Second-Order Induction and the Importance of Precedents

Rossella Argenziano and Itzhak Gilboa

June 15, 2017

Highlights

- How are probabilistic beliefs formed?
  - Where there aren't many "identical" cases
  - Revolutions, elections, economic policy
- Case-based beliefs
  - Similarity-weighted relative frequency
  - As in kernel estimation of probabilities
- Main point: the similarity function is also learnt from the data
  - "Second-order induction"
  - The "empirical similarity"

Features of the Model

- $x^1, ..., x^m$ predicting y, all binary
- Estimating the probability that $y_p = 1$ by

$$\overline{y}_p^s = \frac{\sum_{i \leq n} s(x_i, x_p) y_i}{\sum_{i \leq n} s(x_i, x_p)}$$

- The similarity is binary, and defines a partition
- The question is, thus, which set of variables to use?

Highlight of Main Results

- A larger set of predictors need not provide a better fit
  - Even in-sample, due to the "curse of dimensionality"
  - We provide conditions under which it will
  - Smaller sets are also preferred due to overfitting
- With many predictors, different beliefs (due to different subsets of predictors) is to be expected.
- Finding the best set is NPC
  - Hence even if there is a unique best set it may not be found
- In a toy model, it is easier to establish reputation than to re-establish it

Motivating Example I: President Obama

- A precedent that changes the probability of a non-white president above and beyond its weight in the sample
- In our conceptualization, because it changes the weight of the variable "race" in the similarity function
- This could be relevant also to other races (in a non-binary model).

Motivating Example II: Change of Currency

- The French Franc dropped two zeroes in 1960
- 10 Israeli Lira became 1 Israeli Shekel in 1980
- 1000 Israeli Shekels became 1 New Israeli Shekel in 1985
- These changes used perceptual similarity
- As such, they don't seem to be about the empirical similarity
- But, at least in the Israeli example, when the perceptual change wasn't accompanied by change of policy it didn't work
- People seemed to have been smart enough to "compute" the empirical similarity.

Model

- $M \equiv \{1, ..., m\}$ set of predictors
- $x \equiv \left(x^1, ..., x^m\right) \in X \equiv \{0,1\}^m$; the predicted variable, $y \in \{0,1\}$
- The *prediction problem* is a pair (B, x_p) where $B = \left\{\left(x_i, y_i\right)\right\}_{i \leq n}$ are observations (or "cases"), $x_i = \left(x_i^1, ..., x_i^m\right) \in X$, $y_i \in \{0,1\}$, and $x_p \in X$ is a new data point
- Given a function $s : X \times X \to \{0,1\}$, the probability that $y_p = 1$ is estimated by

$$\overline{y}_p^s = \frac{\sum_{i \leq n} s\left(x_i, x_p\right) y_i}{\sum_{i \leq n} s\left(x_i, x_p\right)}$$

if $\sum_{i \leq n} s\left(x_i, x_p\right) > 0$ and $\overline{y}_p^s = 0.5$ otherwise.

The Similarity Function

- Given weights $\left(w^1, ..., w^m\right) \in X \left(\equiv \{0,1\}^m\right)$, let

$$s_w\left(x_i, x_p\right) = \prod_{\{j | w^j = 1\}} \mathbf{1}_{\left\{x_i^j = x_p^j\right\}}$$

- This is limited in several ways:
  - Similarity is yes/no
  - And assumed transitive
- But it suffices to convey some points.

The Empirical Similarity

- In-sample estimation

$$\overline{y}_i^s = \frac{\sum_{k \neq i} s(x_k, x_i) y_k}{\sum_{k \neq i} s(x_k, x_i)}$$

  if $\sum_{j \neq i} s(x_j, x_i) > 0$ and $\overline{y}_i^s = 0.5$ otherwise

- Define the sum of squared errors to be

$$SSE(s) = \sum_{i=1}^{n} \left(\overline{y}_i^s - y_i\right)^2$$

The Empirical Similarity – cont.

- It will also be convenient to consider the mean (squared) error, that is,

$$MSE\left(s\right) = SSE\left(s\right)/n$$

- When s is defined by the variables in $J \subset M$ $\left(w^j = 1_{\{j \in J\}}\right)$, we simply refer to the above as $MSE\left(J\right)$
- And, in order to deal with overfitting, define also

$$AMSE(J, c) \equiv MSE(J) + c|J|$$

Non-Monotonicity

Example

i	x_i^1	y_i
1	0	0
2	0	1
3	1	0
4	1	1

It can be seen that the MSE's of the subsets of variables are given by

J	$MSE(J)$
$\varnothing$	$4/9$
$\{1\}$	1

Replication

- The above hinges on the "bins" defined by $J = \{1\}$ being very small
- And suggests

Definition

Given two databases $B = \{(x_i, y_i)\}_{i \leq n}$ and $B' = \{(x'_k, y'_k)\}_{k \leq tn}$ (for $t \geq 1$), we say that B' is a *t-replica of B* if, for every $k \leq tn$, $(x'_k, y'_k) = (x_i, y_i)$ where $i = k (\mathrm{mod}\, n)$.

- Yet, for a database B' which is a t-replica of the above

$$MSE(\varnothing) = \left(\frac{2t}{4t-1}\right)^2 < \left(\frac{t}{2t-1}\right)^2 = MSE(\{1\}).$$

Informativeness

- In the Example x^1 added no information regarding y
- Formally,

Definition

A variable $j \in M$ is *informative* relative to a subset $J \subset M \setminus \{j\}$ in database $B = \{(x_i, y_i)\}_{i \leq n}$ if there exists $z \in \{0,1\}^J$ such that $|b(J, z \cdot 0)|, |b(J, z \cdot 1)| > 0$ and

$$\overline{y}^{(J \cdot j, z \cdot 0)} \neq \overline{y}^{(J \cdot j, z \cdot 1)}$$

where the above are the average y's in the corresponding sub-databases ("bins")

Monotonicity

Theorem

Assume that j is informative relative to $J \subset M \setminus \{j\}$ in the database $B = \{(x_i, y_i)\}_{i \leq n}$. Then there exists a $T \geq 1$ such that, for all $t \geq T$, for a t-replica of B, $MSE\,(J \cup \{j\}) < MSE\,(J)$. Conversely, if j is not informative relative to J, then for any t-replica of B, $MSE\,(J \cup \{j\}) \geq MSE\,(J)$, with a strict inequality unless j is a function of J.

Non-Uniqueness

Example

t replications of

i	x_i^1	x_i^2	y_i
1	1	0	0
2	1	0	1
3	0	1	0
4	0	1	1
5-8	0	0	0
9-12	1	1	1

For $t = 1$,

J	$MSE\,(J)$
$\varnothing$	0.297
$\{1\}$	0.2
$\{2\}$	0.2
$\{1, 2\}$	0.333

Differences of Opinions

- Let n be fixed and let m grow with

$$P\left(x_i^j = 1 \,\middle|\, x_k^l,\ l < j \text{ or } (l = j, k < i)\right) \in (\varepsilon, 1-\varepsilon)$$

for a fixed $\varepsilon \in (0, 0.5)$.

Proposition

For every $n \geq 4$ and every $\varepsilon \in (0, 0.5)$, if there are at least two cases with $y_i = 1$ and at least two with $y_i = 0$, then, as $m \to \infty$, the probability that there exist J,J' with $J \cap J' = \varnothing$ and $MSE(J) = MSE(J') = 0$ tends to 1.

Fixed m, growing n, (science) vs. the other way around (art?)

A Complexity Result

Define

Problem

EMPIRICAL-SIMILARITY: Given integers $m, n \geq 1$, *a database* $B = \{(x_i, y_i)\}_{i \leq n}$, *and (rational) numbers* $c, R \geq 0$, *is there a set* $J \subset M \equiv \{1, ..., m\}$ *such that* $AMSE(J, c) \leq R$*?*

Thus, EMPIRICAL-SIMILARITY is the yes/no version of the optimization problem, "Find the empirical similarity for database B and constant c". We can now state

Theorem

EMPIRICAL-SIMILARITY is NPC.

Application: Reputation

- There are $2N$ past cases in which $x_i^j = 0$ (a new player is now joining the scene)
- Among these, N times $y_i = 1$ and N times $y_i = 0$
- There are $k + l$ new cases in which $x_i^j = 1$. In $k \geq 0 - y_i = 1$, and in $l \geq 0 - y_i = 0$
- For N and l, let $k(N, l)$ is the minimal k for which $MSE(J \cup \{j\}) < MSE(J)$.

Proposition

Let there be given $N > 2$ and $l \geq 0$. Then:
(i) For every $l \geq 0$ there exists k_0 such that for all $k \geq k_0$, $MSE(J \cup \{j\}) < MSE(J)$; in particular, $k(N, l)$ is finite (as $k(N, l) \leq k_0$);
(ii) $k(N, 0) = 2$;
(iii) $k(N, 1) = 5$.

A Continuous Model

- $\left(x^1, ..., x^m\right) \in X \subseteq \mathbb{R}^m$, $y \in Y \subseteq \mathbb{R}$

$$\overline{y}_p^s = \frac{\sum_{i \leq n} s\left(x_i, x_p\right) y_i}{\sum_{i \leq n} s\left(x_i, x_p\right)}$$

- Similarity

$$s\left(x, x'\right) = \exp\left(-\sum_{j=1}^{m} w^j \left(x^j - x'^j\right)^2\right)$$

  $w^j \in \mathbb{R}_+ \cup \{\infty\}$

- $w^j = \infty$ means that $s\left(x, x'\right)$ if $x^j \neq x'^j$.

A Continuous Model – cont.

- Define

$$AMSE(w, c) \equiv MSE(w) + c|J(w)|$$

where

$$J(w) = \left\{ j \leq m \;\middle|\; w^j > 0 \right\}$$

- We assume a fixed cost for using a variable
  - Cost of obtaining the data
  - Cost of recalling it and using it in computations.

A Continuous Model – cont.

Problem

CONTINUOUS-EMPIRICAL-SIMILARITY: Given integers $m, n \geq 1$, a database of rational valued observations, $B = \{(x_i, y_i)\}_{i \leq n}$, and (rational) numbers $c, R \geq 0$, is there a vector of extended rational non-negative numbers w such that $AMSE(w, c) \leq R$?

And we can state

Theorem

CONTINUOUS-EMPIRICAL-SIMILARITY is NPC.

Discussion: Equilibrium selection

- Most of our examples are equilibrium selection in a coordination game (revolutions, elections, inflation)
- The empirical similarity can be viewed as a theory of focal points
- It is compatible with agents being very naive or very sophisticated
- As well as with a distribution of the agents across Level-K reasoning for various K's.

Discussion: Bayesianism

- More generally, this theory of belief formation is compatible with
  - A minimal view of Bayesianism (prior restricted to the possible values of y in a given period)
  - A maximal view of Bayesianism (where the prior is defined over the entire history).

Discussion: Associative Rules

- Undoubtedly, people also think in terms of rules
- In particular, many of our examples can be viewed as association rules
  - ("*If* it is a country in the Soviet Bloc, *then* it will not be allowed to break free")
- It is less obvious how one generates probabilities from association rules
  - Rules can contradict each other
  - They can all be silent.

Discussion: Regression

- Much of our stories can be told in the language of regression
- One difference: regression can only improve, in-sample, by adding variables
- But the complexity the non-uniqueness results hold
  - "Fact-Free Learning" (Aragones, Gilboa, Postlewaite, Schmeidler, 2005)
- We find similarity-weighted empirical frequencies somewhat more plausible as a cognitive model.