Developments in Probabilistic Modelling with Neural Networks—Ensemble Learning

David J.C. MacKay
Department of Physics, Cambridge University
Madingley Road, Cambridge CB3 0HE, United Kingdom
mackay@mrao.cam.ac.uk

Abstract

Ensemble learning by variational free energy minimization is a framework for statistical inference in which an ensemble of parameter vectors is optimized rather than a single parameter vector. The ensemble approximates the posterior probability distribution of the parameters.

In this paper I give a review of ensemble learning using a simple example.

1 Ensemble Learning by Free Energy Minimization

A new tool has recently been introduced into the field of neural networks. In traditional approaches to model fitting, a single parameter vector $\mathbf{w}$ is optimized by, say, maximum likelihood or penalized maximum likelihood; in the Bayesian interpretation, these optimized parameters are viewed as defining the mode of a posterior probability distribution $P(\mathbf{w}|D,\mathcal{H})$ (given data D and model assumptions $\mathcal{H}$).

The new concept introduced by Hinton and van Camp (1993) is to work in terms of an approximating *ensemble* $Q(\mathbf{w};\theta)$, that is, a probability distribution over the parameters, and optimize the ensemble (by varying its own parameters θ) so that it approximates the posterior distribution of the parameters $P(\mathbf{w}|D,\mathcal{H})$ well. The objective function chosen to measure the quality of the approximation is a *variational free energy*,[1]

$$F(\theta) = -\int d^k\mathbf{w}\, Q(\mathbf{w};\theta) \log \frac{P(D|\mathbf{w},\mathcal{H})P(\mathbf{w}|\mathcal{H})}{Q(\mathbf{w};\theta)}. \tag{1}$$

The numerator $P(D|\mathbf{w},\mathcal{H})P(\mathbf{w}|\mathcal{H})$ is, within a multiplicative constant, equal to the posterior probability $P(\mathbf{w}|D,\mathcal{H}) = P(D|\mathbf{w},\mathcal{H})P(\mathbf{w}|\mathcal{H})/P(D|\mathcal{H})$. So the free energy $F(\theta)$ can be viewed as the sum of $-\log P(D|\mathcal{H})$ and the Kullback-Leibler divergence between $Q(\mathbf{w};\theta)$ and $P(\mathbf{w}|D,\mathcal{H})$. $F(\theta)$ is bounded below by $-\log P(D|\mathcal{H})$ and only attains this value for $Q(\mathbf{w};\theta) = P(\mathbf{w}|D,\mathcal{H})$. For

[1] Variational free energy minimization is a well-established tool for the approximation of probability distributions in statistical physics (Feynman 1972). The free energy can also be described in terms of description lengths, as in Hinton and van Camp (1993).

certain models and certain approximating distributions, this free energy, and its derivatives with respect to the ensemble's parameters, can be evaluated.

Hinton and van Camp (1993) considered a regression network with one non-linear hidden layer and showed that a *separable* Gaussian approximating distribution $Q(\mathbf{w};\theta)$ can be optimized with a deterministic algorithm.

Hinton and Zemel (1994) have applied the same approach to the optimization of an autoencoder. The hidden-to-output part of an autoencoder is viewed as defining a generative model employing latent variables that live in the hidden layer of the model. The optimization of such a generative model is challenging, requiring, for every given data example, an implicit or explicit computation of the posterior probability distribution P of the latent variables. Hinton and Zemel (1994) view the input-to-hidden 'recognition' part of the autoencoder as defining an approximating distribution Q for this distribution P. A single objective function F can then be defined for simultaneous optimization of the generative model and the recognition model. The Helmholtz machine (Dayan *et al.* 1995) is a further generalization of these ideas.

In a broader statistical context, Neal and Hinton (1993) have shown that it is possible to view the Expectation-Maximization (EM) algorithm in terms of a free energy minimization. The Bayesian (ML II) approach to the optimization of hyperparameters in a hierarchical model (reviewed in (MacKay 1992)) can also be derived as a free energy minimization (MacKay 1995a). The deterministic Boltzmann machine can be derived as a free energy approximation to the Boltzmann machine (Radford Neal, personal communication). And MacKay (1995b) has obtained an algorithm for decoding certain binary codes by variational free energy minimization.

2 Inferring a Gaussian distribution

For background reading on Bayesian methods, the textbook of Box and Tiao (1973) is recommended.

The popular one-dimensional Gaussian distribution is parameterized by a mean μ and a standard deviation σ:

$$P(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \equiv \text{Normal}(x;\mu,\sigma^2). \tag{2}$$

Let us examine the inference of μ and σ given data points x_n, $n = 1 \ldots N$, assumed to be drawn independently from this distribution. When inferring these parameters, we must specify their prior distribution. This gives us the opportunity to include specific knowledge that we have about μ and σ (from independent experiments, or on theoretical grounds, for example). If we have no such knowledge, then we can construct an appropriate prior which embodies our supposed ignorance. In either case, it may be appropriate to consider *conjugate priors*; these are priors which have a functional form which integrates naturally with data measurements, making the inferences have an analytically convenient form. The conjugate prior for a mean μ is a Gaussian, $P(\mu|\mu_0,\sigma_\mu) = \text{Normal}(\mu;\mu_0,\sigma_\mu)$. In the limit $\mu_0 = 0, \sigma_\mu \to \infty$, we obtain the *noninformative prior* for a location parameter, the flat prior. This is 'noninformative' because it is *invariant* under the reparameterization $\mu' = \mu + c$. The prior $P(\mu) = \text{const.}$ is also an *improper* prior, that is, it is not normalizable.

The conjugate prior for a standard deviation σ is a gamma distribution, conveniently defined in terms of the inverse variance $\beta = 1/\sigma^2$:

$$P(\beta) = \Gamma(\beta; b_\beta, c_\beta) = \frac{1}{\Gamma(c_\beta)} \frac{\beta^{c_\beta - 1}}{b_\beta^{c_\beta}} \exp\left(-\frac{\beta}{b_\beta}\right), 0 \leq \beta < \infty \tag{3}$$

This is a simple peaked distribution with mean $b_\beta c_\beta$ and variance $b_\beta^2 c_\beta$. In the limit $b_\beta c_\beta = 1, c_\beta \to 0$, we obtain the noninformative prior for a scale parameter, the $1/\sigma$ prior. This is 'noninformative' because it is invariant under the reparameterization $\sigma' = c\sigma$. The $1/\sigma$ prior is less strange looking if we examine the resulting density over $\log \sigma$, or $\log \beta$, which is flat. This is the prior that expresses ignorance about σ by saying 'well, it could be 10, or it could be 1, or it could be 0.1, ...' Scale variables such as σ are usually best represented in terms of their logarithm. Again, this noninformative prior is improper.

In the following examples, I will use the improper priors for μ and σ.

2.1 Maximum likelihood and marginalization: σ_N and σ_{N-1}

The task of inferring the mean and standard deviation of a Gaussian distribution from N samples is a familiar one, though maybe not everyone understands the difference between the σ_N and σ_{N-1} buttons on their calculator. Let us recap the formulae, then derive them.

Given data $D = \{x_n\}_{n=1}^N$, an 'estimator' of μ is

$$\bar{x} \equiv \sum_{n=1}^N x_n / N, \tag{4}$$

and two estimators of σ are:

$$\sigma_N \equiv \sqrt{\frac{\sum_{n=1}^N (x_n - \bar{x})^2}{N}} \quad \text{and} \quad \sigma_{N-1} \equiv \sqrt{\frac{\sum_{n=1}^N (x_n - \bar{x})^2}{N-1}} \tag{5}$$

There are two principal paradigms for statistics: sampling theory and Bayesian inference. In sampling theory (also known as 'frequentist' or orthodox statistics), one invents 'estimators' of quantities of interest and then chooses between those estimators using some criterion measuring their sampling properties; there is no clear principle for deciding which criterion to use to measure the performance of an estimator; nor, for most criteria, is there any systematic procedure for the construction of optimal estimators. In Bayesian inference, in contrast, once we have made explicit all our modelling assumptions, our inferences are mechanistic. Whatever question we wish to pose, the rules of probability theory give a unique answer which consistently takes into account all the given information. Human-designed estimators and confidence intervals have no role in Bayesian inference; human input only enters into the important tasks of designing the hypothesis space, and implementing inference in that space. The answers to our questions are probability distributions over the quantities of interest. We often find that the estimators of sampling theory emerge automatically as modes or means of these posterior distributions when we turn the handle of Bayesian inference.

In sampling theory, the estimators above can be motivated as follows. $\bar{x}$ is an unbiased estimator of μ which, out of all the possible unbiased estimators of μ, has smallest variance (where this variance is computed by averaging over an ensemble of fictitious experiments in which the data samples are assumed to come from an unknown Gaussian distribution). The estimator $(\bar{x}, \sigma_N)$ is the maximum likelihood estimator for (μ, σ). The estimator σ_N is *biased*, however: the expectation of σ_N, given σ, averaging over many imagined experiments, is not σ. This motivates the invention of σ_{N-1} which can be shown to be an unbiased estimator. Or to be precise, it is σ_{N-1}^2 which is an unbiased estimator of σ^2.

We now look at some Bayesian inferences for this problem, assuming non-informative priors for μ and σ. The emphasis is thus not on the priors, but rather on (a) the likelihood function, and (b) the concept of marginalization. The joint posterior probability of μ and σ is proportional to the likelihood function illustrated by a contour plot in figure 1a. The log likelihood is:

$$\log P(\{x_n\}_{n=1}^N | \mu, \sigma) = -N \log(\sqrt{2\pi}\sigma) - \sum_n (x_n - \mu)^2/(2\sigma^2), \quad (6)$$

$$= -N \log(\sqrt{2\pi}\sigma) - [N(\mu - \bar{x})^2 + S]/(2\sigma^2), \quad (7)$$

where $S \equiv \sum_n (x_n - \bar{x})^2$. Given the Gaussian model, the likelihood can be expressed in terms of the two functions of the data $\bar{x}$ and S, so these two quantities are known as 'sufficient statistics'. The posterior probability of μ and σ is, using the improper priors:

$$P(\mu, \sigma | \{x_n\}_{n=1}^N) = \frac{P(\{x_n\}_{n=1}^N | \mu, \sigma) P(\mu, \sigma)}{P(\{x_n\}_{n=1}^N)} \quad (8)$$

$$= \frac{\frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left(-\frac{N(\mu-\bar{x})^2+S}{2\sigma^2}\right) \frac{1}{\sigma_\mu} \frac{1}{\sigma}}{P(\{x_n\}_{n=1}^N)} \quad (9)$$

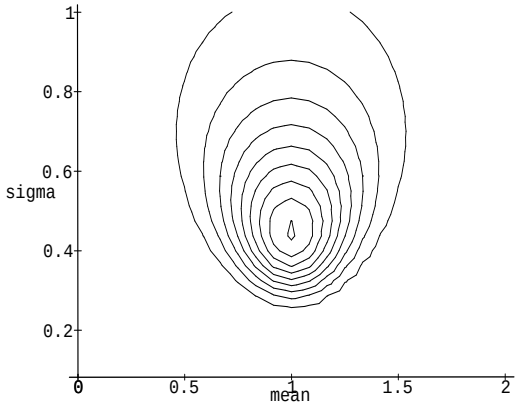
This function describes the answer to the question, 'given the data, and the noninformative priors, what might μ and σ be?' It may be of interest to find the parameter values that maximize the posterior probability (though it should emphasized that posterior probability maxima have no fundamental status in Bayesian inference). Differentiating the log likelihood with respect to μ and $\log \sigma$ we find the maximum likelihood solution: $\{\mu, \sigma\}_{\rm ML} = \left\{\bar{x}, \sigma_N = \sqrt{S/N}\right\}$.

There is more to the posterior distribution than just its mode. As can be seen in figure 1a, the likelihood has a skew peak. As we increase σ, the width of the conditional distribution of μ increases. And if we fix μ to a sequence of values moving away from the sample mean $\bar{x}$, we obtain a sequence of conditional distributions over σ whose maxima move to increasing values of σ.

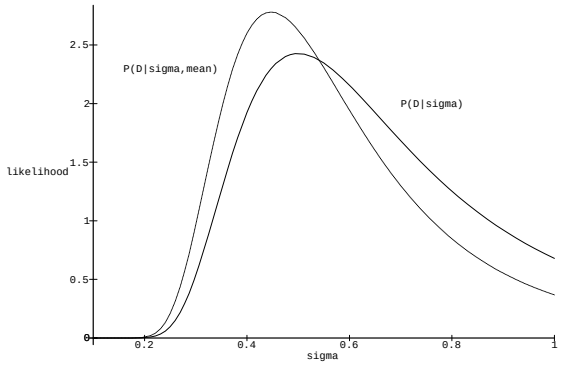
The next question we might ask is 'given the data, and the noninformative prior on μ, and assuming a particular value of σ, what might μ be?'

The posterior probability of μ given σ is

$$P(\mu | \{x_n\}_{n=1}^N, \sigma) = \frac{P(\{x_n\}_{n=1}^N | \mu, \sigma) P(\mu)}{P(\{x_n\}_{n=1}^N | \sigma)} \quad (10)$$



a)



b)

Figure 1: **The likelihood function for the parameters of a Gaussian distribution**.
a) Contour plot of the log likelihood as a function of μ and σ. The data set of $N = 5$ points had mean $\bar{x} = 1.0$ and $S^2 = \sum(x - \bar{x})^2 = 1.0$. Notice that the maximum is skew in σ. The two estimators of standard deviation have values $\sigma_N = 0.45$ and $\sigma_{N-1} = 0.50$.
b) The two graphs show: the likelihood as a function of σ, with μ fixed to $\bar{x}$, *i.e.*, $P(D|\mu = \bar{x}, \sigma)$ [this is a vertical section through the peak in (a)]; and the 'evidence' (marginalized likelihood) for σ, $P(D|\sigma)$, assuming a flat prior on μ (rescaled by an arbitrary constant). The evidence is obtained by projecting the probability mass in (a) onto the σ axis. The maximum of $P(D|\mu = \bar{x}, \sigma)$ is at σ_N. The maximum of $P(D|\sigma)$ is at σ_{N-1}.

$$\propto \quad \exp(-N(\mu - \bar{x})^2/(2\sigma^2)) \tag{11}$$

$$= \quad \mathrm{Normal}(\mu; \bar{x}, \sigma^2/N). \tag{12}$$

We note the familiar $\sigma/\sqrt{N}$ scaling of the error bars on μ.

Let us now ask the question 'given the data, and the noninformative priors, what might σ be?' This question differs from the first one we asked in that we are now not interested in μ. This parameter must therefore be *marginalized* over. The posterior probability of σ is:

$$P(\sigma|\{x_n\}_{n=1}^N) = \frac{P(\{x_n\}_{n=1}^N|\sigma)P(\sigma)}{P(\{x_n\}_{n=1}^N)}. \tag{13}$$

The data-dependent term $P(\{x_n\}_{n=1}^N|\sigma)$ appeared earlier as the normalizing constant in equation (10); one name for this quantity is the 'evidence', or marginalized likelihood, for σ. We obtain the evidence for σ by integrating out μ; a noninformative prior $P(\mu) = 1/\sigma_\mu$ is assumed. The Gaussian integral, $P(\{x_n\}_{n=1}^N|\sigma) = \int P(\{x_n\}_{n=1}^N|\mu,\sigma)P(\mu)\, d\mu$, yields:

$$\log P(\{x_n\}_{n=1}^N|\sigma) = -N\log(\sqrt{2\pi}\sigma) - \frac{S}{2\sigma^2} + \log\frac{\sqrt{2\pi}\sigma/\sqrt{N}}{\sigma_\mu}. \tag{14}$$

The first two terms are the best fit log likelihood (*i.e.*, the log likelihood with $\mu = \bar{x}$). The last term is the log of the 'Occam factor' which penalizes smaller values of σ. When we differentiate the log evidence with respect to $\log\sigma$, to find the most probable σ, the additional volume factor $(\sigma/\sqrt{N})$ shifts the maximum from σ_{N} to

$$\sigma_{N-1} = \sqrt{S/(N-1)} \tag{15}$$

Intuitively, the denominator $(N-1)$ counts the number of noise measurements contained in the quantity $S = \sum_n (x_n - \bar{x})^2$. The sum contains N residuals-squared, but there are only $(N-1)$ effective noise measurements because the determination of one parameter μ from the data causes one dimension of noise to be gobbled up in unavoidable over-fitting. Figure 1b shows the marginalized likelihood as a function of σ along with the likelihood as a function of σ with μ fixed to its most probable value, $\bar{x}$.

The final inference we might wish to make is 'given the data, what is μ?' To answer this, we marginalize over σ and obtain the posterior marginal distribution of μ, which is a Student t-distribution:

$$P(\mu|D) \propto 1/\left(N(\mu - \bar{x})^2 + S\right)^{N/2}. \tag{16}$$

3 An Approximating Ensemble

I now illustrate the concept of ensemble learning by free energy minimization by fitting an approximating ensemble $Q(\mu,\sigma)$ to the posterior distribution (8-9). Let us make the single assumption that the approximating ensemble is separable in the form $Q(\mu,\sigma) = Q_\mu(\mu)Q_\sigma(\sigma)$. No restrictions on the functional form of $Q_\mu(\mu)$ and $Q_\sigma(\sigma)$ are made.

We write down a variational free energy,

$$F(Q) = -\int d\mu \; d\sigma \; Q_\mu(\mu) Q_\sigma(\sigma) \log \frac{P(D|\mu,\sigma)P(\mu,\sigma)}{Q_\mu(\mu)Q_\sigma(\sigma)}. \tag{17}$$

We can find the optimal separable distribution Q by considering separately the optimization of F over $Q_\mu(\mu)$ for fixed $Q_\sigma(\sigma)$, and then the optimization of $Q_\sigma(\sigma)$ for fixed $Q_\mu(\mu)$.

3.1 Optimization of $Q_\mu(\mu)$

As a functional of $Q_\mu(\mu)$, F is:

$$\begin{aligned} F &= -\int d\mu \, Q_\mu(\mu) \left[\int d\sigma \, Q_\sigma(\sigma) \log P(D|\mu,\sigma) + \log[P(\mu)/Q(\mu)] \right] + \text{const.} \\ &= \int d\mu \, Q_\mu(\mu) \left[\int d\sigma \, Q_\sigma(\sigma) N\beta \frac{1}{2}(\mu - \bar{x})^2 + \log Q(\mu) \right] + \text{const.}' \end{aligned}$$

The dependence on Q_σ thus collapses down to a dependence simply on the mean $\bar{\beta} \equiv \int d\sigma \, Q_\sigma(\sigma) 1/\sigma^2$.

Now we can recognize the function $-N\bar{\beta}\frac{1}{2}(\mu - \bar{x})^2$ as the log of a Gaussian identical to the posterior distribution for a particular value of $\beta = \bar{\beta}$. Since a divergence $\int Q \log(Q/P)$ is minimized by setting $Q = P$, we can immediately write down the distribution $Q_\mu^{\text{opt}}(\mu)$ that minimizes F for fixed Q_σ:

$$Q_\mu^{\text{opt}}(\mu) = P(\mu|D, \bar{\beta}, \mathcal{H}) = \text{Normal}(\mu; \bar{x}, \sigma^2_{\mu|D}). \tag{18}$$

where $\sigma^2_{\mu|D} = 1/(N\bar{\beta})$.

3.2 Optimization of $Q_\sigma(\sigma)$

As a functional of $Q_\sigma(\sigma)$, F is (neglecting additive constants):

$$\begin{aligned} F(Q) &= -\int d\sigma \, Q_\sigma(\sigma) \left[\int d\mu \, Q_\mu(\mu) \log P(D|\mu,\sigma) + \log[P(\sigma)/Q_\sigma(\sigma)] \right] \\ &= \int d\sigma \, Q_\sigma(\sigma) \left[(N\sigma^2_{\mu|D} + S)\beta/2 - \left(\tfrac{N}{2} - 1\right) \log \beta + \log Q_\sigma(\sigma) \right] \end{aligned}$$

where the integral over μ is performed assuming $Q_\mu(\mu) = Q_\mu^{\text{opt}}(\mu)$. Here, the β-dependent expression in the brackets can be recognized as the log of a gamma distribution over β (see equation (3)), giving as the distribution that minimizes F for fixed Q_μ: $Q_\sigma^{\text{opt}}(\beta) = \Gamma(\beta; b', c')$, with $1/b' = \frac{1}{2}(N\sigma^2_{\mu|D} + S)$ and $c' = N/2$.

3.3 Joint optimum $Q_\mu(\mu)Q_\sigma(\sigma)$

We now have an implicit equation for the optimal approximating ensemble, with $\sigma^2_{\mu|D} = 1/(N\bar{\beta})$, and $\bar{\beta} = b'c'$. The solution is:

$$1/\bar{\beta} = S/(N-1) \tag{19}$$

Thus we obtain, by ensemble learning, an approximation to the posterior that agrees nicely with the conventional estimators. The approximate posterior distribution over β is a gamma distribution with mean β corresponding to a variance of $\sigma^2 = S/(N-1) = \sigma_{N-1}^2$. And the approximate posterior distribution over μ is a Gaussian with mean $\bar{x}$ and standard deviation $\sigma_{N-1}/\sqrt{N}$.

Acknowledgements

I thank Radford Neal and Geoff Hinton for helpful discussions.

References

Box, G. E. P., and Tiao, G. C. (1973) *Bayesian inference in statistical analysis*. Addison–Wesley.

Dayan, P., Hinton, G. E., Neal, R. M., and Zemel, R. S. (1995) The Helmholtz machine. *Neural Computation*. to appear.

Feynman, R. P. (1972) *Statistical Mechanics*. W. A. Benjamin, Inc.

Hinton, G. E., and van Camp, D., (1993) Keeping neural networks simple by minimizing the description length of the weights. In: *Proceedings of COLT-93*.

Hinton, G. E., and Zemel, R. S. (1994) Autoencoders, minimum description length and Helmholtz free energy. In *Advances in Neural Information Processing Systems 6*, ed. by J. D. Cowan, G. Tesauro, and J. Alspector, San Mateo, California. Morgan Kaufmann.

MacKay, D. J. C. (1992) Bayesian interpolation. *Neural Computation* **4** (3): 415–447.

MacKay, D. J. C., (1995a) Ensemble learning and evidence maximization. submitted to NIPS*95.

MacKay, D. J. C. (1995b) Free energy minimization algorithm for decoding and cryptanalysis. *Electronics Letters* **31** (6): 446–447.

Neal, R. M., and Hinton, G. E. (1993) A new view of the EM algorithm that justifies incremental and other variants. *Biometrika*. submitted.