal weights $w_\nu = 2^{-K(\nu)}$ lead to the smallest performance bounds within an additive (to $\ln w_\mu^{-1}$) constant in all enumerable environments.

Continuous Probability Classes $\mathcal{M}$

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

$$\mathcal{M} := \{\mu_\theta : \theta \in \mathbb{R}^d\}, \quad \xi(x_{1:n}) := \int_{\mathbb{R}^d} d\theta\, w(\theta)\, \mu_\theta(x_{1:n}), \quad \int_{\mathbb{R}^d} d\theta\, w(\theta) = 1$$

We only used $\xi(x_{1:n}) \geq w_\mu \cdot \mu(x_{1:n})$
which was obtained by dropping the sum over μ.

Here, restrict integral over $\mathbb{R}^d$ to a small vicinity N_δ of θ.

For sufficiently smooth μ_θ and $w(\theta)$ we expect

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

Continuous Probability Classes $\mathcal{M}$

Average Fisher information $\bar{\jmath}_n$ measures curvature (parametric complexity) of $\ln \mu_\theta$.

$$\bar{\jmath}_n := \frac{1}{n} \sum_{x_{1:n}} \mu(x_{1:n}) \nabla_\theta \ln \mu_\theta(x_{1:n}) \nabla_\theta^T \ln \mu_\theta(x_{1:n})_{|\theta=\theta_0}$$

Under weak regularity conditions on $\bar{\jmath}_n$ on can prove:

Theorem 7.4 (Continuous entropy bound)

$$D_n \leq \ln w_\mu^{-1} + \tfrac{d}{2} \ln \tfrac{n}{2\pi} + \tfrac{1}{2} \ln \det \bar{\jmath}_n + o(1)$$

i.e. D_n grows only logarithmically with n.

E.g. $\bar{\jmath}_n = O(1)$ for the practically very important class of stationary (k^{th}-order) finite-state Markov processes ($k = 0$ is i.i.d.).

Bayesian Sequence Prediction: Summary

- General sequence prediction: Use known (subj.) Bayes mixture $\xi = \sum_{\nu \in \mathcal{M}} w_\nu \nu$ in place of unknown (obj.) true distribution μ.

- Bound on the relative entropy between ξ and μ.

$\Rightarrow$ posterior of ξ converges rapidly to the true posterior μ.

- ξ is also optimal in a decision-theoretic sense w.r.t. any bounded loss function.

- No structural assumptions on $\mathcal{M}$ and $\nu \in \mathcal{M}$.

Literature

[Hut05] M. Hutter. *Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability*. Springer, Berlin, 2005. http://www.hutter1.net/ai/uaibook.htm.

[Jef83] R. C. Jeffrey. *The Logic of Decision*. University of Chicago Press, Chicago, IL, 2nd edition, 1983.

[Fer67] T. S. Ferguson. *Mathematical Statistics: A Decision Theoretic Approach*. Academic Press, New York, 3rd edition, 1967.

[DeG70] M. H. DeGroot. *Optimal Statistical Decisions*. McGraw-Hill, New York, 1970.