Tutorial on Universal Prediction, Day 2, afternoon

CWI Peter Grünwald

Universiteit Leiden

Centrum Wiskunde & Informatica – Amsterdam
Mathematisch Instituut – Universiteit Leiden

Menu

1. Generalizing Bayes, making it work for general loss
2. Final Remarks about 0/1-loss
3. Log-loss prediction & Kolmogorov Complexity
4. "MDL" & Occam's Razor
5. Philosophy & Relevance

The Clue

- We then have for all n, y^n, θ :

$$\text{log-loss}(y^n, P_{\text{Bayes},\eta}) \leq \text{log-loss}(y^n, P_{\theta,\eta}) - \log W(\theta)$$

$$\text{log-loss}(y^n, \mathbf{A}) \leq \text{log-loss}(y^n, \theta) - \log W(\theta) + \Delta^*_\eta(\mathbf{A})$$

$$\text{loss}(y^n, \mathbf{A}) \leq \text{loss}(y^n, \theta) + \frac{-\log W(\theta)}{\eta} + \frac{\Delta^*_\eta(\mathbf{A})}{\eta}$$

The Clue

- We then have for all n, y^n, θ :

$$\text{log-loss}(y^n, P_{\text{Bayes},\eta}) \leq \text{log-loss}(y^n, P_{\theta,\eta}) - \log W(\theta)$$

$$\text{log-loss}(y^n, \mathbf{A}) \leq \text{log-loss}(y^n, \theta) - \log W(\theta) + \Delta^*_\eta(\mathbf{A})$$

$$\text{loss}(y^n, \mathbf{A}) \leq \text{loss}(y^n, \theta) + \frac{-\log W(\theta)}{\eta} + \frac{\Delta^*_\eta(\mathbf{A})}{\eta}$$

Vovk '90: for so-called "mixable" loss fns, there exists an A (the aggregating algorithm) such that for some $\eta > 0$, we are guaranteed $\Delta^*_\eta(\mathbf{A}) \leq 0$ (hence name mixability gap)

The Clue

- We then have for all n, y^n, θ :

$$\text{log-loss}(y^n, P_{\text{Bayes},\eta}) \leq \text{log-loss}(y^n, P_{\theta,\eta}) - \log W(\theta)$$

$$\text{log-loss}(y^n, \mathbf{A}) \leq \text{log-loss}(y^n, \theta) - \log W(\theta) + \Delta^*_\eta(\mathbf{A})$$

$$\text{loss}(y^n, \mathbf{A}) \leq \text{loss}(y^n, \theta) + \frac{-\log W(\theta)}{\eta} + \frac{\Delta^*_\eta(\mathbf{A})}{\eta}$$

Vovk '90: for so-called "mixable" loss fns, there exists an A (the aggregating algorithm) such that for some $\eta > 0$, we are guaranteed $\Delta^*_\eta(\mathbf{A}) \leq 0$ (hence name mixability gap)

Example: squared loss
$\text{loss}(y, a) = (y - a)^2$ is ½- mixable if $\mathcal{Y} = [-1, 1]$

Mixable Loss Functions

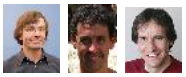
- All 'strictly convex' loss functions with bounded range are mixable for finite $\eta > 0$
- So we can still run generalized Bayes and get regret bounds that are still of order $O(-\log W(\theta))$
- Usually $\eta < 1$:
  prior becomes more, data less important
- Modifying Bayes may seem like 'hack' (relation to Bayes' theorem is lost) but: resulting procedure still minimax optimal up to a constant (and in practice preferable over minimax algorithm)
- Something deep going on...

The Clue, now for nonmixable loss

- We then have for all n, y^n, θ :

$$\text{loss}(y^n, \mathbf{A}) \leq \text{loss}(y^n, \theta) + \frac{-\log W(\theta)}{\eta} + \frac{\Delta^*_\eta(\mathbf{A})}{\eta}$$

0/1-loss is nonmixable
(related to Bayes predictions being 'pushed to boundary')



The Clue

- We then have for all n, y^n, θ :

$$\text{loss}(y^n, \mathbf{A}) \leq \text{loss}(y^n, \theta) + \frac{-\log W(\theta)}{\eta} + \frac{\Delta^*_\eta(\mathbf{A})}{\eta}$$

Hedge-style algorithms: for bounded nonmixable losses functions: there still exist **A** such that $\Delta^*_\eta(\mathbf{A}) \leq \eta^2 \cdot n$
For uniform prior over K predictors, optimizing over η gives $\eta = \tilde{O}\left(\sqrt{\frac{\log K}{n}}\right)$
We get regret bound (and actual regret) $O\left(\sqrt{n \cdot (\log K)}\right)$
(Warmuth, Freund, Schapire)

The Clue

- We then have for all n, y^n, θ :

$$\text{loss}(y^n, \mathbf{A}) \leq \text{loss}(y^n, \theta) + \frac{-\log W(\theta)}{\eta} + \frac{\Delta^*_\eta(\mathbf{A})}{\eta}$$

Hedge-style algorithms: for bounded nonmixable losses functions: there still exist **A** such that $\Delta^*_\eta(\mathbf{A}) \leq \eta^2 \cdot n$
For uniform prior over K predictors, optimizing over η gives $\eta = \tilde{O}\left(\sqrt{\frac{\log K}{n}}\right)$
We get regret bound (and actual regret) $O\left(\sqrt{n \cdot (\log K)}\right)$
(Warmuth, Freund, Schapire)

The Hedge Algorithm

- Basic Algorithm requires knowledge of 'horizon' n : at every time point $t < n$, you use generalized posterior with $\eta = \tilde{O}\left(\sqrt{\frac{\log K}{n}}\right)$
- If n unknown, can still achieve same bound by decreasing learning rate dynamically: at each time t, you use posterior weights you get if you had used $\eta = \tilde{O}\left(\sqrt{\frac{\log K}{t}}\right)$ from time 1 to t

"The older you get, the less attention you pay to all the experiences you had in life!"

Menu

1. Generalizing Bayes, making it work for general loss
2. Final Remarks about 0/1-loss

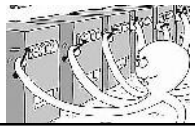
3. Log-loss prediction and Kolmogorov Complexity
4. "MDL" & Occam's Razor
5. Philosophy & Relevance

Varying the Setting

- Predicting better than the Best Expert
  - As good as the best convex combination
  - As good as the best sequence of experts
  - Applied to Prediction of Electricity Consumption in Greater Paris Region by Electricité de France (Devaine, Goude, Stoltz 2012)

Varying the Setting

- Predicting better than the Best Expert
  - As good as the best convex combination
  - As good as the best sequence of experts
  - Applied to Prediction of Electricity Consumption in Greater Paris Region by Electricité de France (Devaine, Goude, Stoltz 2012)
- **Multi-Armed Bandit** & other Limited Feedback types



Google Ads

Major Open Problem(s)

Learning the Learning Rate

Major Open Problem(s)

Learning the Learning Rate

Easy Data

For log-loss, worst-case and average case regret comparable. For 0/1-loss, there is wide gap...

Menu

1. Generalizing Bayes
2. Final Remarks about 0/1-loss
3. Log-loss prediction & Kolmogorov Complexity
4. "MDL" & Occam's Razor
5. Philosophy & Relevance

Universal Log-Loss Prediction and Kolmogorov Complexity

- Let $P_1, P_2, \ldots$ be an effective enumeration of the lower-semi-computable semi-measures on $\mathcal{Y}^\infty$
- The KM version of Kolmogorov complexity can be written as
$$KM(y^n) = -\log \sum_\theta W(\theta) P_\theta(y^n)$$
for a special prior W on $\{1, 2, \ldots\}$
- We have $\sup_{y^n \in \{0,1\}^n} |KM(y^n) - K(y^n)| = O(\log n)$

Universal Log-Loss Prediction and Kolmogorov Complexity

- The KM version of Kolmogorov complexity can be written as $KM(y^n) = -\log \sum_\theta W(\theta) P_\theta(y^n)$
$$= -\log P_{\text{Bayes}}(y^n)$$
- By definition, the log-loss of Solomonoff Prediction (Li and Vitányi, '07) on sequence y^n given by
$$-\log P_{\text{Bayes}}(y^n)$$
- Solomonoff Prediction = Universal log-loss Prediction with the lower-semi-computable semimeasures as 'candidate set'

"Differences"

- **Method same, analysis different:** Solomonoff's converge **theorem** based on assumption that data are sampled from a computable distribution. We avoid that assumption
- **We treat predictors as oracles:** their predictions are just available to us, we don't need to know how they got them
  - They may use side-information unknown to us
  - They may be 'uncomputable', or, more realistically, have exceedingly complex inner workings. Still, if we trust them, we may assign a high prior to them
  - In Solomonoff's theory, such predictors would be considered 'complex' and automatically be assigned small priors

Menu

1. Generalizing Bayes
2. Final Remarks about 0/1-loss
3. Log-loss prediction & Kolmogorov Complexity
4. "MDL" & Occam's Razor
5. Philosophy & Relevance

Minimum Description Length

- **Goal** (or **Means**): individual sequence log-loss prediction, but our set of candidate predictors is a (potentially huge, but usable) statistical model

Minimum Description Length

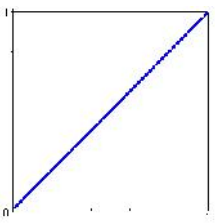
Rissanen '78, '89, '96, G. '07

- **Goal** (or **Means**): individual sequence log-loss prediction, but our set of candidate predictors is a (potentially huge, but usable) statistical model
- If model is still 'parametric' we can define the **minimax optimal prediction** strategy
  - normalized maximum likelihood, Shtarkov '88, Rissanen '96
  - This is what 'modern' MDL approaches are usually based on, rather than Bayesian prediction

Minimum Description Length

- **Goal** (or **Means**): individual sequence log-loss prediction, but our set of candidate predictors is a (potentially huge, but usable) statistical model
- If model is still 'parametric' we can define the minimax optimal prediction strategy
  - normalized maximum likelihood, Shtarkov '88, Rissanen '96
  - This is what 'modern' MDL approaches are usually based on, rather than Bayesian prediction
- One (somewhat crude) way to approximate minimax is to discretize the model and use Bayes on the resulting set of predictors with a **uniform prior**

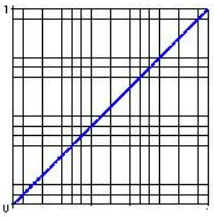
Nested Models



$\mathcal{M}_1$ Class (model) of i.i.d. Bernoulli distributions

$\mathcal{M}_2$ Class (model) of First-Order Markov Chains

Nested Models



$\ddot{\mathcal{M}}_1$ **Discretized** i.i.d. Bernoulli distributions

$\ddot{\mathcal{M}}_2$ **Discretized** First-Order Markov Chains

Prediction relative to simple model is easier

- At each fixed level of discretization, (tight) bound on Bayesian regret relative to $\ddot{\mathcal{M}}_1$:
  $\log |\ddot{\mathcal{M}}_1|$ ($= \log 20$)
  and relative to $\ddot{\mathcal{M}}_2$:
  $\log |\ddot{\mathcal{M}}_2| = \log |\ddot{\mathcal{M}}_1|^2$ ($= 2 \log 20$)
- Smaller regret relative to simple model
- On the other hand, best predictor-with-hindsight in complex model will usually have smaller cumulative loss than best predictor-with-hindsight in simple model
- Which model (set) should we pick?

Meta-Priors

- We solve conundrum by using universal prediction idea at meta-level: we put uniform meta-prior on $\ddot{\mathcal{M}}_1, \ddot{\mathcal{M}}_2$

For $P_\theta \in \ddot{\mathcal{M}}_1$: $W(\theta) = W(\theta \mid \ddot{\mathcal{M}}_1) W(\ddot{\mathcal{M}}_1) = \frac{1}{20} \cdot \frac{1}{2}$

For $P_\theta \in \ddot{\mathcal{M}}_2$: $W(\theta) = W(\theta \mid \ddot{\mathcal{M}}_1) W(\ddot{\mathcal{M}}_1) = \frac{1}{400} \cdot \frac{1}{2}$

- Using the Bayesian algorithm we get simultaneous regret bounds:

For all $P \in \ddot{\mathcal{M}}_1$: $\text{loss}(y^n, P_{\text{Meta-Bayes}}) - \text{loss}(y^n, P) \leq \log |\ddot{\mathcal{M}}_1| + 1$

For all $P \in \ddot{\mathcal{M}}_2$: $\text{loss}(y^n, P_{\text{Meta-Bayes}}) - \text{loss}(y^n, P) \leq \log |\ddot{\mathcal{M}}_2| + 1$

Meta-Priors

- We solve conundrum by using universal prediction idea at meta-level: we put uniform meta-prior on $\ddot{\mathcal{M}}_1, \ddot{\mathcal{M}}_2$

For $P_\theta \in \ddot{\mathcal{M}}_1$: $W(\theta) = W(\theta \mid \ddot{\mathcal{M}}_1) W(\ddot{\mathcal{M}}_1) = \frac{1}{20} \cdot \frac{1}{2}$

For $P_\theta \in \ddot{\mathcal{M}}_2$: $W(\theta) = W(\theta \mid \ddot{\mathcal{M}}_1) W(\ddot{\mathcal{M}}_1) = \frac{1}{400} \cdot \frac{1}{2}$

- Using the Bayesian algorithm we get regret bounds:

For all $P \in \ddot{\mathcal{M}}_1$: $\text{loss}(y^n, P_{\text{Meta-Bayes}}) - \text{loss}(y^n, P) \leq \log |\ddot{\mathcal{M}}_1| + 1$

For all $P \in \ddot{\mathcal{M}}_2$: $\text{loss}(y^n, P_{\text{Meta-Bayes}}) - \text{loss}(y^n, P) \leq \log |\ddot{\mathcal{M}}_2| + 1$

- Prior 'prefers' (puts most mass on) elements of simple model. In practice, posterior 'prefers' simple model at small samples, switches to complex later

Meta-Priors and MDL

- We solve conundrum by using universal prediction idea at meta-level: we put uniform meta-prior on $\ddot{\mathcal{M}}_1, \ddot{\mathcal{M}}_2$

For $P_\theta \in \ddot{\mathcal{M}}_1$: $W(\theta) = W(\theta \mid \ddot{\mathcal{M}}_1) W(\ddot{\mathcal{M}}_1) = \frac{1}{20} \cdot \frac{1}{2}$

For $P_\theta \in \ddot{\mathcal{M}}_2$: $W(\theta) = W(\theta \mid \ddot{\mathcal{M}}_1) W(\ddot{\mathcal{M}}_1) = \frac{1}{400} \cdot \frac{1}{2}$

- MDL is largely based on using universal predictors with such meta-properties. Sometimes directly used for prediction, sometimes indirectly for other forms of inductive inference (long story...)

Ockham+Luckiness

- Ockham+Luckiness Principle: given a large structured 'model' $\mathcal{M}$ you can (repeatedly!) single out a small, less complex subset $\mathcal{M}_{\text{simple}}$ and construct a meta-prior such that for all y^n

$\forall P \in \mathcal{M}_{\text{simple}}$: $\text{regret}(P^{\text{meta}}_{\text{Bayes}}, P, y^n) \leq \text{regret}(P_{\text{Bayes}} \mid \mathcal{M}_{\text{simple}}, P, y^n) + 1$

$\forall P \in \mathcal{M}$: $\text{regret}(P^{\text{meta}}_{\text{Bayes}}, P, y^n) \leq \text{regret}(P_{\text{Bayes}} \mid \mathcal{M}, P, y^n) + 1$

- Rationale:
  - **If you're lucky (some element of the simple model already predicts reasonably well), you'll do much better than with the original prior on the large model at small samples**
  - **If you're *not* lucky, you will hardly do worse than with the original prior on the large model**

Ockham+Luckiness

$\log n$

- Ockham+Luckiness Principle: given a large structured 'model' $\mathcal{M}$ you can (repeatedly!) single out a small, less complex subset $\mathcal{M}_{\text{simple}}$ and construct a meta-prior such that for all y^n

$$\forall P \in \mathcal{M}_{\text{simple}}: \ \text{regret}(P^{\text{meta}}_{\text{Bayes}}, P, y^n) \leq \text{regret}(P_{\text{Bayes}} \mid \mathcal{M}_{\text{simple}}, P, y^n) + 1$$

$$\forall P \in \mathcal{M}: \ \text{regret}(P^{\text{meta}}_{\text{Bayes}}, P, y^n) \leq \text{regret}(P_{\text{Bayes}} \mid \mathcal{M}, P, y^n) + 1$$

$2 \log n$

- Rationale:
  - **If you're lucky (some element of the simple model already predicts reasonably well), you'll do much better than with the original prior on the large model**
  - **If you're *not* lucky, you will hardly do worse than with the original prior on the large model**

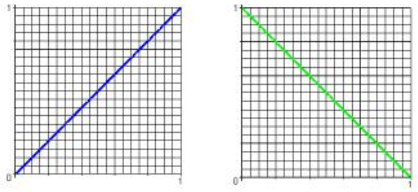
Ockham+Luckiness

Ockham-Luckiness Principle has

- Objective Component:
  1. Given a large model, it is generally a good idea to use a prior that 'prefers' simple sub-models. If the simple submodel already contains reasonably good predictor, you learn fast; if it doesn't, hardly anything is lost
  2. We can precisely quantify what we mean by 'learning fast', 'simple submodel' etc.
- Subjective Component:
  you can decide on what "simple submodel" to choose yourself.

Subjective Aspect

Given $\ddot{\mathcal{M}}_2$, how should we construct $\ddot{\mathcal{M}}_1$?



Why is Bernoulli (left) more natural candidate for $\ddot{\mathcal{M}}_1$ than 'reverse Bernoulli' (right) or any other 1-dimensional submodel of $\ddot{\mathcal{M}}_2$, for that matter?

Objective Aspect

- There really is some objectivity here in that one cannot reverse the role of the large and the simple model in the analysis!
- We designed a prior that 'prefers' the simple model. A prior that would 'prefer' the complex model will **never** do much better than the original prior on the complex model, and sometimes a bit **worse** → **that would be an objectively stupid thing to do...**

An Objective but Weak form of 'Ockham's Razor'

- MDL can be thought of as formalizing particular form of Ockham's razor that can be 'objectively' justified:
  - 'complexity' of a model measured by minimax log-loss regret ('idealized codelength')
  - Preferring simple submodels at small sample sizes leads to predictions that can hardly get worse but much better if you're lucky
  - Just about all modern applications of statistics or machine learning with large models use ideas like this in one form or another!

Contrast to 'Ockham's Razor' in Solomonoff-Hutter Theory

- MDL can be thought of as formalizing a particular 'objectively justified' form of Occam's razor
  - 'complexity' of a model measured by minimax log-loss regret
- However there remain subjective aspects (what is the 'right' submodel) and, most importantly:
- MDL makes no claims about the inherent complexity of individual hypotheses (distributions/predictors) only about **sets** of them (2nd vs.1st degree polynomials)

Contrast to 'Ockham's Razor' in Solomonoff-Hutter Theory

- MDL can be thought of as formalizing a particular 'objectively justified' form of Occam's razor
  - 'complexity' of a model measured by minimax log-loss regret
- However there remain subjective aspects (what is the 'right' submodel) and, most importantly:
- MDL makes no claims about the inherent complexity of individual hypotheses (distributions/predictors) only about **sets** of them (2nd vs.1st degree polynomials)
- Kolmogorov-complexity based approaches to induction such as Solomonoff's do define inherent complexity of individual hypotheses and 'prefer' 'simple' (low Kolmogorov complexity) ones

Menu

1. Generalizing Bayes
2. Final Remarks about 0/1-loss
3. Log-loss prediction & Kolmogorov Complexity
4. "MDL" & Occam's Razor
5. Philosophy & Relevance

Bayes vs. Worst-Case

- Bayesian statistics in its usual interpretation relies on heavy assumptions about the world:
  - " $W(\theta)$ is your prior belief that the world is in state θ "
  - Savage/De Finetti/Cox justification of Bayes: much more subtle, but assumptions still strong: decision maker has a complete preference order on acts (Why should she?)

Bayes vs. Worst-Case

- Bayesian statistics in its usual interpretation relies on heavy assumptions about the world:
  - " $W(\theta)$ is your prior belief that the world is in state θ "
- Our reinterpretation justifies Bayesian approaches in *some* situations with hardly *any* assumptions
  - With prior W, your predictions will never be worse than the predictions made by any θ up to to luckiness term

$$-\log W(\theta)$$

no matter what data you will observe

Bayes vs. Worst-Case

- Bayesian statistics in its usual interpretation relies on heavy assumptions about the world:
  - " $W(\theta)$ is your prior belief that the world is in state θ "
- Our reinterpretation justifies Bayesian approaches in *some* situations with hardly *any* assumptions
  - With prior W, your predictions will never be worse than the predictions made by any θ up to luckiness term

$$-\log W(\theta)$$

no matter what data you will observe

Bayes vs. Worst-Case

- Bayesian statistics in its usual interpretation relies on heavy assumptions about the world:
  - " $W(\theta)$ is your prior belief that the world is in state θ "
- Our reinterpretation justifies Bayesian approaches in *some* situations with hardly *any* assumptions
  - With prior W, your predictions will never be worse than the predictions made by any θ up to luckiness term

$$-\log W(\theta)$$

no matter what data you will observe

Objective Subjectivity: Using Prior **Distributions** without **committing** to Prior **Assumptions**

Beware – we focus on log-loss

- Bayesian statistics in its usual interpretation relies on heavy assumptions about the world:
  - "$W(\theta)$ is your prior belief that the world is in state θ"
- Our reinterpretation justifies Bayesian approaches in *some* situations with hardly *any* assumptions
  - With prior W, your predictions will never be worse than the predictions made by any θ up to luckiness term
    $$-\log W(\theta)$$
    no matter what data you will observe

In other situations (loss functions) need to change Bayes a little

Aggregating Algorithm/Hedge

Vovk '90, Freund & Shapire '98

- Both algorithms work like this:
  1. fix "appropriate" η
  2. For each i, y^i, calculate generalized posterior $W_\eta(\theta \mid y^i)$ and predict y_{i+1} using some fixed function f, $a_{i+1} := f(W_\eta(\theta \mid y^i))$

UPSHOT: the algorithm is not Bayes any more, but the bounds still involve priors!

Regret Bounds for AA/Hedge:

- We have for all n, y^n, θ :

$$\text{regret}(y^n, \mathbf{AA}, \theta) :=$$

$$\text{loss}(y^n, \mathbf{AA}) - \text{loss}(y^n, \theta)$$

$$\leq \begin{cases} -\log W(\theta) & \text{for log-loss} \\ \frac{1}{\eta} \cdot -\log W(\theta) & \text{for } \eta\text{-mixable loss functions} \\ C \cdot \sqrt{-\log W(\theta)} & \text{for other bounded losses, e.g.0/1*} \end{cases}$$

Regret Bounds for AA/Hedge:

Priors remain there even though we have different loss fn!

- We have for all n, y^n, θ :

$$\text{regret}(y^n, \mathbf{AA}, \theta) :=$$

$$\text{loss}(y^n, \mathbf{AA}) - \text{loss}(y^n, \theta)$$

$$\leq \begin{cases} -\log W(\theta) & \text{for log-loss} \\ \frac{1}{\eta} \cdot -\log W(\theta) & \text{for } \eta\text{-mixable loss functions} \\ C \cdot \sqrt{-\log W(\theta)} & \text{for other bounded losses, e.g.0/1*} \end{cases}$$

Regret Bounds for AA/Hedge:

- We have for all n, y^n, θ :

$$\text{regret}(y^n, \mathbf{AA}, \theta) :=$$

$$\text{loss}(y^n, \mathbf{AA}) - \text{loss}(y^n, \theta)$$

$$\leq \begin{cases} -\log W(\theta) & \text{for log-loss} \\ \frac{1}{\eta} \cdot -\log W(\theta) & \text{for } \eta\text{-mixable loss functions} \\ C \cdot \sqrt{-\log W(\theta)} & \text{for other bounded losses, e.g.0/1*} \end{cases}$$

no stochastic assumptions whatsoever!

What we did Today

1. Prediction with difficult loss functions
   - Vovk's mixability, 0/1-loss, the Hedge Algorithm
2. Relations to Kolmogorov Complexity KM(x)
   - Solomonoff prediction, superloss processes
3. Relations to Minimum Description Length, Occam's Razor
4. Bayes and Luckiness
   - "Objective Subjectivity",

Relevance

- We analyzed one aspect of inductive inference (prediction) without making any distributional assumptions.
- Why is this relevant?
  1. Philosophical Reasons
  2. Practical Reasons

Philosophical Relevance I

- Interesting new interpretation of a priori distribution 'hope' vs. 'expectation' (same word in Latin languages!)

Philosophical Relevance II

- Philosophers (e.g. Schultz and Thorn, 2012) have recently discovered universal prediction under the name meta-induction
- They realize: if we set the goal more modestly (learning to be as good as the best comparator, rather than learning to be good 'absolutely', then Hume's 'Fundamental Problem of Induction' **disappears…**
- However, they are reinventing all the wheels of the last 20 years…
- Open "Problem": get philosophers to embrace this!

Philosophical Relevance III

- People not immersed into statistics before a certain critical age often feel ill at ease when reading
  'assume $Y_1, Y_2, \ldots$ distributed according to some P, P being a member of some family $\mathcal{M}$ '
- Does randomness exist in the real world?

Practical Relevance

- People not immersed into statistics before a certain critical age often feel ill at ease when reading
  'assume $Y_1, Y_2, \ldots$ distributed according to some P, P being a member of some family $\mathcal{M}$ '
- Does randomness exist in our application?

  Is it possible to include Tolstoy's ***War and Peace*** in a reasonable way into the set of 'all possible novels' and further to postulate the existence of a certain probability distribution in this set? Must we assume that the individual scenes in this book form a random sequence with stochastic relations that damp out quite rapidly over a distance of several pages?

  - A.N. Kolmogorov (1965)

Practical Relevance

- People not immersed into statistics before a certain critical age often feel ill at ease when reading
  'assume $Y_1, Y_2, \ldots$ distributed according to some P, P being a member of some family $\mathcal{M}$ '
- Does randomness exist in our application?
  - to be fair, I have applications in mind far beyond what was envisaged by the founding fathers of 20th century statistics: