Predictability, Complexity and Learning

Ilya Nemenman

Co-authored with: William Bialek, Naftali Tishby

August 26, 2003

http://xxx.lanl.gov/abs/physics/0007070

Outline

- A curious observation.
- Why a new learning and complexity theory is needed?
- Why and how to use information theory?
- Predictive information and its properties.
- Calculating predictive information for different processes.
- Unique complexity measure through predictive information.
- Possible applications.

Entropy of words in a spin chain

$$S(N) = -\sum_{k=0}^{2^N-1} P_N(W_k) \log_2 P_N(W_k)$$

For this chain, $P(W_0) = P(W_1) = P(W_3) = P(W_7) = P(W_{12}) = P(W_{14}) = 2$, $P(W_8) = P(W_9) = 1$, and all other frequencies (probabilities) are zero. Thus, $S(4) \approx 2.95$ bits.

Entropy of 3 generated chains

- $J_{\mathrm{ij}} = \delta_{\mathrm{i,j+1}}$
- $J_{\mathrm{ij}} = J_0\, \delta_{\mathrm{i,j+1}}$, J_0 is taken

  at random from $\mathcal{N}(0, 1)$

  every 400000 spins
- J_{ij} is taken at random

  from $\mathcal{N}(0, \frac{1}{\mathrm{i-j}})$

  every 400000 spins

$1 \cdot 10^9$ spins total.

Entropy is extensive! It shows no distinction between the cases.

Subextensive component of the entropy

This component is usually neglected in physics and information theory.

Subextensive entropy shows a qualitative distinction between the cases! What is the significance of this difference?

Problems in learning and complexity theories

- many frameworks to study learning
  - all very specialized
- many frameworks to study complexity
  - probabilistic or deterministic?
  - description length (Kolmogorov-style complexities) — but are all bits relevant?
- no clear connection between learning and complexity
- over-universal complexity definitions
- complexity must be zero for a completely random signal, and some measures get it wrong

There is very little known about connections between various views on learning and complexity.

We need a *universal* paradigm created, of which all studied problems are special cases.

We base this approach on the notion of predictability.

Why predictability?

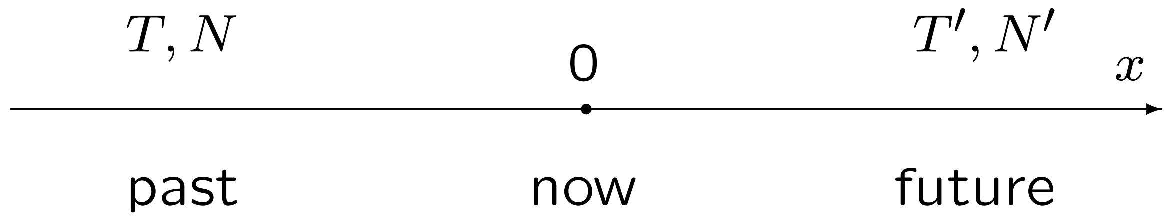
- we learn (estimate parameters, extrapolate, classify, ...) not for the sake of learning; the problem of learning is to *generalize* and *predict* from training examples, and estimation of parameters is only an intermediate step
- nonpredictive features in any signal are useless since we observe *now* and react in the *future*
- high predictability means more features to predict, not easier prediction — this is a problem of intuitively higher complexity
- it is impossible to predict a totally random string, so if complexity is based on predictability, for such a string it is zero

Quantifying predictability

- learning is accrual of *information*
- Shannon's information theory is *the only* nonmetric way to quantify information

Thus we will use information theory to study predictability and will define *predictive information* as
the information that the observed data provides about the data that is coming.

Definitions



$$
\begin{aligned}
\mathcal{I}_{\text{pred}}(T,T') &= \left\langle \log_2 \left[\frac{P(x_{\text{future}}|x_{\text{past}})}{P(x_{\text{future}})} \right] \right\rangle \\
&= S(T) + S(T') - S(T+T') \\
S(T) &= \mathcal{S}_0 \cdot T + S_1(T)
\end{aligned}
$$

extensive component cancels in predictive information

predictability is a deviation from extensivity!

$$
I_{\text{pred}}(T) \equiv \mathcal{I}_{\text{pred}}(T,\infty) = S_1(T)
$$

Properties of $I_{\text{pred}}(T)$

- $I_{\text{pred}}(T)$ is information, so $I_{\text{pred}}(T) \geq 0$
- $I_{\text{pred}}(T)$ is subextensive, $\lim_{T\to\infty} \frac{I_{\text{pred}}(T)}{T} = 0$
- diminishing returns, $\lim_{T\to\infty} \frac{I_{\text{pred}}(T)}{S(T)} = 0$
- prediction and postdiction are symmetric
- it relates to and generalizes many relevant quantities
  - learning: universal learning curves
  - complexity: complexity measures
  - coding: coding length

How can I_{pred} behave?

$\lim_{N\to\infty} I_{\text{pred}} = \text{const}$ no long-range structure

- simply predictable (periodic, constant, etc.) processes
- fully stochastic (Markov) processes

$\lim_{N\to\infty} I_{\text{pred}} = \text{const} \times \log_2 N$ precise learning of a fixed set of parameters

- learning finite-parameter densities
- analyzed as $I(N, \text{parameters}) = I_{\text{pred}}(N)$

$\lim_{N\to\infty} I_{\text{pred}} = \text{const} \times N^{\xi}$ learning more features as N grows

- learning continuous densities
- not well studied

Problem setup

$Q(x\|\boldsymbol{\alpha})$	probability density function for $\vec{x}$ parameterized by unknown parameters $\boldsymbol{\alpha}$
$\dim \boldsymbol{\alpha} = K$	dimensionality of $\boldsymbol{\alpha}$, may be infinite
$\mathcal{P}(\boldsymbol{\alpha})$	prior distribution of parameters
$\vec{x}_1 \cdots \vec{x}_\mathsf{N}$	random samples from the distribution

$$
\begin{aligned}
P(\vec{x}_1, \vec{x}_2, \cdots, \vec{x}_\mathsf{N} | \boldsymbol{\alpha}) &= \prod_{\mathsf{i}=1}^{\mathsf{N}} Q(\vec{x}_\mathsf{i} | \boldsymbol{\alpha}) \\
P(\vec{x}_1, \vec{x}_2, \cdots, \vec{x}_\mathsf{N}) &= \int d^K \alpha \mathcal{P}(\boldsymbol{\alpha}) \prod_{\mathsf{i}=1}^{\mathsf{N}} Q(\vec{x}_\mathsf{i} | \boldsymbol{\alpha}) \\
S(\vec{x}_1, \vec{x}_2, \cdots, \vec{x}_\mathsf{N}) \equiv S(N) &= -\int d\vec{x}_1 \cdots d\vec{x}_\mathsf{N}\, P(\{\vec{x}_\mathsf{i}\}) \log_2 P(\{\vec{x}_\mathsf{i}\})
\end{aligned}
$$

Separating the extensive term

$$
\begin{aligned}
S(N) &= -\int d^K\bar{\boldsymbol{\alpha}}\mathcal{P}(\bar{\boldsymbol{\alpha}}) \left\{ d^N\vec{x} \prod_{\mathsf{j}=1}^{\mathsf{N}} Q(\vec{x}_{\mathsf{j}}|\bar{\boldsymbol{\alpha}}) \log_2 \int d^K\alpha\mathcal{P}(\boldsymbol{\alpha}) \prod_{\mathsf{i}=1}^{\mathsf{N}} Q(\vec{x}_{\mathsf{i}}|\boldsymbol{\alpha}) \right\} \\
&= -\int d^K\bar{\boldsymbol{\alpha}}\mathcal{P}(\bar{\boldsymbol{\alpha}}) \left\{ d^N\vec{x} \prod_{\mathsf{j}=1}^{\mathsf{N}} Q(\vec{x}_{\mathsf{j}}|\bar{\boldsymbol{\alpha}}) \right. \\
&\qquad \left. \times \log_2 \prod_{\mathsf{j}=1}^{\mathsf{N}} Q(\vec{x}_{\mathsf{j}}|\bar{\boldsymbol{\alpha}}) \int d^K\alpha\mathcal{P}(\boldsymbol{\alpha}) \overbrace{\prod_{\mathsf{i}=1}^{\mathsf{N}} \left[\frac{Q(\vec{x}_{\mathsf{i}}|\boldsymbol{\alpha})}{Q(\vec{x}_{\mathsf{i}}|\bar{\boldsymbol{\alpha}})}\right]}^{\exp[-N\mathcal{E}_N(\boldsymbol{\alpha};\{\vec{x}_{\mathsf{i}}\})]} \right\}
\end{aligned}
$$

This separates $S(N)$ into the extensive and the subextensive terms

$$
\begin{aligned}
\mathcal{S}_0 &= \int d^K\alpha\mathcal{P}(\boldsymbol{\alpha}) \left[-\int d^Dx Q(\vec{x}|\boldsymbol{\alpha}) \log_2 Q(\vec{x}|\boldsymbol{\alpha}) \right], \\
S_1(N) &= -\int d^K\bar{\alpha}\, d^N\, \vec{x}_{\mathsf{i}}\mathcal{P}(\bar{\boldsymbol{\alpha}}) \log_2 \left[\int d^K\alpha P(\boldsymbol{\alpha}) \mathrm{e}^{-N\mathcal{E}_N} \right]
\end{aligned}
$$

Annealed approximation

Under some conditions we may have

$$
\begin{aligned}
\psi(\boldsymbol{\alpha}, \bar{\boldsymbol{\alpha}}; \{x_i\}) &\equiv \underbrace{\mathcal{E}_N(\boldsymbol{\alpha}; \{\vec{x}_i\})}_{\text{quenched energy}} - \underbrace{D_{KL}(\bar{\boldsymbol{\alpha}}||\boldsymbol{\alpha})}_{\text{annealed energy}} \\
&\equiv -\frac{1}{N}\sum_{i=1}^{N} \ln\left[\frac{Q(\vec{x}_i|\boldsymbol{\alpha})}{Q(\vec{x}_i|\bar{\boldsymbol{\alpha}})}\right] + \int d\vec{x} Q(\vec{x}|\bar{\boldsymbol{\alpha}}) \ln\left[\frac{Q(\vec{x}|\boldsymbol{\alpha})}{Q(\vec{x}|\bar{\boldsymbol{\alpha}})}\right] \\
&\stackrel{\sim}{\to} 0
\end{aligned}
$$

$$
S_1(N) \stackrel{\sim}{\to} S_1^{(a)}(N) \equiv -\int d^K\bar{\alpha}\mathcal{P}(\bar{\boldsymbol{\alpha}}) \underbrace{\log_2 \overbrace{\int d^K\alpha P(\boldsymbol{\alpha}) e^{-ND_{KL}}}^{\text{annealed partition function, } Z(\bar{\boldsymbol{\alpha}};N)}}_{\text{annealed free energy, } F(\bar{\boldsymbol{\alpha}};N)}
$$

Density of states

We can rewrite the partition function

$$
\begin{aligned}
Z(\bar{\boldsymbol{\alpha}}; N) &= \int dD \rho(D; \bar{\boldsymbol{\alpha}}) \exp[-ND] \\
\rho(D; \bar{\boldsymbol{\alpha}}) &= \int d^K \alpha \mathcal{P}(\boldsymbol{\alpha}) \delta[D - D_{\mathsf{KL}}(\bar{\boldsymbol{\alpha}} || \boldsymbol{\alpha})] \\
\int dD \rho(D; \bar{\boldsymbol{\alpha}}) &= \int d^K \alpha \mathcal{P}(\boldsymbol{\alpha}) = 1
\end{aligned}
$$

The density ρ could be very different for different targets.

Thus learning is annealing at decreasing temperature; properties of predictive information (and learning) almost always depend on $D = 0$ behavior of the density.

Power–law density function

For this case:

$$\begin{aligned} \rho(D \to 0; \bar{\boldsymbol{\alpha}}) &\approx A(\bar{\boldsymbol{\alpha}}) D^{(d-2)/2} \\ S_1^{(\mathrm{a})} &\approx \frac{d}{2} \log_2 N \end{aligned}$$

If $d = d(\bar{\boldsymbol{\alpha}})$, then we can get non half–integer coefficients in front of the logarithm term.

- this behavior is known in MDL and other literature
- speed of approach to this asymptotics is rarely investigated

Examples of the logarithmic predictive information

- Finite parameter models, $\dim \boldsymbol{\alpha} = K$. Then for $\boldsymbol{\alpha} \approx \bar{\boldsymbol{\alpha}}$ and for *sound* parameterization

$$
\begin{aligned}
D_{\mathsf{KL}}(\bar{\boldsymbol{\alpha}} || \boldsymbol{\alpha}) &\approx \frac{1}{2} \sum_{\mu\nu} (\bar{\alpha}_\mu - \alpha_\mu) \mathcal{F}_{\mu\nu} (\bar{\alpha}_\nu - \alpha_\nu) + \cdots \\
\rho(D \to 0; \bar{\boldsymbol{\alpha}}) &\approx \mathcal{P}(\bar{\boldsymbol{\alpha}}) \frac{2\pi^{K/2}}{\Gamma(K/2)} (\det \mathcal{F})^{-1/2} D^{(K-2)/2} \\
\mathcal{F} &\text{ — Fisher information matrix}
\end{aligned}
$$

To avoid complications with *soundness*, we can *define* the phase space dimensionality of the model family through the exponent of the density function.

- Finite parameter Markov process, learn $Q(\vec{x}_1 \cdots \vec{x}_\mathsf{N}|\boldsymbol{\alpha})$. If energy is extensive,

$$D_{\mathsf{KL}}\left[Q(\{\vec{x}_\mathsf{i}\}|\bar{\boldsymbol{\alpha}})||Q(\{\vec{x}_\mathsf{i}\}|\boldsymbol{\alpha})\right] \;\to\; N\mathcal{D}_{\mathsf{KL}}\left(\bar{\boldsymbol{\alpha}}||\boldsymbol{\alpha}\right) + o(N)\,.$$

  and extensive term is replaced by

$$\begin{aligned} S\left[\{\vec{x}_\mathsf{i}\}|\boldsymbol{\alpha}\right] &\equiv -\int d^N\vec{x}\, Q(\{\vec{x}_\mathsf{i}\}|\boldsymbol{\alpha}) \log_2 Q(\{\vec{x}_\mathsf{i}\}|\boldsymbol{\alpha}) \\ &\to N\mathcal{S}_0 + \mathcal{S}_0^*; \qquad \mathcal{S}_0^* = \frac{K'}{2}\log_2 N \end{aligned}$$

  then

$$S_1^{(\mathsf{a})}(N) = \frac{K+K'}{2}\log_2 N$$

Predictive information does not distinguish predictability coming from unknown parameters and from intrinsic long–range correlations. This is similar to describing physical systems with correlations using order parameters.

Essential singularity in the density function

As $d \to \infty$ we may imagine the following behavior

$$
\begin{aligned}
\rho(D \to 0; \bar{\boldsymbol{\alpha}}) &\approx A(\bar{\boldsymbol{\alpha}}) \exp\left[-\frac{B(\bar{\boldsymbol{\alpha}})}{D^{\mu}}\right], \quad \mu > 0 \\
C(\bar{\boldsymbol{\alpha}}) &= [B(\bar{\boldsymbol{\alpha}})]^{1/(\mu+1)} \left(\frac{1}{\mu^{\mu/(\mu+1)}} + \mu^{1/(\mu+1)}\right) \\
S_1^{(\mathrm{a})}(N) &\approx \frac{1}{\ln 2} \langle C(\bar{\boldsymbol{\alpha}}) \rangle_{\bar{\boldsymbol{\alpha}}} N^{\mu/(\mu+1)}
\end{aligned}
$$

- finite parameter model with increasing number of parameters $K \sim N^{\mu/(\mu+1)}$; $S_1(N) \sim N^{\mu/\mu+1}$, not $S_1(N) \sim \frac{N^{\mu/\mu+1}}{2} \log N$
- as $\mu \to \infty$ complexity grows and then vanishes to the leading order when $S_1^{(\mathrm{a})}$ becomes extensive

Example of the power–law I_{pred}

Learning a nonparametric (infinite parameter) density $Q(x) = 1/l_0 \mathrm{e}^{-\phi(x)}$, $x \in [0, L]$, with some smoothness constraints (Bialek, Callan, and Strong 1996).

$$\begin{aligned}
\mathcal{P}[\phi(x)] &= \frac{1}{\mathcal{Z}} \exp\left[-\frac{l}{2}\int dx \left(\frac{\partial\phi}{\partial x}\right)^2\right] \delta\left[\frac{1}{l_0}\int dx\, \mathrm{e}^{-\phi(x)} - 1\right] \\
\rho(D \to 0; \bar{\phi}) &= A[\bar{\phi}(x)] D^{-3/2} \exp\left(-\frac{B[\bar{\phi}(x)]}{D}\right) \\
S_1^{(\mathrm{a})}(N) &\approx \frac{1}{2\ln 2}\sqrt{N}\left(\frac{L}{l}\right)^{1/2}
\end{aligned}$$

- increasing number of ‘effective parameters’ (bins) of adaptive size $\sim \sqrt{l/NQ(x)}$
- heuristic arguments for the dimensionality ζ and the smoothness exponent η give $S_1(N) \sim N^{\zeta/2\eta}$ — demonstrates a crossover from complexity to randomness

Which complexity do we want to define?

- complexity of dynamics that generates a time series (not computational or descriptive complexity); thus it must be zero for totally random and for easily predictable processes
- usable for Occam–style punishment in statistical inference
- expressible in conventional physical terms
- must be attached to an ensemble, not a single realization

Ensemble property?

All pictures could be just random, but we do not perceive them this way. Ensemble is implicit!

Unique measure of complexity!

Complexity measure must be:

- some kind of entropy (we proclaim Shannon's postulates)
  - monotonic in N for N equally likely signals
  - additive for statistically independent signals
  - a weighted sum of measure at branching points if measuring a leaf on a tree
- reparameterization, quantization invariant, thus subextensive
- invertible temporally local transformations (e. g., $x_k \to x_k + \xi x_{k-1}$—measuring device with inertia) and prior insensitive

The divergent subextensive term measures complexity uniquely!

What's next?

- separating predictive information from non–predictive using the 'relevant information' technique
- reflection to physics — finding order parameters for phase transitions using behavior of the predictive information
- reflection to biology — large expansion from receptors to primary sensory cortices may be due to efficient representation of predictive information, not current state of the world
- reflection to psychology — experiments on learning distributions and language (power law complexity class) by humans; what expectations of the world do we have?
- reflection to statistics
  - nonparametric models may be simpler then finite parameter ones (relevant to biology)
  - predictive information is the property of the data (nonparametric extension of the MDL principle)

Summary

We have built a generalizing and unique theory of learning and complexity.