Predictive Hypothesis Identification

ANU

Marcus Hutter

Canberra, ACT, 0200, Australia

http://www.hutter1.net/

NICTA

Summary

- If prediction is the goal, but full Bayes not feasible, one should Identify (estimate/test/select) the Hypothesis (parameter/model/ interval) that Predicts best (PHI).
- What best is can depend on benchmark (Loss, $\widetilde{\text{Loss}}$), distance function (d), how long we use the model (m), compared to how much data we have at hand (n).
- The new principle (PHI) possesses many desirable properties.
- In particular, PHI can properly deal with nested hypotheses, and nicely justifies, reconciles, and blends MAP and ML for $m \gg n$, MDL for $m \approx n$, and SMF for $n \gg m$.

Background & Idea & Principle

The Problem - Information Summarization

- Given: Data $D \equiv (x_1, ..., x_n) \in \mathcal{X}^n$ (any $\mathcal{X}$) sampled from distribution $p(D|\theta)$ with unknown $\theta \in \Omega$.
- Likelihood function $p(D|\theta)$ or posterior $p(\theta|D) \propto p(D|\theta)p(\theta)$ contain **all** statistical information about the sample D.
- Information summary or simplification of $p(D|\theta)$ is needed: (comprehensibility, communication, storage, computational efficiency, mathematical tractability, etc.).
- Regimes: - parameter estimation,
  - hypothesis testing,
  - model (complexity) selection.

Ways to Summarize the Posterior by

- a single point $\Theta = \{\theta\}$ (ML or MAP or mean or stochastic or ...),
- a convex set $\Theta \subseteq \Omega$ (e.g. confidence or credible interval),
- a finite set of points $\Theta = \{\theta_1, ..., \theta_l\}$ (mixture models)
- a sample of points (particle filtering),
- the mean and covariance matrix (Gaussian approximation),
- more general density estimation,
- in a few other ways.

I concentrate on set estimation, which includes (multiple) point estimation and hypothesis testing as special cases.

Call it: Hypothesis Identification.

Desirable Properties

of any hypothesis identification principle

- leads to **good predictions** (that's what models are ultimately for),
- be broadly applicable,
- be analytically and computationally tractable,
- be defined and works also for non-i.i.d. and non-stationary data,
- be reparametrization and representation invariant,
- works for simple and composite hypotheses,
- works for classes containing **nested** and overlapping **hypotheses**,
- works in the estimation, testing, and model selection regime,
- reduces in special cases (approximately) to existing other methods.

Here we concentrate on the first item, and will show that the resulting principle nicely satisfies many of the other items.

The Main Idea

- Machine learning primarily cares about predictive performance.
- We address the problem head on.
- Goal: Predict m future obs. $\boldsymbol{x} \equiv (x_{n+1}, ..., x_{n+m}) \in \mathcal{X}^m$ well.
- If θ_0 is true parameter, then $p(\boldsymbol{x}|\theta_0)$ is obviously the best prediction.
- If θ_0 unknown, then the Bayesian predictive distribution $\boldsymbol{p(x|D) = \int p(x|\theta)p(\theta|D)d\theta = p(D,x)/p(D)}$ is best.
- Approx. full Bayes by predicting with hypothesis $H = \{\theta \in \Theta\}$, i.e.
- Use (composite) likelihood $\boldsymbol{p(x|\Theta) = \frac{1}{P[\Theta]}\int_\Theta p(x|\theta)p(\theta)d\theta}$ for prediction.
- The closer $p(\boldsymbol{x}|\Theta)$ to $p(\boldsymbol{x}|\theta_0)$ or $p(\boldsymbol{x}|D)$ the better H's prediction.
- Measure closeness with some distance function $d(\cdot,\cdot)$.
- Since $\boldsymbol{x}$ and θ_0 are unknown, we must sum or average over them.

Predictive Hypothesis Identification (PHI)

Definition 1 (Predictive Loss) The predictive Loss/ $\widetilde{\text{Loss}}$ of Θ given D based on distance d for m future observations is

$$\text{Loss}_d^m(\Theta, D) := \int d(p(\boldsymbol{x}|\Theta), p(\boldsymbol{x}|D))d\boldsymbol{x},$$
$$\widetilde{\text{Loss}}_d^m(\Theta, D) := \iint d(p(\boldsymbol{x}|\Theta), p(\boldsymbol{x}|\theta))\, p(\theta|D)d\boldsymbol{x}\, d\theta$$

Definition 2 (PHI) The best ($\widetilde{\text{best}}$) predictive hypothesis in hypothesis class $\mathcal{H}$ given D is

$$\hat\Theta_d^m := \arg\min_{\Theta\in\mathcal{H}} \text{Loss}_d^m(\Theta, D)$$
$$(\,\tilde\Theta_d^m := \arg\min_{\Theta\in\mathcal{H}} \widetilde{\text{Loss}}_d^m(\Theta, D)\,)$$

Use $p(\boldsymbol{x}|\hat\Theta_d^m)$ $(p(\boldsymbol{x}|\tilde\Theta_d^m))$ for prediction. That's it!

Example & Exact Properties

A Very Simple Example

- Consider a sequence of n bits from an unknown source. Assume we have observed $n_0 = \#0s = \#1s = n_1$.
- We want to test whether the unknown source is a fair coin: "fair" ($H_f = \{\theta = \frac{1}{2}\}$) versus "don't know" ($H_v = \{\theta \in [0;1]\}$) $\mathcal{H} = \{H_f, H_v\}$, $\theta \in \Omega = [0;1] =$ bias.
- Classical tests involve the choice of some confidence level α.
- Problem 1: The answer depends on the confidence level.
- Problem 2: The answer should depend on the purpose.
- A smart customer wants to predict m further bits. We can tell him 1 bit of information: "fair" or "don't know".
- $m = 1$: The answer doesn't matter, since in both cases customer will predict 50% by symmetry.
- $m \ll n$: We should use our past knowledge and tell him "fair".
- $m \gg n$: We should ignore our past knowledge & tell "don't know", since customer can make better judgement himself, since he will have much more data.
- Evaluating PHI on this simple Bernoulli example $p(D|\theta) = \theta^{n_1}(1-\theta)^{n_0}$ exactly leads to this conclusion!
- Maximum A Posteriori (MAP): $P[\Omega|D] = 1 \geq P[\Theta|D]\, \forall\Theta \implies$ $\Theta^{\text{MAP}} := \arg\max_{\Theta\in\mathcal{H}} \text{P}[\Theta|D] = \Omega = H_v =$"don't know", however strong the evidence for a fair coin! MAP is risk averse finding a likely true model of low pred. power.
- Maximum Likelihood (ML): $p(D|H_f) \geq p(D|\Theta)\, \forall\Theta \implies$ $\Theta^{\text{ML}} := \arg\max_{\Theta\in\mathcal{H}} p(D|\Theta) = \{\frac{1}{2}\} = H_f =$"fair", however weak the evidence for a fair coin! Composite ML risks an (over)precise prediction.
- Fazit: Although MAP and ML give identical answers for uniform prior on simple hypotheses, their naive extension to composite hypotheses is diametral.
- Intuition/PHI/MAP/ML conclusions hold in general.

Some Popular Distance Functions

(f)	f-divergence	$d(p,q) = f(p/q)q$ for convex f with $f(1) = 0$	
(1)	absolute deviation:	$d(p,q) = \|p-q\|$,	$f(t) = \|t-1\|$
(h)	Hellinger distance:	$d(p,q) = (\sqrt{p}-\sqrt{q})^2$,	$f(t) = (\sqrt{t}-1)^2$
(2)	squared distance:	$d(p,q) = (p-q)^2$,	no f
(c)	chi-square distance:	$d(p,q) = (p-q)^2/q$,	$f(t) = (t-1)^2$
(k)	KL-divergence:	$d(p,q) = p\ln(p/q)$,	$f(t) = t\ln t$
(r)	reverse KL-div.:	$d(p,q) = q\ln(q/p)$,	$f(t) = -\ln t$

The f-divergences are particularly interesting, since they contain most of the standard distances and make Loss representation invariant.

Exact Properties of PHI

Theorem 3 (Invariance of Loss) $\text{Loss}_d^m(\Theta, D)$ and $\widetilde{\text{Loss}}_d^m(\Theta, D)$ are invariant under reparametrization $\theta \leadsto \vartheta = g(\theta)$ of Ω. If distance d is an f-divergence, then they are also independent of the representation $x_i \leadsto y_i = h(x_i)$ of the observation space $\mathcal{X}$.

Theorem 4 (PHI for sufficient statistic) Let $\boldsymbol{t = T(x)}$ be a sufficient statistic for θ. Then $\text{Loss}_f^m(\Theta, D) = \int d(p(t|\Theta), p(t|D))dt$ and $\widetilde{\text{Loss}}_f^m(\Theta, D) = \int d(p(t|\Theta), p(t|\theta))p(\theta|D)dtd\theta$, i.e. $p(\boldsymbol{x}|\cdot)$ can be replaced by the probability density $p(t|\cdot)$ of t.

Theorem 5 (Equivalence of $\text{PHI}_{2|r}^m$ and $\widetilde{\text{PHI}}_{2|r}^m$) For square distance ($d\widehat{=}2$) and RKL distance ($d\widehat{=}r$), $\text{Loss}_d^m(\Theta, D)$ differs from $\widetilde{\text{Loss}}_d^m(\Theta, D)$ only by an additive constant $\text{c}_d^m(D)$ independent of Θ, hence PHI and $\widetilde{\text{PHI}}$ select the same hypotheses $\hat\Theta_2^m = \tilde\Theta_2^m$ and $\hat\Theta_r^m = \tilde\Theta_r^m$.

Bernoulli Example

$$p(t|\theta) = \binom{m}{t}\theta^t(1-\theta)^{m-t}, \qquad t = m_1 = \#1s = \text{suff.stat.}$$

For RKL-distance and point hypotheses, Theorems 4 and 5 now yield

$$\tilde\theta_r = \hat\theta_r = \arg\min_\theta \text{Loss}_r^m(\theta|D) = \arg\min_\theta \sum_{t=1}^m p(t|D)\ln\frac{p(t|D)}{p(t|\theta)} =$$

$$... = \tfrac{1}{m}\mathbf{E}[t|D] = \frac{n_1+1}{n+2} = \text{Laplace rule}$$

Asymptotic Properties

PHI for Large m (Still Exact in n)

- $I_1(\theta) := -\int(\partial\partial^\top \ln p(x|\theta))p(x|\theta)dx =$ Fisher information matrix.
- $J := \int\sqrt{\det I_1(\theta)}d\theta =$ intrinsic size of Ω.
- $p_J(\theta) := \sqrt{\det I_1(\theta)}/J =$ Jeffrey's prior is a popular reparametrization invariant (objective) reference prior.

Point Estimation

Theorem 6 ($\widetilde{\text{Loss}}_h^m(\theta, D)$ for large m) Under some differentiability assumptions, for point estimation, the predictive Hellinger loss for large m is

$$\widetilde{\text{Loss}}_h^m(\theta, D) = 2 - 2\left(\frac{8\pi}{m}\right)^{d/2}\frac{p(\theta|D)}{\sqrt{\det I_1(\theta)}}[1 + O(m^{-1/2})]$$
$$\stackrel{J}{=} 2 - 2\left(\frac{8\pi}{m}\right)^{d/2}\frac{p(D|\theta)}{Jp(D)}[1 + O(m^{-1/2})]$$

where the first expression holds for any continuous prior density and the second expression ($\stackrel{J}{=}$) holds for Jeffrey's prior.

PHI = IMAP $\stackrel{J}{=}$ ML for $m \gg n$

Minimizing $\widetilde{\text{Loss}}_h^\infty$ is equivalent to a reparametrization invariant variation of MAP:

$$\tilde\theta_h^\infty = \theta^{\text{IMAP}} := \arg\max_\theta \frac{p(\theta|D)}{\sqrt{\det I_1(\theta)}} \stackrel{J}{=} \arg\max_\theta p(D|\theta) \equiv \theta^{\text{ML}}$$

This is a nice reconciliation of MAP and ML: An "improved" MAP leads for Jeffrey's prior back to "simple" ML.

PHI $\approx$ MDL for $m \approx n$

We can also relate PHI to the Minimum Description Length (MDL) principle by taking the logarithm of the second expression in Theorem 6:

$$\tilde\theta_h^\infty \stackrel{J}{=} \arg\min_\theta\{-\log p(D|\theta) + \tfrac{d}{2}\log\tfrac{m}{8\pi} + J\}$$

For $m = 4n$ this is the classical (large n approximation of) MDL.

Loss for Large m and Composite Θ

Theorem 7 ($\widetilde{\text{Loss}}_h^m(\Theta, D)$ for large m) Under some differentiability assumptions, for composite Θ, the predictive Hellinger loss for large m is

$$\widetilde{\text{Loss}}_h^m(\Theta, D) \stackrel{J}{=} 2 - 2\left(\frac{8\pi}{m}\right)^{d/4}\sqrt{\frac{p(D|\Theta)\text{P}[\Theta|D]}{J\text{P}[D]}} + o(m^{-d/4})$$

MAP Meets ML Half Way

- The expression is proportional to the geometric average of the posterior and the composite likelihood.
- For large Θ, the likelihood gets small, since the average involves many wrong models.
- For small Θ, posterior $\propto$ volume of Θ, hence tends to zero.
- The product is maximal for $|\Theta| \sim n^{-d/2}$ (which makes sense).

Finding $\tilde\Theta_h^\infty$ Explicitly

Contrary to MAP and ML, an unrestricted maximization of ML$\times$MAP over all measurable $\Theta \subseteq \Omega$ makes sense, and can be reduced to a one-dimensional maximization.

Theorem 8 (Finding $\tilde\Theta_h^\infty$ exactly) Let $\Theta_\gamma := \{\theta : p(D|\theta) \geq \gamma\}$ be the γ-level set of $p(D|\theta)$. If $\text{P}[\Theta_\gamma]$ is continuous in γ, then

$$\tilde\Theta_h^\infty = \arg\max_\Theta \frac{\text{P}[\Theta|D]}{\sqrt{\text{P}[\Theta]}} = \arg\max_{\Theta_\gamma:\gamma\geq 0}\frac{\text{P}[\Theta_\gamma|D]}{\sqrt{\text{P}[\Theta_\gamma]}}$$

Theorem 9 (Finding $\tilde\Theta_h^\infty$ for Large n ($m \gg n \gg 1$)) $\tilde\Theta_h^\infty = \{\theta : (\theta-\bar\theta)^\top I_1(\bar\theta)(\theta-\bar\theta) \leq \tilde r^2\} = \text{Ellipsoid}, \quad \tilde r \approx \sqrt{d/n}$

Loss_h^m: Similar to (asymptotic) expressions of $\widetilde{\text{Loss}}_h^m$.

Large Sample Approximations

PHI for large sample sizes $n \gg m$. For simplicity $\theta \in I\!R$.

- A classical approximation of $p(\theta|D)$ is by a Gaussian with same mean and variance.
- Generalization to Sequential Moment Fitting (SMF): Fit first k (central) moments $\bar\theta^A \equiv \mu_1^A := \mathbf{E}[\theta|A]$ and $\mu_k^A := \mathbf{E}[(\theta-\bar\theta^A)^2|A] \quad (k \geq 2)$
- Moments μ_k^D are known and can in principle be computed.

Theorem 10 (PHI for large n by SMF) If $\Theta^* \in \mathcal{H}$ is chosen such that $\mu_i^{\Theta^*} = \mu_i^D$ for $i = 1, ..., k$, then under some technical conditions,

$$\text{Loss}_f^m(\Theta^*, D) = O(n^{-k/2})$$

- Normally, no $\Theta \in \mathcal{H}$ has better loss order, therefore $\hat\Theta_f^m \simeq \Theta^*$.
- $\hat\Theta \equiv \hat\Theta_f^m$ neither depends on m, nor on the chosen distance f.

Large Sample Applications

- $\Theta = \{\theta_1, ..., \theta_l\}$ unrestricted $\implies$ $k = l$ moments can be fit.
- For interval est. $\mathcal{H} = \{[a;b] : a, b \in I\!R, a \leq b\}$ and uniform prior, we have $\bar\theta^{[a;b]} = \frac{1}{2}(a+b)$ and $\mu_2^{[a;b]} = \frac{1}{12}(b-a)^2$ Determine a and b such that $\bar\theta^{[a;b]} = \bar\theta^D$ and $\mu_2^{[a;b]} = \mu_2^D$. $\implies$ $k = 2$ and $\hat\Theta = [\bar\theta^D - \sqrt{3}\mu_2^D\,;\, \bar\theta^D + \sqrt{3}\mu_2^D]$.
- In higher dimensions, common choices of $\mathcal{H}$ are convex sets, ellipsoids, and hypercubes.

Paper: http://arxiv.org/abs/0809.1270