Introduction to Information-Theoretic Learning: Shannon's Entropy, Renyi's Entropy, and Minimum Error Entropy

based on
Principe, J. C., Information theoretic learning: Renyi's entropy and kernel perspectives, *Springer*, 2010

Fabio Massimo Zennaro

Information Theoretic Learning

What is Information Theoretic Learning (ITL)? A *framework* for learning relying on the theory of information.

Centrality of the idea of *information* to construct models.

Why is ITL relevant?

- Grounded in information theory
- Lightweight assumptions
- Mathematically elegant
- Practically applicable

Outline

Outline of the presentation

1. Entropy
2. Renyi's Entropy
3. Minimum Error Entropy

Information (I)

Information as a (scalar) quantitative way to assess *uncertainty* in relationship to random events.

A *measure of information* $I()$ must satisfy the following requirements:

- **Extensive**: given two mutually independent event A and B then $I(A \text{ and } B) = I(A) + I(B)$
- **Zero information**: given an event A such that $P(A) = 1$ then $I(A) = 0$

Information (II)

Information of an event A:

$$I(A) = -\log P(A)$$

- **Extensive**: $I(A \text{ and } B) = -\log P(A \text{ and } B) = -\log P(A)P(B) = -\log P(A) - logP(B) = I(A) + I(B)$
- **Zero information**: $I(A) = -\log P(A) = -\log 1 = 0$

Entropy

Measure of information for a random variable is *Shannon's entropy*:

$$\begin{aligned} H[X] = E[I(X)] &= -\sum_i p(x_i) \log p(x_i) \\ &= -\int_{\mathcal{X}} p(x) \log p(x) \, \mathrm{d}x \end{aligned}$$

- *Unexpectedness weighted by probability*
- Physical analogy (*Gibbs Entropy*)
- Useful mathematical properties (*continuity, symmetry, recursivity...*)
- Statistical descriptor

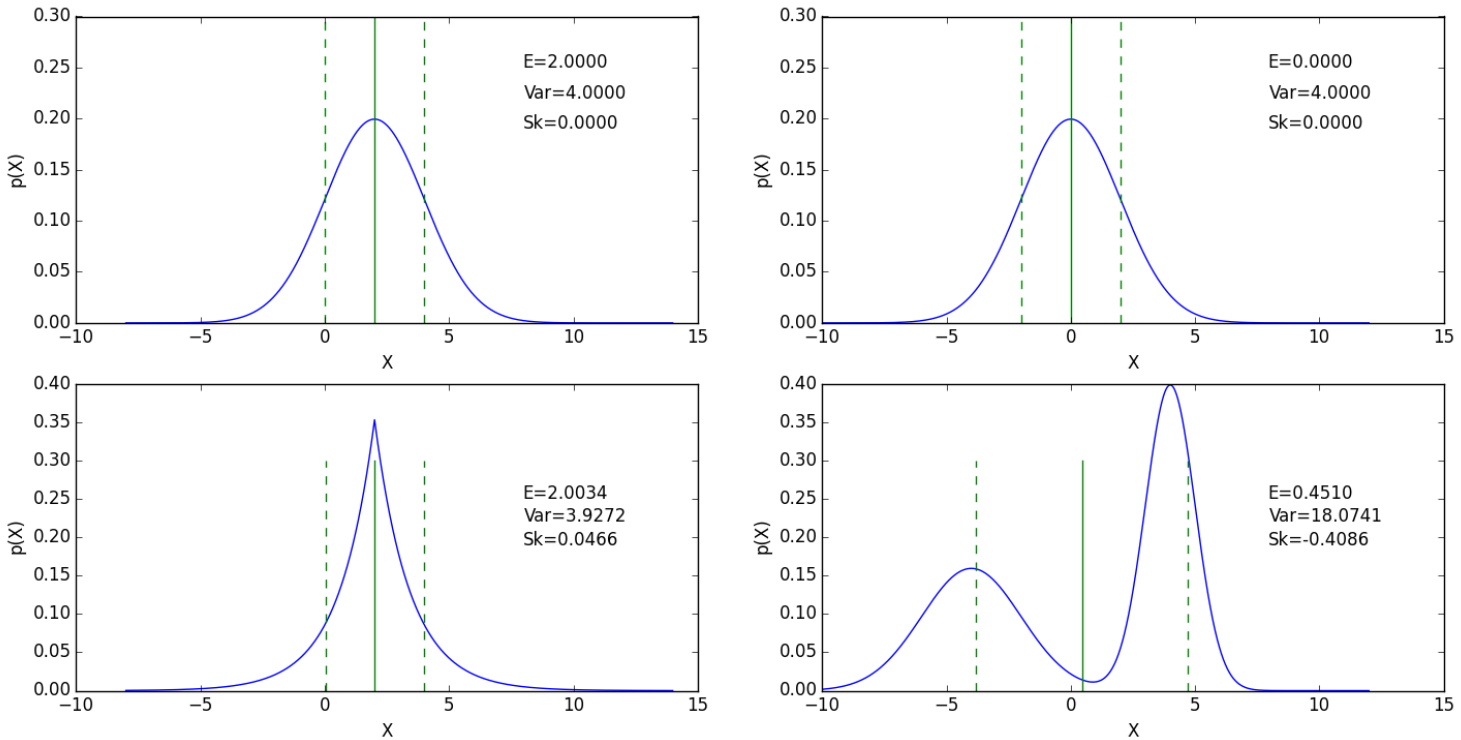
Statistical Interpretation

Uncertainty is modeled through *PDF* or *PMF*

However, working with functions is challenging.

We synthetize PDF via **statistical descriptors**, such as *statistical moments*...

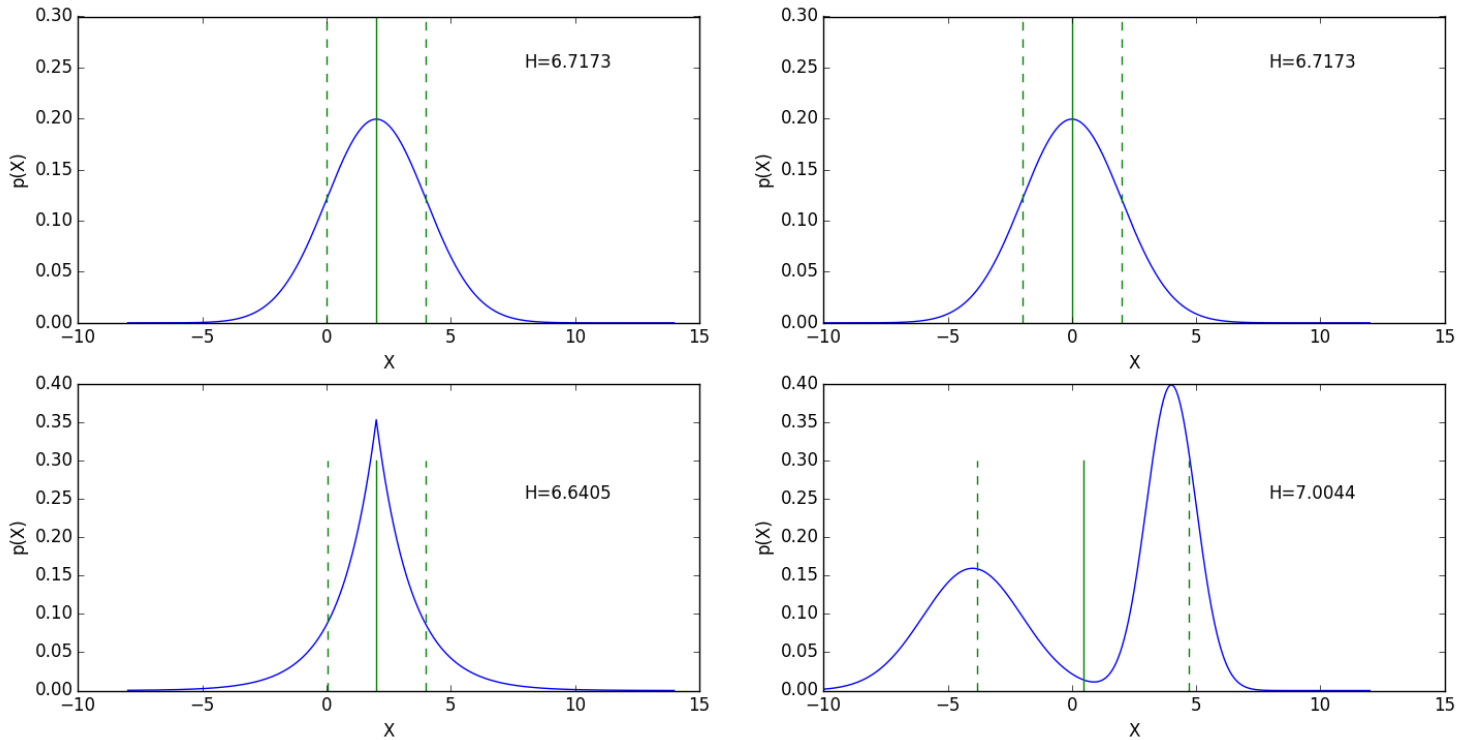
Statistical Description via Moments



$$M_n[X] = \int_{\mathcal{X}} x^n p(x)\,\mathrm{d}x$$

- Dependency on the domain
- Implicit assumptions
- Theoretical need for an infinite number of moments

Statistical Description via Entropy



$$H[X] = -\int_{\mathcal{X}} p(x) \log p(x) \, \mathrm{d}x$$

- Domain-agnostic
- Better quantification of the volume spanned by the pdf
- Harder to estimate

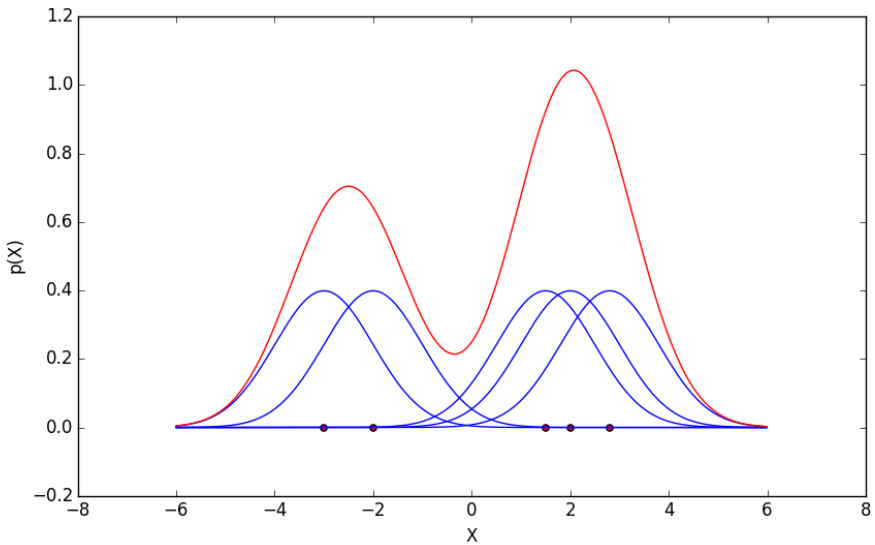
Kernel Density Estimation

Estimating entropy requires estimating pdf.

One approach is **Parzen Kernel Density Estimation (KDE)**:

$$\hat{p}(x) = \frac{1}{N\sigma} \sum_i \kappa \left(\frac{x - x_i}{\sigma} \right)$$

Often, the kernel is a *Gaussian* kernel



Entropy Estimation

Using KDE, the *entropy estimator* is:

$$\begin{aligned}\hat{H}[X] &= -\int_{\mathcal{X}} \hat{p}(x) \log \hat{p}(x) \\ &= -\int_{\mathcal{X}} \left[\frac{1}{N\sigma}\sum_i \kappa\left(\frac{x-x_i}{\sigma}\right)\right] \log \left[\frac{1}{N\sigma}\sum_i \kappa\left(\frac{x-x_i}{\sigma}\right)\right]\end{aligned}$$

- Data $\rightarrow$ PDF estimation $\rightarrow$ Integral estimation
- Composed bias and variance
- In general, unreliable in high dimensions

Recap (1/3)

Where are we?

- We do modelling using PDFs/PMFs...
- ... we use descriptors to synthetize PDFs/PMFs ...
- ... information theory provides powerful descriptors ...
- ... but these descriptors are hard to estimate.

Where next?

- We look for alternative forms of the same descriptors...

Alternative Entropy

Back to the original problem of finding an information function $I()$ that is *extensive* and *zero-information*.

Let the information of an event A be:

$$I(A) = -\log P(A)$$

First solution for information of a random variable X was:

$$H[X] = \sum_i p(x_i) I(x_i)$$

Alternative Entropy

A *more general* formulation for a random variable X is:

$$H[X] = g^{-1}\left(\sum_i p(x_i) g\left(I\left(x_i\right)\right)\right)$$

where $g()$ is monotonic function with inverse $g^{-1}()$.

- If $g(x) = cx$, for a real constant c, then $H[X]$ is Shannon's entropy.

Renyi's Entropy

If we take:

$$g(x) = c2^{(1-\alpha)x}$$

then we get **Renyi's entropy**

$$H_\alpha[X] = \frac{1}{1-\alpha} \log\left(\sum_i p^\alpha(x_i)\right)$$

- Renyi's entropy is parametric in α (*spectrum of Renyi's information*).
- Renyi's entropy *moves out* the logarithm.
- Renyi's entropy retains many properties of Shannon's entropy (*continuity, symmetry...*)

Entropies

Renyi's entropy subsumes other entropies:

- $\alpha = 0$ defines *Hartley's Entropy*:

$$H_0[X] = \log |X|$$

- $\lim \alpha \to 1$ defines *Shannon's Entropy*:

$$H_1[X] = -\sum_i p(x_i) \log p(x_i)$$

- $\alpha = 2$ defines **Quadratic Entropy** or *Collision Entropy*:

$$H_2[X] = -\log \sum_i p^2(x_i)$$

- $\lim \alpha \to \infty$ defines *Min-entropy*:

$$H_\infty[X] = -\log (\max p(x_i))$$

Information Potential

Let us isolate the log argument of $H_\alpha[X]$ and let us call it **information potential**:

$$H_\alpha = \frac{1}{1-\alpha} \log \underbrace{\left(\sum_i p^\alpha (x_i) \right)}_{\substack{V_\alpha [X] \\ IP_\alpha [X]}}$$

We can rewrite Renyi's entropy as:

$$H_\alpha [X] = \frac{1}{1-\alpha} \log V_\alpha [X] = -\log \left(\sqrt[\alpha-1]{V_\alpha [X]} \right)$$

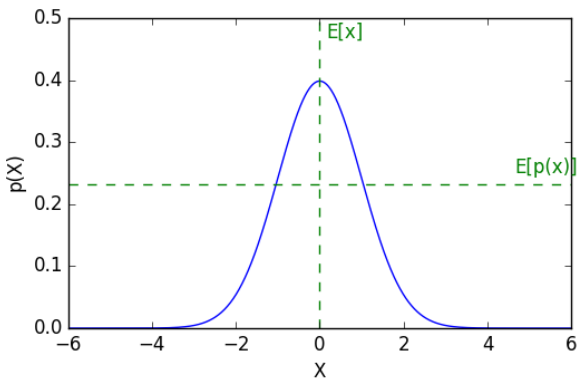
Renyi's Quadratic Information Potential

Let us focus on Renyi's Quadratic Entropy:

$$\begin{aligned} H_2[X] &= -\log \sum_i p^2(x_i) \\ &= -\log \int_{\mathcal{X}} p^2(x_i) \end{aligned}$$

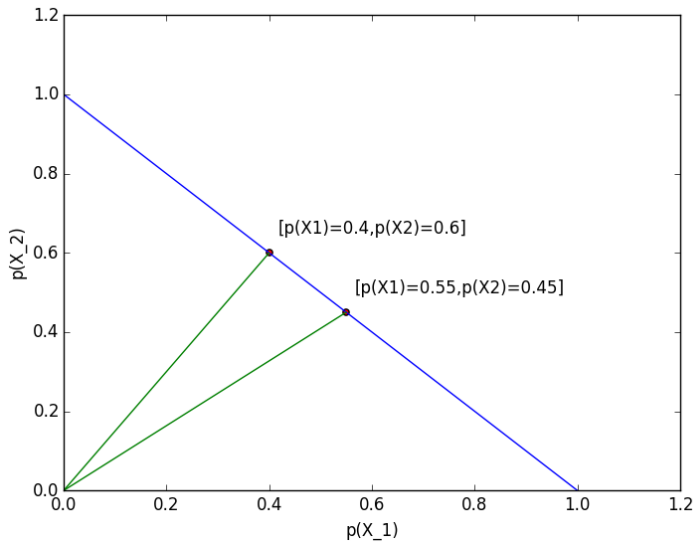
The *information potential* is the *expected value of* $p(x_i)$:

$$IP_2[X] = E[p(x_i)]$$



Geometric Interpretation

$$H_2[X] = -\log\left(\sqrt{V_2[X]}\right) = -\log\left(\sqrt{\sum_i p^2(x_i)}\right)$$



Renyi's Quadratic Entropy identifies probability distributions on a *probability simplex*.

Estimation of Renyi's Quadratic Entropy (I)

Now, estimating Renyi's Quadratic Entropy reduces to *estimating Renyi's Quadratic IP*.

That is, *estimating the expected value of* $p(x_i)$:

$$\begin{aligned}
\hat{H}_2[X] &= -\log \int_{\mathcal{X}} E[\hat{p}(x_i)]\,\mathrm{d}x \\
&= -\log \int_{\mathcal{X}} \hat{p}(x_i)\,\hat{p}(x_i)\,\mathrm{d}x \\
&= -\log \int_{\mathcal{X}} (\hat{p}(x_i))^2\,\mathrm{d}x
\end{aligned}$$

Estimation of Renyi's Quadratic Entropy (II)

Let us use *Parzen Kernel Density Estimation* with a *Gaussian kernel* to estimate $p(x_i)$.

$$\begin{aligned}
\hat{H}_2[X] &= -\log \int_{\mathcal{X}} (\hat{p}(x_i))^2 \, \mathrm{d}x \\
&= -\log \int_{\mathcal{X}} \left(\frac{1}{N} \sum_i G_\sigma (x - x_i) \right)^2 \mathrm{d}x \\
&= -\log \frac{1}{N^2} \int_{\mathcal{X}} \left(\sum_i \sum_j G_\sigma (x - x_i) \cdot G_\sigma (x - x_j) \right) \mathrm{d}x
\end{aligned}$$

Estimation of Renyi's Quadratic Entropy (III)

Property of Gaussians: the integral of the product of two Gaussians can be computed from the Gaussian at the difference of the arguments and with a variance given by the sum of the original variances.

$$\begin{aligned}\hat{H}_2[X] &= -\log \frac{1}{N^2} \int_{\mathcal{X}} \left(\sum_i \sum_j G_\sigma(x - x_i) \cdot G_\sigma(x - x_j) \right) \mathrm{d}x \\ &= -\log \frac{1}{N^2} \sum_i \sum_j G_{\sigma\sqrt{2}}(x_j - x_i)\end{aligned}$$

Quadratic Information Potential Estimator:

$$\hat{V}_2[X] = \frac{1}{N^2} \sum_i \sum_j G_{\sigma\sqrt{2}}(x_j - x_i)$$

Pros of Estimating Renyi's Quadratic Entropy

Quadratic Information Potential Estimator:

$$\hat{V}_2[X] = \frac{1}{N^2} \sum_i \sum_j G_{\sigma\sqrt{2}}(x_j - x_i)$$

- *Original problem*: data $\rightarrow$ PDF estimation $\rightarrow$ Integral estimation
  New problem: data $\rightarrow \hat{V}_2[X]$
- Estimating the expected value of $p(x)$ is easier than estimating $p(x)$
- The estimator $\hat{V}_2[X]$ can be directly computed from data
- Bias and variance depending only on $\hat{V}_2[X]$

Cons of Estimating Renyi's Quadratic Entropy

Quadratic Information Potential Estimator:

$$\hat{V}_2[X] = \frac{1}{N^2} \sum_i \sum_j G_{\sigma\sqrt{2}}(x_j - x_i)$$

- Estimating $\hat{V}_2[X]$ depends on *pairs of samples*, therefore its complexity is $O(N^2)$
- Estimating $\hat{V}_2[X]$ requires setting a *kernel bandwidth*
- Still unreliable in high dimension

Physical Parallelism

ML	ITL	Physics
Data	Information Particle	Mass Particle
Kernel	Information Potential Field	Gravitational Potentential Field
Sum of Kernel	Averaged Informational Field	Total Gravitational Field
Derivative of Kernel	Information Force	Gravitational Force

Average Potential Field: $\hat{V}_2[X]$

Potential Field on x_i: $\hat{V}_2[x_i]$

Potential Field on x_i due to x_j: $\hat{V}_2[x_i, x_j]$

Information Force on x_i: $\hat{F}_2[x_i] = \frac{\partial}{\partial x_i}\hat{V}_2[x_i]$

Information Force on x_i due to x_j: $\hat{F}_2[x_i, x_j]$

Recap (2/3)

Where are we?

- We found alternative information theoretic descriptors...
- ... we realized that one of these descriptors has interesting properties ...
- ... and we can (relatively) easily estimate it.

Where next?

- We consider applying this descriptor to learning...

Error function and Error Entropy Criterion

Learning is usually driven by the minimization of an **error function** describing the distribution of the errors.

Error Entropy Criterion: Minimizing the information contained in the error distribution equates to maximize the information in the model.

This means *minimizing the entropy* of the error distribution.

MSE

A common loss function is the *mean square error*:

$$\mathcal{L}_{MSE}(e(w)) = E\left[e^2(w)\right]$$

- Learn by *minimizing* MSE.
- MSE is an estimator of the *error variance*.
- If the errors are Gaussian, this loss function is optimal.
- If the errors are not Gaussain, this loss function is sub-optimal. Some alternatives have been proposed (least mean fourth, L_p powers)

MEE

We can now use the estimator of the information potential for the loss function:

$$\mathcal{L}_{MEE}\left(e(w)\right) = H_2\left[e(w)\right]$$

- Learn by *minimizing* $\hat{H}_2$ or by *maximizing* $\hat{V}_2$
- No assumptions on the errors
- Shape the error distribution into a delta function
- Move information from e (error) to w (model)

Learn by MEE

Learn the model by *gradient descent*:

$$\frac{d}{dw} V[e(w)]$$

- Error signal $e(w)$ is low-dimensional
- Derivative explicitly affect the overall shape of the error PDF
- Apply gradient descent in any of its flavours (*momentum*, *conjugate*, *quasi-Newtonian*..)

Recap (3/3)

Where are we?

- We implemented a loss function based on an information theoretic estimator.

Where next?

- *Efficient computation* of information potential
- Derivation of Renyi's *mutual information* and *pdf distances*
- Extension of correlation into *correntropy*
- Integration of this theoretical framework with *RKHS*

Thanks

Thank you for listening!