Homework assignment #11

Optional; due in class Fri 12/7

Maximum Entropy (10 points)

The maximum entropy method is an approach to constructing models. It follows a prescription often described as "let's build a model that reproduces certain observations, but is otherwise as unstructured as possible". Let's see what this means.

Consider a system with states described by $\mathbf{x}$. For simplicity, we will think of these states as discrete (and write sums rather than integrals). Assume that our system is at a statistically stationary point, so that there is a well-defined probability distribution $P(\mathbf{x})$ out of which the states are chosen. As we discussed in class, we typically cannot measure this probability distribution from data. What experiments do is estimate the expectation values of various functions. For instance, when studying a gas, we cannot observe the states of all the molecules, but we can measure the average force (pressure) that the gas exerts on a wall.

Let's label the functions that we can measure as $f_\mu(\mathbf{x})$, where index μ accounts for the fact that we may be measuring several things: $\mu = 1 \ldots K$. Our experiments give us estimates of the expectation values $\langle f_\mu(\mathbf{x})\rangle_{expt}$. When modeling the system, we would like to make sure our model reproduces these observations, i.e.:

$$\langle f_\mu(\mathbf{x})\rangle = \sum_{\mathbf{x}} P(\mathbf{x})\, f_\mu(\mathbf{x}) = \langle f_\mu(\mathbf{x})\rangle_{expt} \qquad (1)$$

But other than that, we would like our distribution to have "as little structure as possible" – specifically, we are looking for $P(\mathbf{x})$ that has the largest possible entropy while being consistent with the constraints (1).

The Boltzmann distribution (3 points; questions 1-4 must be complete for credit)

1. To be a probability distribution, $P(\mathbf{x})$ must also satisfy a normalization constraint. Write it out.
2. Recalling the definition of the entropy of a distribution, write the entropy of $P(\mathbf{x})$; this is the quantity we want to maximize, subject to the $K+1$ constraints above.
3. The problem of maximization under constraints is solved using Lagrange multipliers (one per constraint). Solve this problem, and show that the solution takes the form:

$$P(\mathbf{x}) = \frac{1}{Z}\exp\left[-\sum_{\mu=1}^{K}\lambda_\mu f_\mu(\mathbf{x})\right] \qquad (2)$$

   Here λ_μ are the K Lagrange multipliers associated with the constraints (1), and the remaining Lagrange multiplier λ_0 is lurking somewhere in Z; give the explicit relation.
4. What determines the values of λ_μ and Z to be used in the expression (2)?

5. Let's apply this to an example. Let's say that $\mathbf{x}$ is the state of a physical system with energy $E(\mathbf{x})$, and all we know is the expectation value for the energy:

$$\langle E \rangle = \sum_{\mathbf{x}} P(\mathbf{x})\, E(\mathbf{x})$$

Show that the maximum entropy distribution is exactly the Boltzmann distribution, and the Lagrange multiplier is the inverse temperature. (You will not be able to explicitly relate its value to $\langle E \rangle$.)

The Gaussian distribution (2 points)

6. Consider the situation where $\mathbf{x}$ is just a single real number x (and is now continuous). Suppose that we know the mean value of x and its variance. What is the maximum entropy distribution consistent with these constraints? Rewrite it in a familiar form.

Some people see the maximum entropy prescription as the "preferred" way of selecting models. Others caution that the mantra "this is the least structured model consistent with the observations" does not, in fact, confer a special status to this ansatz – it remains an assumption. For instance, people sometimes jokingly refer to problem 6 as the theorem that "proves" that "if you don't know what it is, it's a Gaussian!" – but this, of course, is only a joke.

Still, the family of maximum entropy distributions does have some very appealing properties.

Sufficient statistics (5 points)

Let's think of λ_μ's as parameters that need to be determined by data. (Which is exactly what they are…) The model is a probability distribution parameterized by the set $\{\lambda_\mu\}$, and we will write this as $P(\mathbf{x}\,|\{\lambda_\mu\})$. We observe a set of samples $\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_N$, each of which is drawn independently and at random from $P(\mathbf{x}\,|\{\lambda_\mu\})$. Assume that we know the form of this distribution from Eq. (2), but not the values of the parameters $\{\lambda_\mu\}$. How can we estimate these parameters directly from the set of observations $\{\mathbf{x}_n\}$?

No finite amount of data will determine the exact values of parameters, but we can compute the **distribution** of parameters given the data: $P(\{\lambda_\mu\} \mid \{\mathbf{x}_n\})$.

7. Use the Bayes theorem to express $P(\{\lambda_\mu\} \mid \{\mathbf{x}_n\})$ in terms of $P(\{\mathbf{x}_n\} \mid \{\lambda_\mu\})$. You will need to introduce the prior over lambda; leave it arbitrary, denoting it $P(\{\lambda_\mu\})$.
8. Recall that all datapoints are drawn independently, and write the probability of the set of observations $P(\{\mathbf{x}_n\} \mid \{\lambda_\mu\})$ in terms of the probability function $P(\mathbf{x}\,|\{\lambda_\mu\})$.
9. Finally (not before!), substitute the form of $P(\mathbf{x}\,|\{\lambda_\mu\})$ from equation (2) and show that the distribution of parameters given the data takes the form:

$$P(\{\lambda_\mu\} \mid \{\mathbf{x}_n\}) = \frac{C}{P(\{\mathbf{x}_n\})} \exp\left[-N \sum_{\mu=1}^{K} \lambda_\mu \ \overline{f_\mu}(\{\mathbf{x}_n\})\right]. \quad (3)$$

a. Give the explicit form of $\overline{f_\mu}(\{\mathbf{x}_n\})$ (hint: they are called "the empirical means").
b. Give the explicit form of C and observe that it does not depend on data.

Your reward if you got to here: let's interpret the equation (3). It tells us that aside from the overall normalization constant, the distribution of possible parameter values consistent with the data depends not on all the details of the data, but only on the empirical means $\{\overline{f_\mu}\}$. In other words, all the information that the data points can give about parameters $\{\lambda_\mu\}$ is contained in the average values of the functions f_μ over the observed dataset.

This situation is described by saying that the reduced set of variables $\{\overline{f_\mu}\}$ constitutes "sufficient statistics" for learning the distribution. Thus, for distributions of this form, the problem of compressing N data points into $K \ll N$ variables that are relevant for parameter estimation can be solved explicitly: if we keep track of the running averages $\{\overline{f_\mu}\}$ we can compress our data as we go along, and we are guaranteed that we will never need to go back and examine the data in more detail. A familiar example is that if we know data are drawn from a Gaussian distribution, running estimates of the mean and variance contain all the information available about the underlying parameter values.

The Gaussian example makes it seem that the concept of sufficient statistics is trivial; in reality, this situation is quite unusual. Most of the distributions that we might write down do not have this property—even if they are described by a finite number of parameters, we cannot guarantee that a comparably small set of empirical expectation values captures all the information about the parameter values.

The generic problem of information processing, by the brain or by a machine, is that we are faced with a huge quantity of data and must extract those pieces that are of interest to us. The idea of sufficient statistics is intriguing in part because it provides an example where this problem of 'extracting interesting information' can be solved completely. If the points $D_1, D_2, \ldots, D_N$ are chosen independently and at random from some distribution, the only thing which could possibly be 'interesting' is the structure of the distribution itself (everything else is random, by construction). This structure is described by a finite number of parameters, and there is an explicit algorithm for compressing the N data points $\{D_n\}$ into K numbers that preserve all of the interesting information. The crucial point is that this procedure cannot exist in general, but only for certain classes of probability distributions. This is an introduction to the idea some kinds of structure in data are learnable from random examples, while other structures are not. Intriguingly, this notion of "learnability" appears to be related to the ideas behind renormalization group, where (for renormalizable theories) the expectation values of a finite number of "relevant" operators carry all the information that survives coarse-graining…

Discussion borrowed verbatim from W. Bialek; ask for reference if curious.