Eric Mjolsness

0.1.1 Cox-Jaynes axioms

Compare to: [E. T. Jaynes, Probability Theory: The Logic of Science. Edited by G. Larry Bretthorst Cambridge University Press; partial version at http://bayes.wustl.edu/etj/prob/book.pdf or http://omega.math.albany.edu:8008/-JaynesBook.html]

Popular reference: [Black philosophical essays "Think"].

$$\left\{ \begin{array}{l} p(m|D) = \dfrac{p(D|m)\,p(m)}{p(D)} \\ \text{posterior} = \dfrac{\text{likelihood} * \text{prior}}{\text{evidence}} \end{array} \right.$$

0.2 MAP inference: Optimization formulation

Bayesian inference requires that we calculate the evidence term and keep a whole function $P(M \mid D)$ up to date, recursively using Equation 0.1, as new data comes in. A computationally less demanding goal is to find the mode of this distribution - the single most likely model. Then the evidence term drops out:

D = Data
H = Hidden vbls
M = Model
= $\vec{\theta}$

$$\begin{aligned} \text{argmax}_M\, P(M \mid D) &= \text{argmax}_M \frac{P(D \mid M)\, P(M)}{P(D)} \\ &= \text{argmax}_M\, P(D \mid M)\, P(M) \end{aligned}$$

(As a matter of notation, often the model M is represented by its parameter vector "θ").

Thus, computing the "Maximum a Posteriori" or MAP estimate of M may be tractable when summing up the evidence is not. Because the logarithm function is monotonic,

$$\begin{aligned} \text{argmax}_M\, P(M \mid D) &= \text{argmax}_M \log\left[P(D \mid M)\, P(M)\right] \\ &= \text{argmax}_M \left[\log P(D \mid M) + \log P(M)\right] \end{aligned}$$

If we happen to have a uniform prior on M, this expression simplifies further to maximizing the log likelihood:

$$\text{argmax}_M\, P(M \mid D) = \text{argmax}_M \log P(D \mid M). \tag{2}$$

In this case we have a "Maximum Likelihood Estimator" (MLE, or just ML).

One reason for taking the logarithm is that probabilities in complex models tend to multiply, producing small numbers and unstable algorithms. A related reason is that in statistical mechanics energies are related to log-probabilities, and energies are the sensible and more or less additive quantities.

Often, the likelihood $P(D \mid M)$ depends on some random variables that are in the model but about which we have no data. These variables comprise H, the hidden variables. We may have repeated observartions of D and H for a fixed but unknown model M, which we wish to infer. Our probabilistic model is presumed to give a formula for $P(D, H \mid M)$. Then we just want to maximize

$$\log P(D \mid M) = \log \sum_{\{H\}} P(D, H \mid M)$$

where the sum is over all possible values of all the H variables - often an impossible sum to compute as there are combinatorially many hidden states. Fortunately there is another computational route.

0.3 EM algorithm for ML and MAP

Alternatively we could compute $P(D \mid M)$ from the assumed formula for $P(D, H \mid M)$ if we also had the distribution of the hidden variables $Q(H) \equiv P(H \mid D, M)$:

$$\log P(D \mid M) = \log \frac{P(D, H \mid M)}{P(H \mid D, M)}$$

In this formula, $P(D, H \mid M)$ (and hence the other distributions) may have an product structure over a number of independent experiments with a system given by the common model M.

Again $Q(H)$ is too hard to get directly, but the following approximation method is often effective. We can optimize the log likelihood, averaged over H:

$$\begin{aligned} \langle \log P(D \mid M) \rangle_{Q(H)} &= \left\langle \log \frac{P(D, H \mid M)}{P(H \mid D, M)} \right\rangle_{Q(H)} \\ &= \langle \log P(D, H \mid M) \rangle_{Q(H)} - \langle \log P(H \mid D, M) \rangle_{Q(H)} \\ &= -\left(\langle E(D, H \mid M) \rangle_{Q(H)} - S[Q] \right) \\ &= -F(D, M, Q) \end{aligned}$$

where $S[Q]$ is the entropy

$$S[Q] = -\langle \log Q(H) \rangle_{Q(H)} = -\sum_{\{H\}} Q(H) \log Q(H)$$

and E is the energy or negative log likelihood

$$E(D, H, M) = -L(D, H, M) = -\log P(D, H \mid M)$$

and finally where we define the free energy F as a sum over the hidden states:

$$\begin{aligned} F(D, M, Q) &= \langle E(D, H \mid M) \rangle_{Q(H)} - S[Q] \\ &= \sum_{\{H\}} Q(H)[-\log P(D, H \mid M) + \log Q(H)] \end{aligned}$$

So the ML estimator is

$$M^* = \text{argmin}_M F(D, M, Q)$$

with the correct $Q(H) = P(H \mid D, M)$ substituted in. Now, however, there is a trick for finding the correct Q. If we minimize $F(D, M, Q)$ with respect to probability distributions Q, satisfying the constraint $\sum_{\{H\}} Q(H) = 1$, we find:

$$Q^*(H) = \text{argmin}_{Q(H)} \sum_{\{H\}} Q(H)[-\log P(D, H \mid M) + \log Q(H) + \lambda]$$

Differentiating with respect to Q and seeking a minimum,

$$0 = -\log P(D, H \mid M) + \log Q^*(H) + 1 + \lambda$$

so we get the Boltzmann distribution

$$Q^*(H) = \frac{\exp(-E(D, H, M))}{\sum_{\{H\}} \exp(-E(D, H, M))}$$

$$Q^*(H) = \frac{P(D, H \mid M)}{\sum_{\{H\}} P(D, H \mid M)} = \frac{P(D, H \mid M)}{P(D \mid M)} = P(H \mid D, M)$$

which is in fact the correct value of $Q(H)$ to substitute into F. Thus, minimizing F with respect to Q automatically substitutes in the right value of Q. F must be minimal simultaneoulsy in M and in Q, in order to yeild the MLE model M^*. This can usually be achieved by alternatively minimizing in M and in Q until a local minimum is reached:

initialize M;

repeat {

$Q(H) = \text{argmin}_{Q \mid \sum Q = 1} F(D, M, Q)$

(* "Expectation" step computes the distribution Q , and thus updates the expectation of the log likelihood*)

$M = \text{argmin}_M F(D, M, Q)$

(* "Maximization" step maximize the log likelihood with respect to the model *)

} until satisfactory convergence in M and Q

This is the Expectation-Maximization (EM) algorithm. It can also be used for MAP, by adding the log prior $P(M)$ to the log likelihood expression.