Maximum Entropy Distributions and Their Relationship to Maximum Likelihood

Tutorial by James Coughlan, revised Jan. 2010

1 Maximum Entropy

The maximum entropy principle is used as a means of estimating probability distributions from data. Specifically, given data samples $x^1, \ldots, x^M$, the maximum entropy distribution $P(x)$ is chosen to be the distribution with maximum entropy that satisfies the following constraint:

$$< f(x) >_P = < f(x) >_{emp} \tag{1}$$

where $f(x)$ is a (scalar or vector) function of x, $< . >_P$ denotes expectation with respect to distribution $P(.)$, and $< . >_{emp}$ denotes expectation with respect to the empirical samples. More specifically, $< f(x) >_P = \sum_x P(x) f(x)$ and $< f(x) >_{emp} = (1/M) \sum_{\mu=1}^{M} f(x^\mu)$. (For simplicity we assume x is discrete, but the procedure is easily generalized to the case where x is continuous, in which case sums over x are replaced by integrals.)

Since the entropy of $P(x)$ is $S = -\sum_x P(x) \log P(x)$, we can use Lagrange multipliers to cast the appropriate constrained optimization problem:

$$E = -\sum_x P(x) \log P(x) + \tau((\sum_x P(x)) - 1) + \lambda((\sum_x P(x) f(x)) - < f(x) >_{emp}) \tag{2}$$

where the second term enforces the fact that probability distributions must sum to 1 and the third term enforces the match between the maximum entropy distribution and the empirical data. (If $f(.)$ and λ are vectors, then anywhere they are multiplied together, a dot product is understood.)

Differentiating E yields:

$$\partial E / \partial P(x) = -\log P(x) - 1 + \tau + \lambda f(x), \tag{3}$$

and setting $\partial E / \partial P(x) = 0$ yields the following closed-form expression for $P(x)$:

$$P(x|\lambda) = \frac{e^{\lambda f(x)}}{Z(\lambda)} \tag{4}$$

where $Z(\lambda)$ is a normalization factor (i.e. $Z(\lambda) = \sum_x e^{\lambda f(x)}$) and we now write the probability conditioned on the value of λ.

The value of λ is chosen to satisfy Eq. 1, which we will discuss later.

As an example, if x is a (continuous) scalar and $f(x)$ is the vector-valued function $f(x) = (x, x^2)$, it is straightforward to show that the maximum entropy distribution $P(x)$ will be a Gaussian.

2 Maximum Likelihood

In this section we review maximum likelihood estimation and show how it relates to maximum entropy. Suppose that we accept that our probability distribution has the form given in Eq. 4, and we are provided empirical samples $x_1, \ldots, x_M$. Then the maximum likelihood estimate of λ is given by:

$$argmax_\lambda \prod_{\mu=1}^{M} P(x^\mu|\lambda) \tag{5}$$

where we are assuming that the samples are conditionally independent given λ. This is the value of λ that best explains all the empirical data. Taking logs, we see this is equivalent to:

$$argmax_\lambda L(\lambda) \tag{6}$$

where

$$L(\lambda) = \sum_{\mu=1}^{M} \log P(x^\mu|\lambda) = \sum_{\mu=1}^{M} (\lambda f(x^\mu) - \log Z(\lambda)) \tag{7}$$

and $L(\lambda)$ is the log likelihood function. We re-express $L(\lambda)$ as:

$$L(\lambda) = M(\lambda < f(x) >_{emp} - \log Z(\lambda)) \tag{8}$$

If we differentiate $L(\lambda)$ we get the following expression:

$$\partial L(\lambda)/\partial\lambda = M(< f(x) >_{emp} - \sum_x P(x|\lambda) f(x)) = M(< f(x) >_{emp} - < f(x) >_{P(.|\lambda)}) \tag{9}$$

which we note attains 0 when eq. 1 is satisfied. In other words, maximum entropy and maximum likelihood lead to the same learned distribution but flow from different assumptions.

It can be shown that $L(\lambda)$ is convex (because its second derivative is negative) and thus has a unique maximum. It can be solved for numerically using gradient ascent:

$$\lambda^{new} = \lambda^{old} + k(< f(x) >_{emp} - \sum_x P(x|\lambda) f(x)) \tag{10}$$

where k is a constant that sets the step size. Other methods are available (e.g. GIS, generalized iterative scaling) that can find the solution with fewer calculations. In practice, the bottleneck in solving for λ with any of these methods is the fact that the expectation $\sum_x P(x|\lambda) f(x)$ can be difficult to evaluate.

3 Comments

The maximum entropy principle is sometimes regarded as an ideal learning method that makes minimal assumptions in arriving at an estimate of a distribution learned from data. However, it is important to realize that the form of $f(x)$ is an important (implicit) assumption that can affect the outcome of learning from data. For example, the maximum entropy distribution corresponding to $f(x)$ will in general be different from one corresponding to $g(x) = f(x)^2$, even if $f(x)$ is always positive (and can thus be deduced from the value of $g(x)$).

Finally, note that maximum likelihood is sometimes regarded as non-Bayesian because there is no explicit prior given on λ (which is implicitly uniform). It is easy to extend the estimation of λ to cases in which a prior is placed on λ, so that $P(x|\lambda)$ is replaced with $P(\lambda|x)$ in Eq. 5.