Informing by Example

Pieter Adriaans

SNE, complexity group, IvI, Department of Computer Science
University of Amsterdam,
Science Park 107
1098 XG Amsterdam,
The Netherlands.
P.W.Adriaans@science.uva.nl

Abstract. The best way to explain a theory is to give a simple example that illustrates the theory: that is the fundamental insight behind informing by example (IbE). Using a systematic separation between datasets and their models it is shown that IbE has a wide field of applications both in empirical setting (learning from small data sets using standard compression algorithms) as in the more theoretical context of Shannon information systems and Turing systems. For Turing systems the models selected by IbE satify Zellner's information conservation criterion.

1 Introduction

In this paper we introduce, what we believe to be, a new approach to machine learning and theory formation that we call 'Informing by Example": the best model for a system is a typical example.

Example 1. An example would be Bohr's model of the atom as a small solar system. The example is very simple (one sun, one planet) but it allows us to apply classical mechanics on a sub-atomic level. It also helps us to formulate new theories (electron jumps)

Simple examples of complex things have value. We can use them to explain phenomena and to discover new knowledge.

Example 2. Any mathematician will immediately recognize the Fibonacci sequence $0, 1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, 144, 233, 377, 610, 987, 1597, 2584, 4181....$ A relevant question is: How many examples do we need to recognize this sequence. If we use WoframAlpha as our pattern recognizer the answer is 5. The program recognizes the sequence $0, 1, 1, 2, 3$ as Fibonacci, but interprets $0, 1, 1, 2$ as just a sequence of numbers without any significance. Apparently the sequence $0, 1, 1, 2, 3$ contains, in the mathematical context of WolframAlpha, enough information to identify the basic recursive operation that defines the Fibonacci sequence. It contains, in this context, sufficient information to tell us everything there is to know about Fibonacci numbers. Since it is also the simplest example it is in some sense the 'best' example.

A theory about quality of models can be derived from Bayes' law and Shannon's notion of information. Let $\mathcal{M}$ be a set of hypotheses and let x be a data set. Using Bayes' law, the optimal computational model under this distribution would be:

$$M_{map}(x) = argmax_{M \in \mathcal{M}} \frac{P(M)P(x|M)}{P(x)} \tag{1}$$

under the constraint $P(x) \geq P(M)P(x|M)$. This is equivalent to optimizing:

$$argmin_{M \in \mathcal{M}} - \log P(M) - \log P(x|M) \tag{2}$$

under the constraint $-\log P(x) \leq -\log P(M) - \log P(x|M)$. Here $-\log P(M)$ can be interpreted as the length of the optimal *model code* in Shannon's sense and $-\log P(D|M)$ as the length of the optimal *data-to-model code*; i.e. the data interpreted with help of the model.

This suggests that learning an optimal model is a form of data compression under constraints. For the optimal model the following should hold $-\log P(M|x) = -\log P(M) - \log P(x|M) + \log P(x)$ which is equivalent to:

Definition 1 (Zellner's Information conservation principle).

$$I(x) + I(M|x) = i(M) + I(x|M) \tag{3}$$

This suggests the following interpretation: the information needed to extract a theory from a data set is the surplus of information that the theory adds to the data set. If the theory is optimal, then the sum of the model code and the data-to-model code are in balance with the optimal code of the data. In this case the model can be derived directly from the data. The additional information is zero bits which makes the conditional probability of the theory given the data 1.

The best model for a data set is the smallest data set from which an optimal theory can be derived.

Definition 2 (Informing by Example (IbE)). *The best theory to explain a set of data is the one which minimizes the sum of:*

- *the length of an example data set from which the theory can be derived and*
- *the length of the data when encoded with the help of the theory.*

The essence of informing by example is that we do not measure the model complexity, instead we rank the models by their example sets.

Example 3. Consider this data set:

$$3.14159265358979323846264338327950288419716939937510582097494$$

which consists of the 60 digits of the number π. We measure the complexity of the example as a sequence of digits (i.e. $l(S) log_2 10$). It could be from various classes of data sets (i.e. initial segments of real numbers, or initial segments of wellknown mathematical constants). Using IbE we need a model compressor q that recognizes the data set as an expansion of π. We are not specifically interested in the complexity of q nor in the complexity of a small generating program p for the digits of π. What is of interest is the following: if q can identify π from its first 60 digits of π then probably also from the first 40 or 30. So what is the smallest amount of digits we need to feed to q before it greedily guesses π? Note that, even with 60 or any other finite number of digits, identifying the sequence as π is greedy. Lets use WolframAlpha as our greedy data constructor q. We find that it needs at least 4 digits. Then our IbE complexity becomes something like "Expand example '3.141' to 60 digits", where '3.141' is the model code and "Expand example ... to 60 digits" is the data to model code. Even in plain ASCII this is shorter than the original data set. A more formal estimate would be: $3.141 + \log_2 60 + O(1) = 4 \log_2 10 + \log_2 60 + O(1) \approx 26 + c$ bits, where c is a constant number of bits for the expand instruction.

2 Formal Definition of IbE

Definition 3 (IbE problem class).
An IbE problem class is a tuple $p =< \mathcal{D}, \mathcal{M}, dm, mm, ms >$ *with*

- *A class of data sets* $\mathcal{D}$.
- *A class of models* $\mathcal{M}$.
- *A data measurement function* dm.
- *A data-to-model measurement function* mm.
- *A model selector* ms

with de following defininitions:

Definition 4 (A class of Data Sets). *A countable class of data sets* $\mathcal{D}$ *with a partial order* $\preceq$ *and a minimal element* $\emptyset$.

Definition 5 (A class of Models). *A possibbly uncountable class of models* $\mathcal{M}$.

Definition 6 (A Data Measure). *A data measurement function* $dm : \mathcal{D} \to \mathbb{R}$, *where* $\mathcal{D}$ *is a class of data sets, with* $\forall x, y, \in \mathcal{D}(x \preceq y \Rightarrow dm(x) \leq dm(y))$ *and* $dm(\emptyset) = 0$.

Definition 7 (A Data to Model Measure). *A data to model measurement function* $mm : (\mathcal{D} \times \mathcal{M}) \to \mathbb{R}$. *This function gives the conditional compresion of a data set given a model.*

Definition 8 (Model Selector). *Given a set of data sets* $\mathcal{D}$ *and a set of models* $\mathcal{M}$, *a model selector is a function* $mc : \mathcal{D} \to \mathcal{M}$. *More data sets can have the same model. The* Representeation Bias Function $ms^{-1} : \mathcal{M} \to \mathcal{P}(\mathcal{D})$ *gives us the set of data sets with the same model.*

The operation $ms^{-1}(ms(M))$ gives the *Representional Bias* of a data selector, which indicates how 'greedy' the compressor is. It gives us for any set the set of data sets with the same model as a given data set.

Definition 9 (IbE problem).
Given an IbE problem class $p =< \mathcal{D}, \mathcal{M}, dm, mm, ms >$ an IbE problem for $x \in \mathcal{D}$ is:

$$M_{IbE}(x,p) = argmin_{d \in \mathcal{D}}\ dm(d) + mm(x, ms(d)) \tag{4}$$

The solution to a IbE problem is the set of smallest examples that minimize the sum of the model code and the data to model code. It is now possible to give an objective measure for the relative amount of meaningful or useful information in a data set.

Definition 10 (Facticity, relative to an IbE problem).
Let $p =< \mathcal{D}, \mathcal{M}, dm, mm, ms >$ be an IbE problem class, given the fact that all the smallest examples i in $M_{IbE}(x,p)$ have the same length $l = dm(i)$ the amount of model information or facticity *of the problem $M_{IbE}(x,p)$ is:*

$$\phi_p(x) = -\log_2 \Sigma_{i \in M_{IbE}(x,p)} 2^{-dm(i)} = -\log_2 |M_{IbE}(x,p)| 2^{-l} \tag{5}$$

i.e. uniform length $l = dm(i)$ of all small examples penalized by their bias (i.e. the number of optimal examples $|M_{IbE}(x,p)|$).

The definition is descriptive, not constructive: it specifies what the best models are, not how to find them. For an implementation of IbE we need to specify a search routine that constructs an optimal example set. The next lemma specifies the conditions under which this is possible:

Theorem 1 (Facticity is recursive for IbE).
Let $p =< \mathcal{D}, \mathcal{M}, dm, mm, ms >$ be an IbE problem. If the functions $\preceq$, dm, mm, ms and mc are recursive then $M_{IbE}(x,p)$ and $\phi_p(x)$ are recursive.

Proof: We only need to investigate data sets smaller then d. For all $d \in \mathcal{D}$ the number of elements $d_i \preceq d$ is finite, so the set can be enumerated. Since the functions are recursive we can compute the IbE estimate for each element and collect the smallest ones. □

2.1 Discussion

Informing by example deploys two notions of compression or data abstraction. An effective form of weak compression performed by the model selector function ms (weak generalization) and a compression based on an exhausitive search by the *argmin* function that, depending on the knowlegde representation, often will be exponential or worse in complexity (strong generalization). The exhaustive search will in most cases be replaced by some heuristics. The balance between these two forms of generalization is essential. The exhaustive search searchest the example that makes the best use of the bias of the model compressor. The more greedy the compressor is, the smaller the guiding example can be, but the bigger the risk is that it overgeneralizes beyond the optimal model. In terms of human cognition one could say that the model compression function is the pattern recognition function our brain and the exhaustive search process is science itself. Once we have found a regularity in nature we memorize it as a paradigm: a small example form which we can derive the pattern. The bias of our pattern recognition skills are optimized by a process of evolution. The advantages of the IbE approach are clear:

- It gives objective computable estimates of the model code length.
- Model code and data to model code both are measured in terms of complexity of data sets and therefore balanced.
- It works for small data sets.
- It is computable for a large domain of problem classes.
- The notion of a simplest example seems from a cognitive point of view in line with the way human cognition works (c.f. Kuhn's notion of a paradigm).

There are also disadvantages:

- There might not be a small example that represents the true model class.
- Simple examples might be very large in comparison to their optimal descriptive complexity.

3 IbE for Shannon systems of messages

Suppose we have a discrete random variable X with possible values $\{x_1, ..., x_k\}$ and probability mass function $P(X)$ then the entropy is:

$$H(P) = H(X) = -\sum_{x \in X} P(x) \log_2 P(x) \tag{6}$$

The average entropy per message is the rate:

$$r = \lim_{n \to \infty} \frac{1}{n} X_1, X_, ..., X_n \tag{7}$$

The absolute rate is $R = \log |X|$. The absolute redundancy is $D = R - r$.

Definition 11 (IbE Shannon sequence problem class). *An IbE Shannon problem class is a tuple*

$$p_{ShannonSequence} =< \mathcal{D}, \mathcal{M}, dm, mm, ms >$$

with

- *A class of data sets $\mathcal{D} = \{x_1, ..., x_k\}^*$, i.e. the set of all finite sequences of messages, with a partial order $x \preceq y$ iff $|x| < |y|$ and $\forall x_i (x_i \in x \rightarrow x_i \in y)$.*
- *A class of models the set $\mathcal{M} = \mathcal{P}_X$ of all finite probability distributions on X.*

- *A data measurement function* $dm : \mathcal{D} \rightarrow \mathbb{R}$*, with* $x \preceq y \Rightarrow dm(x) \leq dm(y)$*. In this case* $\forall x \in \mathcal{D}\ dm(x) = l(x)\log_2 k$*.*
- *A data-to-model measurement function* $mm : (\mathcal{D} \times \mathcal{M}) \rightarrow \mathbb{R}$*.* $\forall x \in \mathcal{D}\ \forall P \in \mathcal{M}\ mm(x) = l(x)r_P$ *where* r_P *is the rate defined by* P*.*
- *A model selector* $ms : \mathcal{D} \rightarrow \mathcal{M}$*. In this case the Maximum Likelyhood function for* $x \in \mathcal{M} = \mathcal{P}_X$ *the estimate is* $P(x_i) = \frac{freq(x_i)}{l(x)}$*.*

Lemma 1. *For the Shannon problem class* $p_s =< \mathcal{D}, \mathcal{M}, dm, mm, ms >$ *the problem* $M_{IbE}(x, p_s)$ *and its facticity* $\phi_{p_s}(x)$ *are recursive.*

Proof: instance of theorem **??**. □

Heuristics for finfing a good estimate of the optimal IbE model could involve discretization of the distribution. A model M with k parameters contains uncountably many distributions but only in the order of $2^{(\frac{k}{2})\log n} = n^{\frac{k}{2}}$ differ signiticantly on a sequence of length n **?**] **?**].

Example 4. Consider an *unfair* 4-sided dice, with $P(1) = 0.5, P(2) = 0.125, P(3) = 0.125, P(4) = 0.25$. A simple sequence from which we can learn this distribution is: 11114423: i.e. $ms(\text{”11114423”}) = P$. This gives $P(1) = \frac{4}{8} = 0.5, P(2) = \frac{1}{8} = 0.125, P(3) = \frac{1}{8} = 0.125, P(4) = \frac{2}{8} = 0.25$. The entropy then is $H(X) = 0.5\log_2 0.5 + 0.25\log_2 0.25 + 0.125\log_2 0.125 + 0.125\log_2 0.125 = 1.75$ with $D = \log_2 4 - 1.75 = 0.25$. The length of a simle example is $8\log_2 4 = 16$ bits. $ms^{-1}(ms(\text{”11114423”}))$ is the set with the same distributions.

The *argmin* function will construct the set of smallest examples with the same distribution as ”11114423”. This class has $\binom{6}{2} + 7 + 8 = 30$ elements, which gives a facitiscty in bits of

$$\phi_p(\text{”44444423”}) = -\log_2 \Sigma_{i \in M_{IbE}(x,p)} 2^{-|i|} = -\log_2 \frac{30}{2^{16}} \approx 11$$

The simplest example for a uniform distribution of k messages is a string of k different messages

$$\phi_p(\text{”}x_1 x_2 ... x_k\text{”}) = -\log_2 \Sigma_{i \in M_{IbE}(x,p)} 2^{-|i|} = -\log_2 \frac{k!}{2^{k\log_2 k}} = -\frac{\log k^{-k} k!}{\log_2 k}$$

This is in the same order of magnitude as k. (e.g. for a sequence x with 1000 different messages $phi_p(x) \approx 1436$. So uniform discributions have in this interpretation model codes with a length in the order of the number of messages. A single string with 1000 different message would have a length of $1000 log_2 1000 \approx 10.000$ bits. According to the coupon collector theorem we need at least $k \log k$ samples in order to collect each messages. which gives a sequence of $k(\log k)^2$ bits, which leads to a sample of 100.000, so the 1400 bit facticity is quite acceptable.

Note that the facticity is depreciated by the fact that many strings generate the same estimate for a distribution. This could be mended by definining IbE in multisets of messages in stead of sequences. The minimal example then would be a multiset of messages that has a one to one correspondence to an estimated distribution and a much more compact code. The general picture that emerges is that a richer model class and a better model selector produce better learning results. In the following paragraph we investigate IbE in the context the most general class of model: partial recursive functions.

4 IbE for Turing systems

In this section we use latin lowercas x, y, z to indicate strings and greek lowercase α, β, γ to indicate models (i.e. pre-fix free strings that act as programs for a universal Turing machine)

Definition 12 (IbE Turing problem class).
An IbE Turing problem class is a tuple

$$p_{Turing} =< \mathcal{D}, \mathcal{M}, dm, mm, ms >$$

with

- *A class of data sets $\mathcal{D} = \{0,1\}^*$, i.e. the set of all finite binary strings, with a partial order $x \preceq y$ iff $l(x) < l(y)$.*
- *A class of models: A prefix-free self delimiting code $\mathcal{M}$ for a universal Turing machine T_U.*
- *A data measurement function $dm(x) = l(x)$,*
- *A data-to-model measurement function $mm(x, \alpha) = l(z)$ where $z = \min_i(i \in \mathcal{D} \wedge T_U(\alpha i) = x)$.*
- *A model selector $ms : \mathcal{D} \rightarrow \mathcal{M}$. In this case a function that assigns to each binary string exactly one corresponding prefix-free string: $ms(x) = \overline{x}$.*

For the Turing problem class $p_t =< \mathcal{D}, \mathcal{M}, dm, mm, ms >$ the problem $M_{IbE}(x, p_s)$ and its facticity $\phi_{p_t}(x)$ are not recursive. Specifically the model measurement function mm is not recursive because it would involve running any program on any input.

Lemma 2. *The models selected by $M_{IbE}(x, p_t)$ satisfy Zellner's information conservation criterion $I(x) + I(M|x) = i(M) + I(x|M)$.*

Proof:

$$M_{IbE}(x, p_t) = argmin_{d \in \mathcal{D}}\ dm(d) + mm(x, ms(d))$$

$$= argmin_{d \in \mathcal{D}}\ l(d) + mm(x, \overline{d})$$

Since $mm(x, \overline{d}) = l(z)$ where $z = \min_i(i \in \mathcal{D} \wedge T_U(\overline{d}i) = x)$ this amounts to:

$$M_{IbE}(x, p_t) = argmin_{d,e \in \mathcal{D}} \begin{cases} l(d) + l(e) \\ \text{where } T_U(\overline{d}e) = x \end{cases}$$

Since the model selector is one to one the extra bits making the code self-delimiting do not contain any information $I(x) = I(\overline{d}e) = l(d) + l(e)$, $I(M) = l(d)$ and $I(x|M) = l(e)$ this satisfies Zellner's Information conservation principle (c.f. definition **??**), for $I(M|x) = 0$, i.e. the model is defined by the data. □

Definition 13 (IbE problem).
Given an IbE problem class $p =< \mathcal{D}, \mathcal{M}, dm, mm, ms >$ *an IbE problem for* $x \in \mathcal{D}$ *is:*

$$M_{IbE}(x,p) = argmin_{d\in\mathcal{D}}\ dm(d) + mm(x, ms(d)) \tag{8}$$

Bibliography

Schwarz, Gideon E. (1978), "Estimating the dimension of a model", Annals of Statistics 6 (2): 461464

Krichevsky, R.E. and Trofimov V.K. (1981), 'The Performance of Universal Encoding', IEEE Trans. Information Theory, Vol. IT-27, No. 2, pp. 199207)

Kuhn, TS. 1970. *The Structure of Scientific Revolutions.* The University of Chicago Press, Chicago.

P.W. Adriaans and A. Golan, (2011) Generalized Kolmogorov Complexity, Rapport of the Info-metrics Institute.

P.W. Adriaans, (2009) Between Order and Chaos: The Quest for Meaningful Information, Theory of Computing Systems, Volume 45 , Issue 4 (July 2009), Special Issue: Computation and Logic in the Real World; Guest Editors: S. Barry Cooper, Elvira Mayordomo and Andrea Sorbi, 650-674.

Pieter Adriaans, Facticity as the amount of self-descriptive information in a data set, arXiv:1203.2245 [cs.IT], 2012.

P.W.Adriaans, (2007) The philosophy of learning, in Handbook of the philosophy of information, P.W. Adriaans, J. van Benthem eds. in Handbook of the philosophy of science, Series edited by D. M. Gabbay , P. Thagard and J. Woods.

P.W. Adriaans and P. Vitányi, (2009) Approximation of the Two-Part MDL Code, Comput. Sci. Dept., Univ. of Amsterdam, Amsterdam; Information Theory, IEEE Transactions on, Volume: 55, Issue: 1, On page(s): 444-457.

P.W. Adriaans and C. Jacobs, (2006) Using MDL for grammar induction, , In Proceedings of the 8th International Colloquium on Grammatical Inference (ICGI). Lecture Notes in Artificial Intelligence, Sakaibara, Y.; Kobayashi, S.; Sato, K.; Nishino, T.; Tomita, E. (Eds.). Tokyo, Japan, September 21, LNAI, vol. 4201

L. Antunes and L. Fortnow, (2003) Sophistication Revisited. In Proceedings of the 30th International Colloquium on Automata, Languages and Programming, volume 2719 of Lecture Notes in Computer Science, pages 267-277. Springer.

L. Antunes, L. Fortnow. D. Van Melkebeek and N. V. Vinodch, (2006) Computational depth: Concept and application, Theoretical Computer Science, volume, 354.

R. Arnheim (1971), Entropy and Art, an essay on order and disorder, 40th Anniversary edition, University of California Press.

F.A. Bais and J.D. Farmer, (2007) The physics of information, Handbook of the philosophy of information, P.W.Adriaans, J. van Benthem eds. in Handbook of the philosophy of science, Series edited by D. M. Gabbay , P. Thagard and J. Woods.

C. H. Bennett, (1988) Logical depth and physical complexity. In R. Herken, editor, The Universal Turing Machine: A Half-Century Survey, pages 227-257. Oxford University Press.

Max Bense: Aesthetica. Einfhrung in die neue Aesthetik. Baden-Baden: Agis-Verlag, 1965.

G.D. Birkhoff: Collected Mathematical Papers, New York: American Mathematical Society, 1950.

R. Cilibrasi, P. Vitányi, (2005) Clustering by compression, IEEE Trans. Inform. Theor. 51 (4) 1523 1545.

Cover T.M. and Thomas, J.A. (2006) Elements of Information theory, Wiley.

J.P. Crutchfield and K. Young, (1989) Inferring Statistical Complexity. Physical Review Letters 63:105.

J.P. Crutchfield and K. Young, (1990) Computation at the Onset of Chaos, in Entropy, Complexity, and the Physics of Information, W. Zurek, editor, SFI Studies in the Sciences of Complexity, VIII, Addison-Wesley, Reading, Massachusetts. pp. 223-269.

J. P. Crutchfield, (1994) The Calculi of Emergence: Computation, Dynamics, and Induction, Physica D 75, pg. 11-54.

R.A. Fisher (1925), Theory of statistical estimation, Proc. Cambridge Philos. Soc. 22, 700-725.

Foley, D.K. (2010) Notes on Bayesian inference and effective complexity, unpublished manuscript.

Gell-Mann, M. and Lloyd, S. (2003) Effective complexity. In Murray Gell-Mann and Constantino Tsallis, eds. *Nonextensive entropy–Interdisciplinary applications*, Oxford University Press, 387-398.

P.D. Grünwald, (2007) The Minimum Description Length Principle. MIT Press.

Hopcroft, J. E., Motwani, R., Ullman, J. D., (2001) Introduction to Automata Theory, Languages, and Computation Second Edition. Addison-Wesley.

M. Koppel, (1987) Complexity, Depth, and Sophistication", in Complex Systems 1, pages = 1087-1091.

C.G. Langton, (1990) Computation at the edge of chaos: Phase Transitions and Emergent Computation. Physica D, 42, 1990.

Li M., Vitányi P.M.B. (2008) An Introduction to Kolmogorov Complexity and Its Applications, 3rd ed., Springer-Verlag, New York.

James W. McAllister, (2003) Effective Complexity as a Measure of Information Content, Philosophy of Science, Vol. 70, No. 2, pp. 302-307.

J. J. Rissanen, (1978) Modeling by Shortest Data Description, Automatica, volume 14, no. 5, pg. 465-471.

J. J. Rissanen, (1989) Stochastic Complexity in Statistical Inquiry, World Scientific, Singapore.

R. Scha and R. Bod (1993) "Computationele Esthetica", Informatie en Informatiebeleid 11, 1 (1993), pp. 54-63.

N.K. Vereshchagin, Vitányi P.M.B., (2004) Kolmogorov's structure functions and model selection, IEEE Transactions on Information Theory, vol. 50, nr. 12, 3265–3290.

P. Vitányi, (2006) Meaningful information, IEEE Trans. Inform. 52:10, 4617 - 4626.

Wolpert, David H. and Macready, William (2007) Using self-dissimilarity to quantify complexity: Research Articles, Complexity, volume 12,number 3, pages 77–85.

A. Zellner (1988), Optimal information processing and Bayes' theorem, American Statistician 42, 278-284.

A. Zellner (2002), Information processing and Bayesian analysis, Journal of Econometrics 107, 41-50.