Entropy

January 19, 2022

Abstract

The language-learning diary entertains a recurring theme of entropy and various related principles. Although there are many, many, many resources for these concepts, it seems convenient to put them here, all in one place, in an overview form, as a handy quick-reference and refresher. The content here is extracted from various texts.

1 Fast Overview

A generic fast overview. The rest of this text is organized as ...

1.1 Partition Function

See Wikipedia "Partition function"[6] for more.

- States denoted by σ (spins, from Ising model) with distribution $p\left(\sigma\right)$.
- In machine learning, one writes x for σ. In probability theory, one writes $X=x$ for σ, where X is distribution, and x is a specific sampling of that distribution. That is, $p\left(\sigma\right)$ is the same thing as $P\left(X=x\right)$.
- σ is an indexed set, with$N$ elements. As such, one can pretend that it is an N-dimensional vector, which is fine "for most practical purposes", but in rare cases can lead to confusion.
- One is often (usually) interested in the large-N limit, i.e. $N\gg1$ i.e. $N\to\infty$ states.
- The "energy" of a state is $E\left(\sigma\right)=-\log p\left(\sigma\right)+const.$
- The density of states is: $\rho(E)=\sum_{\sigma}\delta(E-E(\sigma))$
- Total entropy is $S\left(E\right)=\log\rho\left(E\right)$
- Both the energy and the entropy contain leading large-N term i.e. they are extensive properties.

- Without loss of generality, can write the Boltzmann distribution

$$p\left(\sigma|\beta\right)=\frac{1}{Z\left(\beta\right)}\exp-N\sum_{i}\beta_{i}H_{i}\left(\sigma\right)$$

  where there are M parameters β_i called order parameters, Lagrange multipliers, etc. and the $H_i\left(\sigma\right)$ are constants of motion. "Without loss of generality" means that any probability distribution can always be written in the above form.

- In probability theory and information geometry, one often writes θ instead of β as the parameter, and f_i instead of H_i.

- In machine learning, one often writes w instead of β. In this case, w is a "weight vector", allowing a neural-net interpretation.

- The partition function is

$$Z\left(\beta\right)=\sum_{\sigma}\exp-N\sum_{i}\beta_{i}H_{i}\left(\sigma\right)$$

- The above describes a "pure state", where the parameters β_i are fixed constants.

1.2 Fisher Information Metric

See Wikipedia, "Fisher Information Metric: for details. For a finite set of probabilities, we have

- Normalization: $\sum_i p_i = 1$
- Entropy: $H = \sum_i p_i \log p_i$
- Metric: write $\psi_i = \sqrt{p_i}$ Then the normalization becomes $\sum_i \psi_i^2 = 1$ is an octant of a sphere. The flat space Eucliden metric, projected onto the sphere, is the Fisher information metric.

2 Various definitions of entropy

XXX TODO there are other entropies, e.g. microcanonincal, etc. define them too.

2.1 From Banach norms

I've never seen the below set into writing before. I'm not sure what it means. Given a set of probabilities p_i, define the sum

$$s_q=\sum_{i}p_{i}^{q}$$

for some number q. The first derivative is

$$\begin{aligned}\left.\frac{d}{dq}s_q\right|_{q=1} &= \left.\frac{d}{dq}\sum_i \exp q\log p_i\right|_{q=1} \\ &= \left.\sum_i p_i^q \log p_i\right|_{q=1} \\ &= \sum_i p_i \log p_i\end{aligned}$$

which is the conventional entropy. The Banach space ℓ_q norm is

$$\ell_q = [s_q]^{1/q}$$

and so

$$\begin{aligned}\frac{d}{dq}\ell_q =& \frac{d}{dq}\exp\frac{1}{q}s_q \\ =& \ell_q\frac{d}{dq}\frac{s_q}{q} \\ =& \ell_q\left(\frac{-s_q}{q^2}+\frac{1}{q}\frac{ds_q}{dq}\right)\end{aligned}$$

Then evaluating at $q=1$ one gets

$$\begin{aligned}\left.\frac{d}{dq}\ell_q\right|_{q=1} =& -\left(\sum_i p_i\right)^2 + \sum_i p_i \log p_i \\ =& -1+\sum_i p_i\log p_i\end{aligned}$$

assuming that $\sum_i p_i = 1$ for conventional probabilities.

This reinterprets the entropy as a kind of tangent vector. What is the interpretation of that tangent vector? What does it “mean”?

3 Zipf’s Law, Hidden variable models

Zipf’s law can arise whenever one has that some (not necessarily all) of the order parameters are “rapidly fluctuating”, or are “unknown”, or are “latent” and must be “averaged over” to obtain a distribution. The primary reference for this section is Schwab, *et al*, “Zipf’s law and criticality in multivariate data without fine-tuning”.[4] See also Aitchison *et al.*, “Zipf’s Law Arises Naturally When There Are Underlying, Unobserved Variables”[1] for a less physics-oriented exposition.

Both Aitchison and also Mora *etal.* “Are biological systems poised at criticality?”[3] articulate relationships to Ising models.

3.1 Zipf's Law

A quick articulation of Zipf's law, based on [1] and [3]:

- Zipf's law is the statement that $p\left(\sigma\right) \sim 1/\text{rank}\left(\sigma\right)$.
- Converting to energy notation, where $E = E\left(\sigma\right) = -\log p\left(\sigma\right)$ as before, one can write Zipf's law as
$$\log\ \text{rank}\left(E\right) = E + const$$
- The rank of a given, fixed state σ can directly understood as the number of states $n\left(E\right)$ with energy less than $E = E\left(\sigma\right)$. That is,
$$\text{rank}\left(\sigma\right) = n\left(E\left(\sigma\right)\right) = \int_{-\infty}^{E(\sigma)} dE'\rho\left(E'\right)$$
where $\rho(E)$ is the density of states, as before.
- Equivalently, the derivative of the rank is exactly the density of states:
- $$\frac{d\ \text{rank}\left(E\right)}{dE} = \rho(E) = \sum_\sigma \delta(E - E(\sigma))$$
- Combining the above expressions and solving gives that
$$\log\ \text{rank}\left(E\right) = E + \log P_s\left(E\right)$$
where $P_s\left(E\right) = e^{-E}n\left(E\right)$ is a smoothed, energy-weighted probability of states. This relation is exact (i.e. is independent of Zipf's law).
- Zipf's law can thus be written as
$$P_s\left(E\right) = const.$$
This enables practical calculations on distributions (next section).
- Equivalently, Zipf's law may be written as
$$n\left(E\right) \sim \exp E$$
That is, the number of states below energy E is expanding exponentially.

3.2 Deriving Zipf's Law

The derivation of Zipf's law from latent variables is given by Schwab *etal.*[4] and is summarixed below. Aitchison *etal.*[1] claim to have a more general proof.

- A “latent variable” θ (or a set θ_i of them) are hidden parameters that govern the observed distribution; namely, that

$$p(\sigma) = \int d\theta\, p(\sigma|\theta)\, p(\theta)$$

- Assume that some (maybe all, but at least one) of the order parameters β_i is a latent variable θ_i. That is, write θ_i for β_i when β_i is latent.
- Resuming the notation from the revious section, if the order parameter has some distribution $p(\theta)$, then one has a “mixed state” and must write

$$p(\sigma) = \int d\theta\, p(\theta)\, e^{-N\mathcal{H}(\sigma,\beta)}$$

  where the integral is over the K latent parameters: $\int d\theta = \int d\theta_1 d\theta_2 \cdots d\theta_K$ and $\mathcal{H}(\sigma,\beta) = \sum_i \beta_i H_i(\sigma) + \frac{1}{N}\log Z(\beta)$.

With some mild assumptions, one can approximate the above integral

- If $p(\theta)$ is smooth, if $p(\theta)$ does not depend on N and if $p(\theta)$ has non-vanishing support at the saddle point β^*, then the above can be approximated using saddle-point methods, giving

$$E(\sigma) = -\frac{1}{N}\log p(\sigma) = \sum_i \beta_i^* H_i(\sigma) + \frac{1}{N}\log Z(\beta^*)$$

- The saddle point β^* is the solution to

$$\frac{1}{N}\left.\frac{\partial \log Z(\beta)}{\partial \beta_i}\right|_{\beta^*} = -H_i(\sigma)$$

  when $\beta_i = \theta_i$ is one of the hidden variables, and otherwise is just the overt, non-hidden value β_i.
- Note that $\beta^* = \beta^*(\sigma)$ that is, β^* is a function of σ. This comes from the right-hand-side, above.
- The microcanonical entropy is given by

$$S(\{H_i(\sigma)\}) = \inf_\beta \left[\sum_i \beta_i H_i(x) + c(\beta)\right]$$

- The “multi-dimensional form of the Gartner-Ellis theorem“ (see [4]) states that the microcanonical ensemble is given as the “Legendre-Fenchel transform of the cumulant generating function”.

- The cumulant generating function is

  $$c(\beta) = \lim_{N\to\infty} \frac{1}{N} \log Z(\beta) - C$$

  where $C = \frac{1}{N} \log \int d\sigma$

- Up to the overall constant C, one thus has Zipf's law

  $$S(\{H_i(\sigma)\}) = E(\sigma)$$

The above is a quick proof that Zipf's law arises when there are one or more hidden variables, allowing the energy to be written as a mixture of multiple different "models" $H_i(\sigma)$, each of which might not, in itself, be Zipfian.

3.3 Proportion of Explained Energy Variance (PEEV)

The derivation above required that the distribution of the latent variables $p(\theta)$ be non-zero smooth near the fixed point. The degree to which these need to be smooth depends inversely in how peaked the components $p(\sigma|\theta)$ are. Aitchison *etal.*[1] articulate how PEEV works to blend together latent distributions.

4 Graph Factorization and Belief Propagation

The graph factorization problem can be posed as a constraint problem. When there are relatively few constraints, or when the coupling is weak, this can be solved in $\mathcal{O}(N)$ time by belief propagation or message passing. The general setting is outlined by Mezard and Mora in a very readable paper,[2] recaped here.

A factorizable graph is written in the form

$$p(\vec{x}) = \frac{1}{Z} \prod_a \psi_a \left(x_{i_1(a)}, x_{i_2(a)}, \cdots, x_{i_{k(a)}(a)} \right)$$

Here, the vector $\vec{x} = (x_1, x_2, \cdots, x_N)$ describes the state of the system, so that each x_i is some variable. Thus, $p(\vec{x})$ describes the probability of finding the system in state $\vec{x}$. Its is assumed that the system must satisfy a set of constraints, which are reflected in the ψ_a, with the index a running over the set of constraints. The constraints are assumed to run over only a subset of the full vector $\vec{x}$, so that, for fixed a, only $k = k(a)$ variables are involved. Which of these variables are involved are indicated by the indicator function $\vec{i}(a) = (i_1(a), i_2(a), \cdots, i_k(a))$. The indicated variables for constraint a are then $x_{i_1(a)}, x_{i_2(a)}, \cdots, x_{i_k(a)}$.

Mezard describes how belief propagation can be used to quickly solve graph factorization problems.[2]

I'm confused; belief propagation can also solve problems that don't have a factorization, e.g. single-layer perceptrons. So, really, the point is that belief propagation can be used to solve for probability distributions, and it works well on factorizable distributions, too. Which is kind-of the only point of the quoted paper. They illustrate for the specific case of the Ising mode.

XXX TODO: clean up this section.

5 Ising Models and Markov Random Fields

Ising models are a simple case of MRF. So cover them first.

5.1 Ising Models

Ising models with more than two states are called Potts models.

The Ising Hamlitonian can be derived as a minimum entropy model, by applying constraints that the probability of singletons and pairs must produce the actual, observed frequency distribution of singletons and pairs. The technique for forcing this agreement is called "lagrange multipliers". XXX TODO find a reference that explains how the below is actually done.

The primary issue is that the $H_i(\sigma)$ depend on the full σ, which in general makes the problem intractable. So instead, write the entropy as

$$S = -\sum_\sigma p(\sigma)\log p(\sigma)$$

and write σ as an indexed set; that is, $\sigma = (\sigma_1, \sigma_2, \cdots, \sigma_N)$. Each σ_i can be interpreted as a value at position i, for example, a word at position i in a sentence, an amino acid in position i in a protein, *etc.* In a basic Potts model, one has observed frequencies $f_a(i)$ which counts the frequency at which the location i had the value $\sigma_i = a$ and the pair frequencies $f_{ab}(i,j)$ that location i had $\sigma_i = a$ and also location j had $\sigma_j = b$. ... Its common to ignore the position dependence, i.e. to observe only the averages, independent of position...

The goal is to create a model such that the singleton constraints

$$f_a(i) = \sum_{\sigma_1}\sum_{\sigma_2}\cdots\sum_{\sigma_N} p(\sigma_1, \sigma_2, \cdots, \sigma_N)\,\delta_{\sigma_i,a}$$

and the pairwise constraints

$$f_{ab}(i,j) = \sum_{\sigma_1}\sum_{\sigma_2}\cdots\sum_{\sigma_N} p(\sigma_1, \sigma_2, \cdots, \sigma_N)\,\delta_{\sigma_i,a}\delta_{\sigma_j,b}$$

are obeyed. Here, the $\delta_{\sigma_i,a}$ is the Dirac delta function, such that

$$\delta_{\sigma_i,a} = \begin{cases} 1 & \text{if } \sigma_i = a \\ 0 & \text{otherwise} \end{cases}$$

This is done by "adding a multiple of zero" ... etc. XXXTODO show how this is done.

This causes the model probability to factorize:

$$p(\sigma) = p(\sigma_1, \sigma_2, \cdots, \sigma_N) = \frac{1}{Z}\prod_i \phi(\sigma_i) \prod_{jk} \phi(\sigma_j, \sigma_k)$$

where each factor is given by a Boltzmann model

$$\phi(\sigma_i) = \exp -H_i(\sigma_i)$$

and

$$\phi(\sigma_i, \sigma_j) = \exp -H_{ij}(\sigma_i, \sigma_j)$$

which now satisfies the constraints on the observed frequency counts.

5.2 Gauge fixing

The frequencies are not independent; one has constraints

$$f_a(i) = \sum_{j,b} f_{ab}(i,j)$$

which means that there's ambiguity in how the Hamiltonian is split up, i.e. there is gauge invariance; and so one needs gauge fixing. This is usually done by forcing

$$0 = \sum_i H_i(\sigma_i)$$

and

$$0 = \sum_{i,j} H_{ij}(\sigma_i, \sigma_j)$$

XXX show the details of this; firm it up.

5.3 Markov Random Fields

See Wikipedia, "Markov Random Field"[5] for more.

Factorization of the Hamiltonian into cliques.

Cliques correspond to synonyms. i.e. words (and word-phrases) that are synonymous (i.e. can replace one-another with low energy.)

Cliques are the base-space for the sheaves; sheaves project down to cliques.

TODO flesh all this out.

References

[1] Laurence Aitchison, Nicola Corradi, and Peter E. Latham. Zipf's law arises naturally when there are underlying, unobserved variables. *PLoS Computational Biology*, 12:e1005110, 2016.

[2] Marc Mézard and Thierry Mora. Constraint satisfaction problems and neural networks: a statistical physics perspective. ArXiv abs/0803.3061, 2008.

[3] Thierry Mora and William Bialek. Are biological systems poised at criticality? *Journal of Statistical Physics*, 144:268–302, 2011.

[4] David J. Schwab, Ilya Nemenman, and Pankaj Mehta. Zipf's law and criticality in multivariate data without fine-tuning. *Physical Review Letters*, 113:068102, 2014.

[5] Wikipedia. Markov random field.

[6] Wikipedia. Partition function.