Notebooks

Maximum Entropy Methods (MaxEnt)

11 Jul 2011 16:06

The maximum entropy method is usually stated in a deceptively simple way: from among all the probability distributions compatible with empirical data, pick the one with the highest information-theoretic entropy. To really understand where this comes from, and appreciate it at its proper worth, we need to look at its origins in equilibrium statistical mechanics.

We start with an assemblage of N particles; they each have a three-dimensional position and momentum, and possibly some internal degrees of freedom, so the dimension of the microscopic state space, or phase space, is at least 6N. We define some macroscopic observables; these are functions of the microscopic state, so they partition phase space into regions where the macrovariables are constant. A macroscopic state is a value for the macroscopic variables, which, in extension, corresponds to one of these regions of phase space. Boltzmann's postulate is that the probability of a macroscopic state is proportional to the *volume* of microscopic phase space compatible with it. The equilibrium macrostate is the most-probable, largest volume macrostate. (There is no notion of an equilibrium microstate.) The logarithm of this volume is (proportional) to the Boltzmann entropy, which is the same as the thermodynamic entropy. If we could somehow fix the value of the macroscopic variables, Boltzmann predicts a distribution over phase space which is uniform over the compatible microstates.

A remarkable result comes from this. (The essence of this goes back to Boltzmann.) Suppose that the macroscopic observables are extensive, so that if we divide the assemblage into two parts, the over-all value of the macrovariables is the sum of separate contributions from the two sub-assemblages. Make one sub-assemblage, the "system", very small in comparison to the other, the "environment". Fix the vector of extensive macro-variables in the over-all assemblage at the value t; then $t = t_{\mathrm{sys}} + t_{\mathrm{env}}$. Then the *marginal* distribution for the microstate of the system, x_{sys}, is not uniform but an an exponential family distribution, with the macroscopic observables being the sufficient statistics. The reason is that any system microstate x_{sys} will have a probability proportional to the volume of phase space in the *environment*, V_{env}, which is compatible with the corresponding macrovariable:

$$p(x_{\mathrm{sys}}) \propto V_{\mathrm{env}}(t - t_{\mathrm{sys}}(x_{\mathrm{sys}})) = e^{H_{\mathrm{env}}(t - t_{\mathrm{sys}}(x_{\mathrm{sys}}))}$$

where $H = \log V$ is the Boltzmann entropy. To make use of this, we compare the probability of two system microstates, x_1 and x_2, with two different values of the macroscopic observables, t_1 and t_2:

$$\frac{p(x_1)}{p(x_2)} = \frac{e^{H_{\text{env}}(t-t_1)}}{e^{H_{\text{env}}(t-t_2)}} = e^{H_{\text{env}}(t-t_1)-H_{\text{env}}(t-t_2)}$$

The crucial step is the next one: because the system is much smaller than the environment, and the macroscopic observables are extensive, the total value of the observables for the whole assemblage t should be much greater than either t_1 or t_2. Assuming that H_{env} is a nice function, we can then expand it in a Taylor series about t, and discard terms after first order:

$$\frac{p(x_1)}{p(x_2)} \approx e^{H_{\text{env}}(t)-t_1 \cdot H' - H_{\text{env}}(t)+t_2 \cdot h'} = e^{(t_2-t_1)\cdot H'}$$

where $H' = \nabla H_{\text{env}}|_t$. (Notice that we are doing a Taylor expansion *inside* an exponential, so the approximation is going to be rather loose, unless the system is indeed truly quite small compared to the environment.) Thus

$$p(x_{\text{sys}}) \propto e^{-t_{\text{sys}}(x_{\text{sys}})\cdot H'}$$

which is an exponential family. The natural sufficient statistics are the extensive macroscopic variables, and the natural parameters are $-H'$, the partial derivatives of the environment's Boltzmann entropy H_{env} with respect to the extensive variables. Thermodynamics tells us to call the components of $-H'$ the intensive variables conjugate to the extensive ones, so we get (inverse) temperature as the derivative of entropy with respect to energy, pressure for the volume derivative, chemical potential for molecular species number, etc. (Note that the temperature of the *system* depends on the derivative of the *environment's* entropy. At equilibrium, since the entropy of the whole assemblage is maximized, the gradient of environment entropy has to equal the gradient for system entropy, which sounds better. It also tells us that equilibrium requires constancy of the intensive thermodynamic variables.) In this derivation, the system will have certain expectation values for the extensive variables, but these are secondary consequences of the natural parameters/intensive variables, which come from the environment's entropy and the global totals of the assemblage. The system will also fluctuate around these expectation values.

If we want to deal with the whole assemblage of particles, the uniform distribution over a sub-space is really a very awkward mathematical object. It would be much more convenient, calculationally, to have an exponential family. Gibbs had a brilliant idea about how to get one. Instead of fixing the *actual* values of the macroscopic observables, fix their *expectation* values. This does not pick out a unique distribution, but, acting by analogy from equilibrium maximizing the Boltzmann entropy, maximize the **Gibbs entropy**,

$$S[p] = -\int p(x) \log p(x) dx$$

where x is of course an abbreviation for the huge vector of coordinates in phase space.

Boltzmann's global microstate distribution is uniform over the set

$$\{x : T(x) = t\}$$

Gibbs's distribution is the solution to

$$\max_p -\int p(x) \log p(x) dx \text{ such that } \int p(x)T(x)dx = t, \ \int p(x)dx = 1$$

We can turn this constrained optimization problem into an unconstrained one through introducing a Lagrange multiplier for each macroscopic observable (and another to make sure the distribution integrates to 1):

$$\max_p -\int p(x) \log p(x) dx + \lambda_0(\int p(x)dx - 1) + \sum_{i=1}^{d} \lambda_i(\int p(x)T_i(x)dx - t_i)$$

Take the derivative with respect to $p(x)$, and set it equal to zero at the maximum. (If you are worried about functional-calculus issues, I applaud your mathematical caution, but you will never make it as a physicist.) The solution, p^*, satisfies

$$-\log p^*(x) - 1 + \lambda_0 + \sum_{i=1}^{d} \lambda_i T_i(x) = 0$$

yielding

$$p^*(x) = e^{\lambda_0 - 1 + \lambda \cdot T(x)}$$

which is an exponential family again. The factor $e^{\lambda_0 - 1}$ is just the normalization constant, and not very interesting. The intensive parameters in λ are implicit functions of the constrained expectation values t, and are the thermodynamic variables conjugate to the extensive macroscopic observables.

Since we now have an exponential family again, calculation is very easy. Moreover, one can show that if N is very large and t is the equilibrium value, then $S[p^*(t)] \approx H[t]$. (More exactly, the difference between the two is $o(N)$, so the specific entropies coincide.) So for many calculational purposes, we can replace the awkward Boltzmann distribution with the slick Gibbs distribution. Moreover, we get exactly the same form as Boltzmann gives us for the marginal distribution of a small part of an equilibrium assemblage. Thus we get to use the same math twice, which always makes physicists happy. So, three cheers for Gibbs, and for maximum entropy reasoning in equilibrium statistical mechanics.

Now, ever since the pioneering work of E. T. Jaynes in the 1950s, the idea of maximum entropy has come to mean something different, or at least broader. Jaynes gave, or claimed to give, a completely general prescription for inferring distributions from data, which started with the observation that the Gibbs entropy looks *exactly the same* as Shannon's information-theoretic entropy. (That's why Shannon called it "entropy", after all.) According to Jaynes, then, the "least biased" guess at a distribution is to find the distribution of highest entropy such that the expectation

value of our selected observables equals their observed values. Mathematically, of course, this just leads to the same exponential family as before, with parameters set to enforce the expectation-value constraints. But within an exponential family, the maximum-likelihood estimate of the parameters is the one where observations match expectations, so the maximum entropy distribution is the maximum-likelihood estimate within that exponential family. For the Jaynesian, max-ent school, statistical mechanics follows from a general rule of the logic of inductive inference, namely to maximize entropy under constraints. The exponential family distribution is primary, and Boltzmann's uniform-within-a-macrostate distribution is idle. Moreover, this perspective opens, or seems to open, all sorts of connections between informational and physical quantities. I think it is safe to say that this idea is pretty thoroughly entrenched among physicist who are interested in the intersection of their subject with information theory and computation, as I am.

As a general prescription for statistical inference, however, I think max-ent sucks. Yet lots of the distributions we encounter in practice are not exponential families. (A subtlety: you can always get your favorite probability density f by claiming to constrain $\log f$, but generally this will not give you the right *family* of distributions.) Something then is obviously lacking in a prescription which always and only gives us exponential families.

One of the places it goes wrong is at the beginning, with the constraint rule. The idea of constraining expectation values to match observations is completely unmotivated in logic or probability theory, but without it of course one gets very different distributions indeed. (See Uffink.) It gets a deceptive amount of plausibility from our experience with unimodal (and indeed sharply peaked) distributions, where the expectation value is close to the most probable value (the mode), but this is very far from being the general case. In exponential families of random graphs, for instance, it is very common for their to be *two* modes, with the expectation value in between them --- and deeply improbable. So even though one is maximizing entropy and likelihood, and making expectations match observations, the only natural conclusion is that there's no way in Hell that model would give you what you actually observed. (Jaynes was canny enough to recognize the importance of testing his models, but the procedures he used were of course inconsistent with his supposedly-Bayesian ideology, and it does nothing to redeem the general prescription.) As Seidenfeld points out, taking into account exactly this sort of *distribution* of observations is crucial to proper statistical inference, but systematically neglected in this recipe.

More broadly, there is a deep disanalogy between the situation faced by the statistical data analyst and that faced by the statistical mechanic. The data analyst *has all the data* --- measurements on each sample; also, generally, some idea of the dependence structure between samples. Assuming independence for simplicity, the data analyst can always reduce the data by taking the empirical distribution (= order statistics or histogram, as applicable), but no more. Constrain the expected empirical distribution and maximize the entropy, and in general you get --- the empirical distribution. Any further reduction of the data, beyond the empirical distribution, imposes assumptions about what the true distribution is --- it's saying that certain statistics alone are sufficient, which is implicitly ruling out all the models in which those are *not* sufficient. This may be a good idea, and it might even be a good idea to

use an exponential family, but dictated by the fundamental logic of inductive inference it is not.

The situation facing us in statistical mechanics is very different. Rather than having lots of observations on comparable units or samples, we have only measurements of the macroscopic observables, which are a small number of coarse-grained and *collective* degrees of freedom. To be in the same situation as the data analyst, we would have to have lots of measurements of individual molecules. Or rather, to be in the same position as the statistical mechanic, the data analyst would have to be forbidden from looking at individual samples, and allowed to see only certain more-or-less complicated functionals of the empirical distribution.

This last point is in fact the clue to why, mathematically, maximum entropy often works as an approximate method. (What follows is shamelessly ripped off from writers like Richard Ellis and Imre Csiszar.) One of the fundamental results in large deviations theory is something called Sanov's Theorem, which concerns the deviations of the empirical distribution away from the true distribution. (I'll go over the independent-samples version now for simplicity.) It needs a moment or two of set-up. If we have two distributions P and Q, with densities p and q, then their **relative entropy** or **Kullback-Leibler divergence** is

$$D(P\|Q) \equiv \int p(x) \log \frac{p(x)}{q(x)} dx \geq 0$$

with $D(P\|Q) = 0$ if and only if P=Q. If we write P_n for the empirical distribution after n samples, then Sanov's theorem concerns the probability that it falls into some set of distributions, generically A. Specifically, it says that

$$\frac{1}{n} \log P(\hat{P}_n \in A) \rightarrow - \inf_{Q \in A} D(Q\|P)$$

(Actually, the precise statement is a little more complicated, to handle some subtle points about differences between open and closed sets.) Now suppose that the empirical distribution is already known to satisfy some constraints, say on the averages of certain quantities, i.e., we know that Q is in C. Then, under some further regularity conditions,

$$\frac{1}{n} \log P(\hat{P}_n \in A | \hat{P}_n \in C) \rightarrow - \inf_{Q \in A \cap C} D(Q\|P) + \inf_{R \in C} D(R\|P)$$

Very approximately, then, for Q within the constraint set C,

$$P(\hat{P}_n \approx Q) \approx \exp\{-n[D(Q\|P) - D(R(C)\|P)]\}$$

where $R(C)$ is the distribution in the constraint set which minimizes the relative entropy. If the true, generating distribution P is uniform, then $D(Q\|P) = \text{constant} - S[Q]$, so

$$P(\hat{P}_n \approx Q) \propto \exp\{nS[Q]\}$$

at least to leading order in the exponent.

Let me sum up the previous long and complicated paragraph. If we draw a very large sample from a uniform distribution, and throw out all the samples which do not have certain average values, then with exponentially-large probability, the *empirical distribution* of the remaining samples will be very close to the one which maximizes the Gibbs-Shannon entropy under the constraints. Maximizing the entropy under the constraints then provides a good approximation to the sample distribution, though one which ignores sampling fluctuations. Max-Ent, in other words, works not because of inductive logic, but as a short-cut for exact probability calculations under special circumstances --- much as Boltzmann would have said. Indeed, statistical mechanics as a whole can be built up in a mathematically sound and (to my eyes) pleasing manner on the basis of large deviations theory, with maximum-entropy reasoning as a useful asymptotic hack.

See also: Exponential Families of Probability Measures; Large Deviations; Foundations of Statistical Mechanics; Statistical Mechanics; Statistics; Tsallis Statistics

Recommended:

- I. Csiszár, "Maxent, Mathematics, and Information Theory", pp. 35--50 in Kenneth M. Hanson and Richard N. Silver (eds.), *Maximum Entropy and Bayesian Methods: Proceedings of the Fifteenth International Workshop on Maximum Entropy and Bayesian Methods* [How large deviations results sometimes give maximum entropy distributions as large sample approximations.]
- E. T. Jaynes
  - "Information Theory and Statistical Mechanics I," *Physical Review* **106** (1957): 620--630
  - "Information Theory and Statistical Mechanics II," *Physical Review* **108** (1957): 171--190
  - *Papers on Probability, Statistics, and Statistical Physics* [Reprints both those papers, with many other important ones by Jaynes]
- Robert E. Kass and Larry Wasserman, "The Selection of Prior Distributions by Formal Rules", *Journal of the American Statistical Association* **91** (1996): 1343--1370 [PDF reprint]
- Benoit Mandelbrot, "The Role of Sufficiency and of Estimation in Thermodynamics", *Annals of Mathematical Statistics* **33** (1962): 1021--1038 [See comments under Sufficient Statistics]
- Teddy Seidenfeld [Demonstrations that max-ent methods are, in fact, plagued by the same problems as the old Principle of Insufficient Reason, and not consistent with Bayesian inference. Claims from the Albany group that maximum entropy is always compatible with Bayesian updating are thus incorrect.]
  - "Why I Am Not an Objective Bayesian: Some Reflections Prompted by Rosenkrantz", *Theory and Decision* **11** (1979): 413--440
  - "Entropy and Uncertainty", pp. 259--287 in I. B. MacNeill and G. J. Umphrey (eds.), *Foundations of Statistical Inference* (1987)
- Jos Uffink

  - "Can the Maximum Entropy Principle be explained as a Consistency Requirement?", *Studies in History and Philosophy of Modern Physics* **26B** (1995): 223-261 [Abstract, with links to PDF and PS]
  - "The Constraint Rule of the Maximum Entropy Principle," *Studies in History and Philosophy of Modern Physics* **27** (1996): 47--79 [Abstract, with links to PDF and PS]

Modesty forbids me to recommend:

- CRS, "The Backwards Arrow of Time of the Consistently Bayesian Statistical Mechanic", cond-mat/0410063

To read (thanks to Edward Burns for recommendations):

- P. Dias and A. Shimony, "A Critique of Jaynes' Maximum Entropy Principle," *Advances in Applied Mathematics* **2** (1981): 172--211
- K. Friedman, and A. Shimony, "Jaynes' Maximum Entropy Prescription and Probability Theory," *Journal of Statistical Physics* **3** (1971): 381-384.
- Peter Grunwald and A. Philip Dawid, "Game Theory, Maximum Entropy, Minimum Discrepancy and Robust Bayesian Decision Theory", *Annals of Statistics* **32** (2004): 1367--1433
- Patrick Haffner, Steven Phillips and Rob Schapire, "Efficient Multiclass Implementations of L1-Regularized Maximum Entropy", cs.LG/0506101
- Prakash Ishwar and Pierre Moulin, "On the existence and characterization of the maxent distribution under general moment inequality constraints", cs.IT/0506013 = *IEEE Transactions on Information Theory* **51** (2005): 3322--3333 ["A broad set of sufficient conditions that guarantees the existence of the maximum entropy (maxent) distribution consistent with specified bounds on certain generalized moments is derived. Most results in the literature are either focused on the minimum cross-entropy distribution or apply only to distributions with a bounded-volume support or address only equality constraints. The results of this work hold for general moment inequality constraints for probability distributions with possibly unbounded support, and the technical conditions are explicitly on the underlying generalized moment functions."]
- Oliver Johnson and Christophe Vignat, "Some results concerning maximum Renyi entropy distributions", math.PR/0507400
- Jill North, "Symmetry and Probability", phil-sci/2978 [I heard Prof. North talk about this at PSA 2006, and it sounded good, but I need to read the details.]

Previous versions: 2011-07-11; 2011-01-15
