Probability, Information Theory, and Physics

From Information Theory to Physics
and from Physics to Deep Learning

Sanha Cheong
sanha@stanford.edu

Department of Physics
Stanford University

July 11, 2017

Agenda

Today's main goal:

- Introduce basic concepts used in *information theory*
- Go through some examples to demonstrate the capabilities of information theory
- Relate the above to well-known results in physics—particularly, some *Statistical Mechanics*
- Use physics to motivate some *Deep Learning* algorithms

For this, the prerequisites are:

- Understanding of calculus
- Exposure to basic ideas in probabilities and statistics
- Physics will be derived, but previous exposure will help

Probability measures how likely an event is to occur or a proposition be true. In other words, it represents *uncertainty*.

There is some subtlety here, however...

- Frequentist: relative *occurrence* of the event under consideration after repeated (infinite) trials
- Bayesian: the *confidence* in a belief or a prediction

This is a very interesting and important debate, but we will use 'probability' interchangeably.

Mathematically, the probability of a state x in the set of all possibilities X is denoted as $P_X(x)$ and $0 \leq P_X(x) \leq 1$. Probability density $p_X(x)$ is the generalization of this concept to continuous variable (uncountable set of possibilities) and is unbound.

Quantifying Information with Probability

Property 1: Range & Limits

The statement "sun rose from the east this morning" doesn't really mean much; it has *zero* information. However, a surprising/rare event holds *a lot* of information. Mathematically,

$$P(x) \to 1^- \implies I(x) \to 0^+$$
$$P(x) \to 0^+ \implies I(x) \to \infty$$

Property 2: Independence $\iff$ Additivity

Suppose that two events x and y are *independent*. When we learn that x *and* y happened, the information gained must be a *sum* of information held by each. Hence,

$$p(x, y) = p(x)p(y) \implies I(x, y) = I(x) + I(y)$$

Self-information & Shannon Entropy

In addition, we assume that $I(x)$ is continuous. Then, such a function is *unique* (up to a constant > 0). Hence, we define:

$$I(x) \equiv -\log(P(x))$$

and call it the *self-information* (also called surprisal) of an event x. Note that the base of the log is irrelevant.

Another crucial quantity for today is the *Shannon entropy*:

$$\mathcal{H}[P] \equiv \mathbb{E}[I(x)] = -\sum_{x \in X} P(x) \log P(x)$$

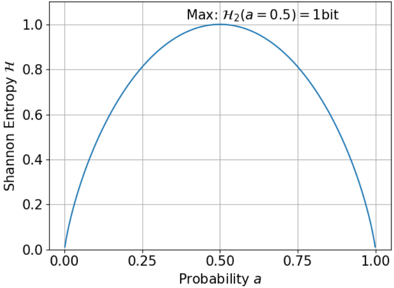
$$\mathcal{H}[p] \equiv \mathbb{E}[I(x)] = -\int_X p(x) \log p(x) \, \mathrm{d}x$$

and we define $0 \log 0 = 0$ (continuous extension).

(The generalization to continuous variables is sometimes called the *differential entropy* and has some caveats.)

Consider a *Bernoulli process* (i.e., coin flip) such that $X = \{0, 1\}$, $P(x = 0) = a$, and $P(x = 1) = 1 - a$. Then:

$$\mathcal{H}_2 = -a \log_2 a - (1-a) \log_2 (1-a)$$



Moreover, for any discrete distribution $P(x)$ on $X = \{x_1, ..., x_n\}$,

$$\mathcal{H}[P] \leq \log n$$

and equality holds if and only if P is a *uniform* distribution. In other words, for a discrete distribution, the entropy is *maximized* when it is uniform.

Now, let x be a *continuous* variable taking values from an interval $X \subset \mathbb{R}$ with a finite total length ℓ. Then, similarly, the uniform distribution

$$p : x \in X \longmapsto \frac{1}{\ell}$$

has the *maximum* entropy: $\mathcal{H}[p] = \log \ell$.

The proof is analogous to that of the discrete case, but replace all $\sum\limits_{i=1}^{n}$ to $\int\limits_{X} \mathrm{d}x$.

Hence, if there are *no other constraints*, the probability distribution over X with maximum entropy is a uniform distribution.

... but

so what?

We just re-derived the principle of *indifference* (equal a priori probability), which is almost common sense and has been known for long time. Are we doing anything *new*?

Yes. We can re-formulate the same (equilibrium) statistical mechanics differently (with less assumption)!

Statistical Mechanics with Maxwell, Boltzmann, and Gibbs

Equal a priori probability & physical knowledge
$\Rightarrow$ Thermodynamics & statistical mechanics
e.g. large # of degree of freedoms, microstates, etc.

Principle of Maximum Entropy states:

Given some testable information, the probability distribution that best represents our current knowledge is the one that maximizes (Shannon) entropy.

Applications: *equilibrium statistical mechanics*, coding theory (FEC), Bayesian inference, *deep learning*, etc.

MaxEnt Statistical Mechanics

Edwin T. Jaynes, Physical Review (1957)

Principle of Maximum Entropy

$\Rightarrow$ physical results as statistical inference

Known macroscopic physical quantities are merely constraints to the entropy maximization problem.

Known Mean μ and Variance σ^2

As a first 'non-trivial' example, consider $p : \mathbb{R} \to \mathbb{R}$ with *extra information*: its mean μ and variance σ^2. i.e.,

$$g_1(p;x) = \int_{-\infty}^{\infty} p \,\mathrm{d}x - 1 = 0$$

$$g_2(p;x) = \int_{-\infty}^{\infty} xp \,\mathrm{d}x - \mu = 0$$

$$g_3(p;x) = \int_{-\infty}^{\infty} (x-\mu)^2 p \,\mathrm{d}x - \sigma^2 = 0$$

Then, consider:

$$\begin{aligned} F[p] &= \int_{-\infty}^{\infty} -p \log p \,\mathrm{d}x + \sum_i \lambda_i g_i(p;x) \\ &= \int_{-\infty}^{\infty} -p \log p + \lambda_1 p + \lambda_2 xp + \lambda_3 (x-\mu)^2 p \,\mathrm{d}x - (\text{cons.}) \end{aligned}$$

and define $\mathcal{L} \equiv -p \log p + \lambda_1 p + \lambda_2 xp + \lambda_3 (x-\mu)^2 p$.

Known Mean μ and Variance σ^2 (cont'd)

The entropy is *maximized* when:

$$\frac{\partial \mathcal{L}}{\partial p} = -1 - \log p + \lambda_1 + \lambda_2 x + \lambda_3 (x-\mu)^2 = 0\,.$$

$$\therefore\, p(x) = \exp\big(\lambda_1 - 1 + \lambda_2 x + \lambda_3 (x-\mu)^2\big)$$

Since $\int_{-\infty}^{\infty} p(x)\,\mathrm{d}x$ must be finite, $\lambda_2 = 0$ and $\lambda_3 < 0$. Re-defining the constants, we can re-write: $p(x) = C\exp\big(-b(x-\mu)^2\big)$. Then, the constraints require $C = \sqrt{\frac{b}{\pi}}$ and $b = \frac{1}{2\sigma^2}$. Therefore,

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

which is the *Gaussian distribution* with the specified mean and variance. Gaussian distribution is the *MaxEnt* distribution when the mean and the variance are known.

Multi-dimensional Gaussian Distribution

Now, consider a n-dimensional distribution $p : \mathbb{R}^N \to \mathbb{R}$. As in Example 2, we have constraints: the *mean* and the *covariance* are $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. Then, the *MaxEnt* distribution is the *multi-variate Gaussian*:

$$p(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n |\boldsymbol{\Sigma}|}} \exp\left[-\frac{1}{2}\left(\mathbf{x} - \boldsymbol{\mu}\right)^T \boldsymbol{\Sigma}^{-1} \left(\mathbf{x} - \boldsymbol{\mu}\right)\right] .$$

In particular, if x_i's are independent of each other,

$$\Sigma = \mathrm{diag}\left(\sigma_1^2, ..., \sigma_n^2\right)$$

and therefore the *MaxEnt* distribution becomes:

$$p(\mathbf{x}) = \left(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_i}\right) \exp\left[-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i - \mu_i}{\sigma_i}\right)^2\right].$$

Let's Get 'Physical'

Since this is a physics talk after all, let's consider a case found in the natural world. Consider a collection of independent point particles with a common mass m moving around *randomly and isotropically* with velocity $\mathbf{v}$.

Under the constraints $\mathbb{E}(\mathbf{v}) = \mathbf{0}$ and $\Sigma = \text{diag}(\sigma^2, \sigma^2, \sigma^2)$, the *MaxEnt* distribution over velocity is:

$$p(v_x, v_y, v_z) = \frac{1}{(2\pi\sigma^2)^{3/2}} \exp\left(-\frac{v_x^2 + v_y^2 + v_z^2}{2\sigma^2}\right) .$$

Empirically, however, it is more useful to consider a distribution over *speed*, not velocity:

$$v = \sqrt{v_x^2 + v_y^2 + v_z^2} \quad \text{and} \quad \mathrm{d}\mathbf{v} = v^2 \sin\theta \, \mathrm{d}v \, \mathrm{d}\theta \, \mathrm{d}\phi \, .$$

Let's Get 'Physical' (cont'd)

Integrating over all solid angle, we obtain:

$$p(v) = \frac{1}{(2\pi\sigma^2)^{3/2}} 4\pi v^2 \exp\left(-\frac{v^2}{2\sigma^2}\right) .$$

Note that σ^2 has dimension of v^2. Since $KE = \frac{1}{2}mv^2$ classically, it is useful to define:

$$\sigma^2 \sim \frac{KE}{m} \Longrightarrow \sigma^2 = \frac{\epsilon}{m}$$

where ϵ is some energy scale. (Thermodynamcially, $\epsilon = k_B T$.) Then, we retain the 'familiar' expression in physics:

$$p(v) = \sqrt{\left(\frac{m}{2\pi\epsilon}\right)^3} 4\pi v^2 \exp\left(-\frac{mv^2}{2\epsilon}\right)$$

which is the *Maxwell-Boltzmann Distribution* over speed.

Boltzmann Statistics

Let us go back to *discrete* probabilities. Consider a random variable $s \in S = \{s_1, ..., s_N\}$. Also, let there be a function $E : S \to \mathbb{R}$ and denote $E(s_i) \equiv E_i$. This time, let our constraint be on $\mathbb{E}(E)$, instead of $\mathbb{E}(s)$. i.e., $\sum_{i=1}^{N} P_i E_i = \langle E \rangle$.

The corresponding *MaxEnt* distribution is:

$$P_i = \frac{e^{-\beta E_i}}{\sum_{j=1}^{N} e^{-\beta E_j}}$$

which is often called the *Boltzmann Statistics* (distribution).

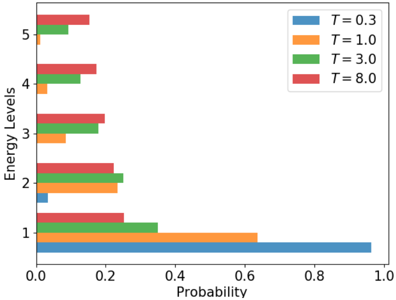
Physically, s_i's are discrete (quantum) states and E_i's their respective *energies*. β defines (absolute) *temperature*: $\beta \equiv 1/k_B T$.

Properties of Boltzmann Statistics

In particular, assuming no degeneracy (no different s_i's have same energy E_i),

$$0 < k_B T \ll \bar{E} \implies P_1 \approx 1$$

$$k_B T \gg \bar{E} \implies P_i \approx \frac{1}{N}$$



However, E need not be one-to-one. That is, E_i's are not necessarily distinct. Hence,

Physically *different* states can be *equally* likely.

e.g. a collection of gas molecules, magnetic moment (spin) alignment, pixelated images, etc.

‘Summarizing’ States and Statistical Learning

In other words, for statistical purposes, energy effectively *summarizes* states.

- Microscopic: each s_i, fully specified to the smallest scale, phenomenologically indistinguishable
- Global/Macroscopic: all s_i's with same E_i, only the total energy is specified, *meaningful* difference

This is at the core philosophy of *statistics* and also is a key challenge of *learning*.

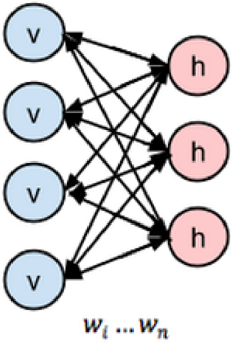
- Statistics: deducing meaningful overall features from many individual components
- Learning: must distinguish and extract meaningless and meaningful features in order to generalize properly

Our last application is a basic *deep learning* algorithm.

Consider a system with *many* degrees of freedom, and suppose we want to extract meaningful features from a data set.

For this, we use to binary vectors:

$$\text{Visible: } \mathbf{v} = [v_1, ..., v_n]$$
$$\text{Hidden: } \mathbf{h} = [h_1, ..., h_m]$$



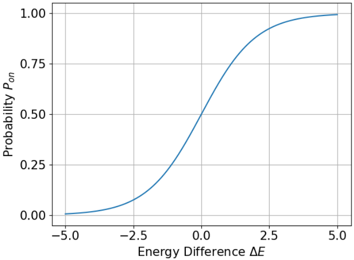
and define:

$$E(\mathbf{v}, \mathbf{h}) = -\sum_{i=1}^{n}\sum_{j=1}^{m} v_i w_{ij} h_j - \sum_{i=1}^{n} a_i v_i - \sum_{j=1}^{m} b_j h_j$$

Now, we need to *train* the energy function such that it has stably low values for a *meaningful* set of microscopic states.

First, we need to know how to *update* states. Consider a hidden unit h_j being on v.s. off. Then,

$$P_{\text{on}} = \frac{e^{-E_{\text{on}}/T}}{e^{-E_{\text{on}}/T} + e^{-E_{\text{off}}/T}} = \frac{1}{1 + \exp\left(-\frac{\Delta E}{T}\right)}$$



Hence, given initial training data $\mathbf{v}$ and prior guesses on a_i, b_j, w_{ij}, we can stochastically *generate* $\mathbf{h}$, *reconstruct* $\mathbf{v}'$, and onwards.

Boltzmann Machine: It works!

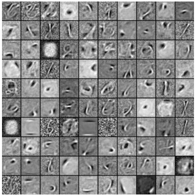
Now that we can generate stochastic *neighboring* (microscopic) states, we tune/update parameters such that their energies are *minimized* globally.

$$\Delta W = \epsilon \left(\mathbf{v}\mathbf{h}^{\mathsf{T}} - \mathbf{v}'\mathbf{h}'^{\mathsf{T}} \right) ,$$

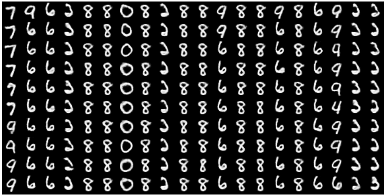
$$\Delta \mathbf{a} = \epsilon(\mathbf{v} - \mathbf{v}') , \quad \Delta \mathbf{b} = \epsilon(\mathbf{h} - \mathbf{h}')$$

Here are some results applied on *text recognition*.

(a) Training Data



(b) Filters (w_{ij}'s)



(c) Samples

To sum up today's talk, we explored:

- How to quantify *information* with (Shannon) entropy
- *Principle of Maximum Entropy* (epistemic modesty)
- Jaynes formalism of (equilibrium) statistical mechanics, one key result being *Boltzmann Statistics*
- Application of Boltzmann statistics in *deep learning*

... and there are MANY more interesting applications and research topics in the area. Feel free to talk to me later!

THANK YOU!

References

[1] I. Goodfellow, Y. Benglo, and A. Courville. *Deep Learning (Adaptive Computation and Machine Learning series)*. The MIT Press, November 2016.

[2] E. T. Jaynes. *Information theory and statistical mechanics*. The Physical Review, 106(4):620–630, May 1957.

[3] M. Mézard and A. Montanari. *Information, Physics, and Computation*. Oxford University Press, 2009.

[4] Steve Presseé and Kingshuk Ghosh and Julian Lee and Ken A. Dill. *Principles of maximum entropy and maximum caliber in statistical physics*. Reviews of Modern Physics, 85:1115-1141, July 2013.