The Expectation-Maximisation Algorithm

ELEM041 - Machine Learning

February 27, 2009

In these notes we will examine the EM algorithm in general and its application to Gaussian mixture models in particular. In § 1 we will define a mixture model; in § 2 we'll attempt a direct maxmimum likelihood estimation of the parameters; in § 3 we'll introduce the EM algorithm and then apply it to mixture models in § 4.

1 Mixture models

First, let us define a mixture model. Assume we have an observed random variable $X : \Omega \to \mathcal{X}$, which we're assuming has a probability density function (pdf) which is a weighted sum of some other density functions; that is,

$$p_X(x) = \sum_{k=1}^{K} \phi_k f_k(x), \tag{1}$$

where $\phi \in \mathbb{R}^K$ so that $\sum_{k=1}^{K} \phi_k = 1$ and f_k is the k^{th} component density function, properly normalised such that $\int_{\mathcal{X}} f_k(x)\,\mathrm{d}x = 1$.

At this point, all we've done is write a probability density a certain way. The next thing is to show how this model is equivalent to a new model which includes a new random variable $U : \Omega \to \{1, \dots, K\}$. Given such a random variable, we can examine the joint probability of the events $X < x$ and $U = k$:

$$P(X < x \cap U = k) = P(X < x | U = k) P(U = k). \tag{2}$$

Since exactly one the events $U = k$ for $k \in \{1, \dots, K\}$ must happen, the total probability theorem applies, and

$$P(X < x) = \sum_{k=1}^{K} P(X < x | U = k) P(U = k). \tag{3}$$

Differentiating both sides by x gives us the pdf of X:

$$p_X(x) = \frac{\mathrm{d}P(X < x)}{\mathrm{d}x} = \sum_{k=1}^{K} \frac{\mathrm{d}P(X < x | U = k)}{\mathrm{d}x} P(U = k). \tag{4}$$

This is the same as the mixture density (1) if we set $\phi_k = P(U = k)$ and $f_k(x) = \mathrm{d}P(X < x | U = k)/\mathrm{d}x$, the latter being a probability density function conditioned on a particular value of U.

2 Direct maximum likelihood estimation

Next, we'll try to do maximum likelihood estimation in the mixture model and see how far we can get with that. The first step is to be a bit more specific about the form of the component mixture densities, which are written as $f_k(x)$ in (1). We'll assume that are all instances of the same parametric density but with different parameters, so that $f_k(x) = f(x; \theta_k)$, where θ_k contains the parameters for the k^{th} component. Given an observed sequence $\mathbf{x} = (x_1, \ldots, x_N)$, the log likelihood is

$$\log p(\mathbf{x}|\theta) = \sum_{i=1}^{N} \log p_X(x_i|\theta) = \sum_{i=1}^{N} \log \left(\sum_{k=1}^{K} \phi_k f(x_i; \theta_k) \right). \tag{5}$$

Direct differentiation gives

$$\begin{aligned} \frac{\partial \log p(\mathbf{x}|\theta)}{\partial \theta_j} &= \sum_{i=1}^{N} \frac{\phi_j}{p_X(x_i|\theta)} \frac{\partial f(x_i; \theta_j)}{\partial \theta_j} \\ &= \sum_{i=1}^{N} \frac{\phi_j f(x_i; \theta_j)}{p_X(x_i|\theta)} \frac{\partial}{\partial \theta_j} \log f(x_i; \theta_j) \end{aligned}$$

and, introducing a Lagrangian multiplier λ to take account of the constraint that $\sum_{k=1}^{K} \phi_k = 1$,

$$\frac{\partial}{\partial \phi_j} \left(\log p(\mathbf{x}|\theta) - \lambda \sum_{k=1}^{K} \phi_k \right) = -\lambda + \sum_{i=1}^{N} \frac{f(x_i; \theta_j)}{p_X(x_i|\theta)}.$$

Though these don't look too bad, setting them all simultaneously to zero leads to equations in which the parameters are too closely coupled and which cannot be solved in closed form. It would be possible to implement gradient ascent, but this would be very slow and we would like to do better if possible. Notice, however, that

$$\frac{\phi_j f(x_i; \theta_j)}{p_X(x_i|\theta)} = \frac{p_{X,U}(x_i, j|\theta)}{p_X(x_i|\theta)} = p_{U|X}(j|x_i, \theta),$$

that is, the posterior distribution of the class assignments given the data and the parameters. This allows us to write the conditions for a stationary point of the log likelihood as

$$0 = \sum_{i=1}^{N} p_{U|X}(k|x_i, \theta) \frac{\partial}{\partial \theta_k} \log f(x_i; \theta_k) \tag{6}$$

$$\lambda = \sum_{i=1}^{N} p_{U|X}(k|x_i, \theta) \frac{1}{\phi_k}. \tag{7}$$

If the posterior distribution $p_{U|X}$ was known or somehow fixed, these conditions would become decoupled and could be solved directly. The EM algorithm is built on this idea: we take a guess at the posterior, solve for the parameters, and then use these new parameters to take a better guess at the posterior, and so on iteratively.

3 The EM Algorithm

In this section we'll examine the EM algorithm in its most general form that applies not just to mixture models but to any model in which there are unobserved or *latent* variables. Our starting point is the joint probability density specified by the model we are using, which, for a parameterised latent variable model, is $p(\mathbf{x}, \mathbf{u}|\theta)$. To formalise this a bit, we suppose we have two (usually multidimensional) random variables $\mathbf{X} : \Omega \rightarrow \boldsymbol{\mathcal{X}}$ and $\mathbf{U} : \Omega \rightarrow \boldsymbol{\mathcal{U}}$, with the given parameterised joint probability density. Given that we cannot observe $\mathbf{U}$, our aim is to maximise the likelihood $p(\mathbf{x}|\theta)$, or equivalently, the log likelihood $\mathcal{L}(\theta) = \log p(\mathbf{x}|\theta)$, where

$$p(\mathbf{x}|\theta) = \int p(\mathbf{x}, \mathbf{u}|\theta) \, \mathrm{d}\mathbf{u}. \tag{8}$$

This is just a process of marginalisation, where we integrate out one variable in a joint pdf to obtain the pdf of the remaining variable. **NB**: the above is written as an integral rather than a sum, but it should be understood to apply both to continuous and discrete latent variables. In the discrete case, the probability densities become weighted sums of δ-distributions (see Appendix § B) and the integrals reduce to the expected sums. This applies equally to the rest of the development below.

Starting with the likelihood $p(\mathbf{x}|\theta)$, we do various things to it that don't change its value:

$$p(\mathbf{x}|\theta) = \frac{q(\mathbf{u})p(\mathbf{u}|\mathbf{x},\theta)p(\mathbf{x}|\theta))}{q(\mathbf{u})p(\mathbf{u}|\mathbf{x},\theta)} = \frac{q(\mathbf{u})p(\mathbf{x},\mathbf{u}|\theta)}{q(\mathbf{u})p(\mathbf{u}|\mathbf{x},\theta)}, \tag{9}$$

where an arbitrary probability density function q has been introduced. Since q is a probability density, $\int_{\mathcal{U}} q(\mathbf{u}) \, \mathrm{d}\mathbf{u} = 1$, and since the above expression is independent of $\mathbf{u}$ (despite the fact that it contains $\mathbf{u}$), we can take logs and write

$$\log p(\mathbf{x}|\theta) = \int q(\mathbf{u}) \log \frac{q(\mathbf{u})p(\mathbf{x},\mathbf{u}|\theta)}{q(\mathbf{u})p(\mathbf{u}|\mathbf{x},\theta)} \, \mathrm{d}\mathbf{u}. \tag{10}$$

Now we break this integral into two separate terms:

$$\log p(\mathbf{x}|\theta) = \int q(\mathbf{u}) \log \frac{p(\mathbf{x},\mathbf{u}|\theta)}{q(\mathbf{u})} \, \mathrm{d}\mathbf{u} + \int q(\mathbf{u}) \log \frac{q(\mathbf{u})}{p(\mathbf{u}|\mathbf{x},\theta)} \, \mathrm{d}\mathbf{u}. \tag{11}$$

The second term is the Kullback-Leibler divergence (see Appendix § A) between the two distributions $q(\mathbf{u})$ and the posterior $p(\mathbf{u}|\mathbf{x},\theta)$. Since it is non-negative, we have

$$\mathcal{D}(q||p_{U|X}) = \int q(\mathbf{u}) \log \frac{q(\mathbf{u})}{p(\mathbf{u}|\mathbf{x},\theta)} \, \mathrm{d}\mathbf{u} \geq 0. \tag{12}$$

The first term is the focus of the EM algorithm and gets its own symbol:

$$\mathcal{L}^*(\theta, q) = \int q(\mathbf{u}) \log \frac{p(\mathbf{x},\mathbf{u}|\theta)}{q(\mathbf{u})} \, \mathrm{d}\mathbf{u}. \tag{13}$$

In these terms we find that $\mathcal{L}(\theta) = \mathcal{L}^*(\theta, q) + \mathcal{D}(q||p_{\mathbf{U}|\mathbf{X}}) \geq \mathcal{L}^*(\theta)$ and hence that $\mathcal{L}^*(\theta, q)$ is a lower bound on the log likelihood, with equality when the KL divergence between q and $p_{U|X}$ is zero, which happens when q is in fact the posterior density, $q(\mathbf{u}) = p(\mathbf{u}|\mathbf{x}, \theta)$.

The plan now is to search for values of θ and q that maximise $\mathcal{L}^*(\theta, q)$. On the face of it, it would seem that this is not an improvement since we have an optimisation problem in a bigger space than we had before. However, this is the well-known tactic of introducing an *auxiliary* function specially designed to be easier to optimise than the original function. In this case, the auxiliary function $\mathcal{L}^*(\theta, q)$ can be optimised effectively by alternately optimising q and θ while holding the other fixed, resulting in the two step EM algorithm:

$$\text{E step :} \quad q^{(t)} = \arg\max_q \mathcal{L}^*(\theta^{(t)}, q) \tag{14}$$

$$\text{M step :} \quad \theta^{(t+1)} = \arg\max_\theta \mathcal{L}^*(\theta, q^{(t)}). \tag{15}$$

By writing $\mathcal{L}^*(\theta, q)$ in two forms, we can understand a bit more about each step. Firstly, we have already established that

$$\mathcal{L}^*(\theta, q) = \mathcal{L}(\theta) - \mathcal{D}(q||p_{\mathbf{U}|\mathbf{X}}). \tag{16}$$

Since $\mathcal{L}(\theta)$ is independent of q, the E step is equivalent to *minimising* $\mathcal{D}(q||p_{\mathbf{U}|\mathbf{X}})$, which reaches zero when q is the same as the posterior $p_{\mathbf{U}|\mathbf{X}}$. Alternatively, we can write

$$\mathcal{L}^*(\theta, q) = \int q(\mathbf{u}) \log p(\mathbf{x}, \mathbf{u}|\theta) \, \mathrm{d}\mathbf{u} + \mathcal{H}(q), \tag{17}$$

where $\mathcal{H}(q)$ is the *entropy* of the distribution q (see § A), Since $\mathcal{H}(q)$ is independent of θ, the M step is equivalent to maximising

$$\int q(\mathbf{u}) \log p(\mathbf{x}, \mathbf{u}|\theta) \, \mathrm{d}\mathbf{u},$$

which can be thought of as the *expectation* of the *complete data* log likelihood $p(\mathbf{x}, \mathbf{u}|\theta)$ with respect to the distribution over $\mathbf{u}$ specified by q. We can formalise this notion by introducing a new random variable $\mathbf{U}^q$ with precisely this distribution function q, that is, $p_{\mathbf{U}^q}(\mathbf{u}) = q(\mathbf{u})$. In that case, we can write

$$\mathcal{L}^*(\theta, q) = \mathrm{E} \, \log p(\mathbf{x}, \mathbf{U}^q|\theta) + \mathcal{H}(q), \tag{18}$$

It may look like a small difference, but maximising this quantity is much more tractable problem than maximising the true log likelihood

$$\mathcal{L}(\theta) = \log \int p(\mathbf{x}, \mathbf{u}|\theta) \, \mathrm{d}\mathbf{u}.$$

We will see that, when applied to a mixture model, the M step reduces to a number of independent weighted maximum likelihood problems which can be solved in closed-form.

4 EM algorithm for mixture models

The preceeding development applies to any latent variable model. In this section we'll apply the theory to data consisting of multiple independent samples from a mixture model. In that case the observed variable $\mathbf{X}$ in the EM algorithm is a *sequence* of independent identically distributed (iid) copies of the variable X in the mixture model and similarly for the latent variables:

$$\begin{aligned} \mathbf{X} &= (X_1, \ldots, X_N) \qquad & \mathbf{x} &= (x_1, \ldots, x_N) \in \mathcal{X}^N \qquad (19) \\ \mathbf{U} &= (U_1, \ldots, U_N) \qquad & \mathbf{u} &= (u_1, \ldots, u_N) \in \{1, \ldots, K\}^N. \qquad (20) \end{aligned}$$

The parameters of the model are $\theta = (\phi, \theta_1, \ldots, \theta_K)$, where $\phi \in \mathbb{R}^K$ contains the weights of the mixture components and θ_k contains the parameters of the k^{th} mixture component. The model is fully specified by giving the 'probability of everything', $p(\mathbf{x}, \mathbf{u}|\theta)$:

$$p(\mathbf{x}, \mathbf{u}|\theta) = \prod_{i=1}^{N} p(x_i|u_i, \theta)p(u_i|\theta) = \prod_{i=1}^{N} \phi_{u_i} f(x_i; \theta_{u_i}) \qquad (21)$$

The auxiliary function of the EM algorithm is therefore

$$\mathcal{L}^*(\theta, q) = \mathcal{H}(q) + \sum_{\mathbf{u}\in\boldsymbol{\mathcal{U}}} q(\mathbf{u}) \sum_{i=1}^{N} \log \phi_{u_i} f(x_i; \theta_{u_i}), \qquad (22)$$

where q is now a distribution over all possible *sequences* of length N. The best way to represent this object will emerge as we develop the M step, but before that we will derive some general principles for simplifying the expression for $\mathcal{L}^*(\theta, q)$.

4.1 The expectation part

The definition of $\mathcal{L}^*$ involves a sum of the form

$$\sum_{\mathbf{u}\in\boldsymbol{\mathcal{U}}} q(\mathbf{u}) \sum_{i=1}^{N} g_i(u_i) = \sum_{i=1}^{N} \sum_{\mathbf{u}\in\boldsymbol{\mathcal{U}}} q(\mathbf{u}) g_i(u_i)$$

We can introduce an extra summation into this without changing its value by using the Kronecker δ; this allows us to isolate the sum over $\mathbf{u}$:

$$\begin{aligned} \sum_{i=1}^{N} \sum_{\mathbf{u}\in\boldsymbol{\mathcal{U}}} q(\mathbf{u}) g_i(u_i) &= \sum_{i=1}^{N} \sum_{\mathbf{u}\in\boldsymbol{\mathcal{U}}} q(\mathbf{u}) \sum_{k=1}^{K} \delta_{k,u_i} g_i(k) \\ &= \sum_{k=1}^{K} \sum_{i=1}^{N} \left(\sum_{\mathbf{u}\in\boldsymbol{\mathcal{U}}} q(\mathbf{u}) \delta_{k,u_i} \right) g_i(k) \end{aligned}$$

The expression in brackets is the sum of the probabilities associated with all the sequences $\mathbf{u}$ that have k in the i^{th} position and is therefore, as a function of k, the marginal probability distribution, according to q, of the i^{th} element in the sequence. Writing this as $q_i(k)$ gives

$$\sum_{i=1}^{N} \sum_{\mathbf{u}\in\boldsymbol{\mathcal{U}}} q(\mathbf{u}) g_i(u_i) = \sum_{k=1}^{K} \sum_{i=1}^{N} q_i(k) g_i(k), \tag{23}$$

which is a sum over NK terms, a considerable improvement over the NK^N terms of the original sum. We can reach the same conclusion by a couple of different arguments. If q represents a hypothetical distribution over the latent space $\boldsymbol{\mathcal{U}}$, then sums of the form $\sum_{\mathbf{u}} q(\mathbf{u})g(\mathbf{u})$ correspond to expectations taken with respect to that distribution. Thus, we can introduce a hypothetical random variable $\mathbf{U}^q : \Omega \to \boldsymbol{\mathcal{U}}$ with precisely this distribution and write

$$\sum_{\mathbf{u}\in\boldsymbol{\mathcal{U}}} q(\mathbf{u}) g(\mathbf{u}) = \mathrm{E}\ g(\mathbf{U}^q).$$

Now, with one eye on the definition of $\mathcal{L}^*(\theta, q)$, we look at what happens if $g(\mathbf{u})$ decomposes into a sum of terms: we get

$$\mathrm{E}\ g(\mathbf{U}^q) = \mathrm{E} \sum_{i=1}^{N} g_i(U_i^q) = \sum_{i=1}^{N} \mathrm{E}\ g_i(U_i^q).$$

Since each expectation in the sum involves only one component U_i^q, it can be written as a sum over the range of U_i^q, which is just $\{1, \dots, K\}$, yielding, as before,

$$\sum_{i=1}^{N} \mathrm{E}\ g_i(U_i^q) = \sum_{i=1}^{N} \sum_{k=1}^{K} P(U_i^q = k) g_i(k) = \sum_{i=1}^{N} \sum_{k=1}^{K} q_i(k) g_i(k),$$

where we $q_i : \{1, \dots, K\} \to [0, 1]$ is again the marginal distribution of U_i^q, obtained by summing $q(\mathbf{u}) \equiv q(u_1, \dots, u_N)$ over all arguments *except* u_i.

Finally, we can use the δ summation trick to obtain

$$\sum_{i=1}^{N} \mathrm{E}\ g_i(U_i^q) = \sum_{i=1}^{N} \mathrm{E} \left(\sum_{k=1}^{K} \delta_{k,U_i^q} g_i(k) \right) = \sum_{i=1}^{N} \sum_{k=1}^{K} \left(\mathrm{E}\ \delta_{k,U_i^q} \right) g_i(k).$$

Now, δ_{k,U_i^q} is a function of a random variable U_i^q and is either 1 if $U_i^q = k$ or 0 otherwise; hence it is itself a binary random variable. It is easy to show that the expectation of a binary random variable is simply the probability that it takes the value 1, which in this case, is the probability of the event $U_i^q = k$, so once again,

$$\sum_{i=1}^{N} \mathrm{E}\ g_i(U_i^q) = \sum_{i=1}^{N} \sum_{k=1}^{K} P(U_i^q = k) g_i(k).$$

Applying this to the problem at hand, we obtain

$$\mathcal{L}^*(\theta, q) = \mathcal{H}(q) + \sum_{i=1}^{N} \sum_{k=1}^{K} q_i(k) \log \phi_k f(x_i; \theta_k), \tag{24}$$

which is readily optimisable as we will see in the next section.

4.2 The maximisation part

Optimisation of ϕ: We wish to find the value of ϕ that maximises $\mathcal{L}^*(\theta, q)$, but since ϕ represents a probability distribution $P(U = k) = \phi_k$, we also have the constraint that $\sum_{k=1}^K \phi_k = 1$. Using the method of Lagrange multipliers, we define a Lagrangian objective function incorporating a Lagrange multiplier λ:

$$L^\lambda(\phi) = \mathcal{L}^*(\theta, q) - \lambda \sum_{k=1}^K \phi_k. \tag{25}$$

Its derivative with respect to some component of ϕ is

$$\frac{\partial L^\lambda(\phi)}{\partial \phi_j} = \left(\sum_{k=1}^K \sum_{i=1}^N q_i(k) \frac{\partial \log \phi_k}{\partial \phi_j} \right) - \lambda \sum_{k=1}^K \frac{\partial \phi_k}{\partial \phi_j} \tag{26}$$

$$= \left(\sum_{i=1}^N q_i(j) \frac{1}{\phi_j} \right) - \lambda. \tag{27}$$

Therefore, the constrained optimum $\hat{\phi}$ is defined by

$$\sum_{i=1}^N q_i(k) \frac{1}{\hat{\phi}_k} = \lambda \quad \Longrightarrow \quad \hat{\phi}_k = \frac{1}{\lambda} \sum_{i=1}^N q_i(k).$$

The normalisation constraint yields

$$\sum_{k=1}^K \hat{\phi}_k = \frac{1}{\lambda} \sum_{k=1}^K \sum_{i=1}^N q_i(k) = 1 \quad \Longrightarrow \quad \lambda = \sum_{i=1}^N \sum_{k=1}^K q_i(k) = N,$$

so the end result is

$$\hat{\phi}_k = \frac{1}{N} \sum_{i=1}^N q_i(k). \tag{28}$$

Optimisation of θ_k: The derivative of $\mathcal{L}^*(\theta, q)$ with respect to θ_k is found to be, using methods which should be familiar by now,

$$\frac{\partial \mathcal{L}^*(\theta, q)}{\partial \theta_k} = \sum_{i=1}^N q_i(k) \frac{\partial}{\partial \theta_k} \log f(x_i; \theta_k). \tag{29}$$

Rather than solving this directly for a particular component density model f, we can show that finding the zeros of this derivative is equivalent to solving a *weighted* maximum likelihood problem.

Imagine we have just one component density that we wish to fit to the data set $\mathbf{x}$ using standard maximum likelhood methods, except that we wish to simulate a data set in which the i^{th} sample x_i appears not once, but w_i times. These w_i function as non-uniform weightings in the log likelihood function, which we can write as follows

$$\mathcal{L}^w(\theta) = \log \prod_{i=1}^N p(x_i|\theta)^{w_i} = \sum_{i=1}^N w_i \log p(x_i|\theta). \tag{30}$$

This weighted log likelihood is maximised by some $\hat{\theta}$ which is a solution of the zero gradient condition:

$$\sum_{i=1}^{N} w_i \frac{\partial}{\partial\theta} \log p(x_i|\theta) = 0.$$

This is exactly like one the conditions defined by (29) if we set $w_i = q_i(k)$ and allow them to be non-negative real numbers rather than just integers. Hence, the optimisation of θ_k in the M step is equivalent to solving a weighted maximum likelihood problem where the weights are the values $q_i(k)$ for that particular k. Each component density model will have its own weighted maximum likelihood solver $\mathrm{WML}_f : \mathbb{R}^N \times \mathcal{X}^N \to \Theta_f$, where Θ_f denotes the space of parameters for that density model, in terms of which, the solution is

$$\hat{\theta}_k = \mathrm{WML}_f([q_i(k)]_{i=1}^N, \mathbf{x}). \tag{31}$$

The definition of WML_f for Gaussian components is derived in § 5.

4.3 Inferring the posterior

The only piece missing from our solution is the process of setting the distribution q equal to the posterior $p(\mathbf{u}|\mathbf{x}, \theta)$. Thanks to the factorisation of the model, this is relatively simple to write:

$$q(\mathbf{u}) = p(\mathbf{u}|\mathbf{x}, \theta) = \prod_{i=1}^{N} \frac{p(u_i, x_i|\theta)}{p(x_i|\theta)} = \prod_{i=1}^{N} \frac{f(x_i; \theta_{u_i})\phi_{u_i}}{\sum_{k=1}^{K} f(x_i; \theta_k)\phi_k}. \tag{32}$$

However, we have seen from the M step that we never need to represent this function in its entirety; all we require are the marginal distributions q_i, which are just the factors of the full distribution:

$$q_i(k) = \frac{f(x_i; \theta_k)\phi_k}{\sum_{j=1}^{K} f(x_i; \theta_j)\phi_j}. \tag{33}$$

Incidentally, this means that the entropy of q is easily computed as

$$\mathcal{H}(q) = \sum_{i=1}^{N} \sum_{k=1}^{K} -q_i(k) \log q_i(k). \tag{34}$$

4.4 Summary

The EM algorithm is a cyclic process and we can start at any point in the cycle given an appropriate initialisation. Assuming we are given an initial estimate of the parameters, the steps are:

1. Compute the posterior marginals q_i using (33);

2. Optionally compute the current value of $\mathcal{L}^*(\theta, q)$ using (24) and (34).

3. Re-estimate ϕ using (28) and each of the θ_k using (31);
4. Go back to step 1.

However, we could just as easily begin with an arbitrary estimate of the posterior marginals $q_i(k)$ and jump straight in at step 3. The iteration can be terminated on the basis of a convergence test, e.g., the parameters aren't changing very much, or the value of $\mathcal{L}^*(\theta, q)$ is not changing very much.

5 Weighted maximum likelihood for Gaussians

In this section, we solve the weighted maximum likelihood problem for a multi-dimensional Gaussian distribution, since this will enable us to complete the EM algorithm for learning GMMs. The derivation is almost identical to the unweighted case.

Given a sequence of observations $\mathbf{x} = (x_1, \ldots, x_N)$, and a sequence of weights $(w_1, \ldots, w_N)$, which need not sum to one, the weighted log likelihood for a Gaussian with mean μ and covariance C is

$$\begin{aligned}\mathcal{L}^w(\mu, C) &= \log \prod_{i=1}^N [p(x_i|\mu, C)]^{w_i} = \sum_{i=1}^N w_i \log p(x_i|\mu, C) \\ &= -\tfrac{1}{2}\left\{ m\xi_0 \log 2\pi + \xi_0 \log |C| + \sum_{i=1}^N w_i (x_i - \mu)^{\mathrm{T}} C^{-1} (x_i - \mu) \right\},\end{aligned} \tag{35}$$

where we have defined $\xi_0 = \sum_{i=1}^N w_i$. The maximum likelihood estimates of μ and C are defined as

$$(\hat{\mu}, \hat{C}) = \arg\max_{(\mu, C)} \mathcal{L}^w(\mu, C), \tag{36}$$

which can be found by computing the partial derivatives of $\mathcal{L}^w(\mu, C)$ and setting them all to zero to find stationary points of $\mathcal{L}^w(\mu, C)$.

Estimation of the mean The derivative with respect to μ is

$$\frac{\partial \mathcal{L}^w(\mu, C)}{\partial \mu} = -\tfrac{1}{2} w_i \sum_{i=1}^N \frac{\partial}{\partial \mu} (x_i - \mu)^{\mathrm{T}} C^{-1} (x_i - \mu) = \sum_{i=1}^N w_i C^{-1} (x_i - \mu). \tag{37}$$

Hence, $\hat{\mu}$ satisfies $\sum_{i=1}^N w_i C^{-1}(x_i - \hat{\mu}) = 0$. Multiplying by C yields

$$\sum_{i=1}^N w_i (x_i - \hat{\mu}) = 0 \quad \Longrightarrow \quad \hat{\mu} = \frac{1}{\xi_0} \sum_{i=1}^N w_i x_i. \tag{38}$$

Thus the maximum likelihood estimator of the mean is the *weighted* sample mean.

Estimation of the covariance To estimate C, we write the log likelihood in terms of the precision matrix $A = C^{-1}$ and find the value of A that maximises the likelihood.

$$\mathcal{L}^w(\mu, A) = -\tfrac{1}{2}\left\{ m\xi_0 \log 2\pi - \xi_0 \log |A| + \sum_{i=1}^{N} w_i (x_i - \mu)^{\mathrm{T}} A (x_i - \mu) \right\}. \tag{39}$$

$$\frac{\partial \mathcal{L}^w(\mu, A)}{\partial A} = -\tfrac{1}{2}\left\{ -\xi_0 (A^{-1})^{\mathrm{T}} + \sum_{i=1}^{N} w_i (x_i - \mu)(x_i - \mu)^{\mathrm{T}} \right\}. \tag{40}$$

Note that C is symmetric, so $(A^{-1})^{\mathrm{T}} = C$. Since we have already determined that the optimal mean is $\hat{\mu}$, the maximum likelihood estimator $\hat{C}$ satisfies

$$\xi_0 \hat{C} - \sum_{i=1}^{N} w_i (x_i - \hat{\mu})(x_i - \hat{\mu})^{\mathrm{T}} = 0 \qquad \Longrightarrow \qquad \hat{C} = \frac{1}{\xi_0} \sum_{i=1}^{N} w_i (x_i - \hat{\mu})(x_i - \hat{\mu})^{\mathrm{T}}, \tag{41}$$

which is the weighted sample covariance.

Summary The maximum likelihood parameters $\hat{\mu}$ and $\hat{C}$ are both expressed in terms of weighted sums. In fact, we can define three statistics which summarise the data completely as far as this model is concerned:

$$\xi_0 = \sum_{i=1}^{N} w_i, \qquad \xi_1 = \sum_{i=1}^{N} w_i x_i, \qquad \xi_2 = \sum_{i=1}^{N} w_i x_i x_i^{\mathrm{T}}. \tag{42}$$

In terms of these, the estimated covariance is

$$\begin{aligned} \hat{C} &= \frac{1}{\xi_0} \sum_{i=1}^{N} \left(w_i x_i x_i^{\mathrm{T}} - \hat{\mu} x_i^{\mathrm{T}} - x_i \hat{\mu}^{\mathrm{T}} + \hat{\mu}\hat{\mu}^{\mathrm{T}} \right) \\ &= \frac{1}{\xi_0} \left(\xi_2 - \hat{\mu} \xi_1^{\mathrm{T}} - \xi_1 \hat{\mu}^{\mathrm{T}} + \xi_0 \hat{\mu}\hat{\mu}^{\mathrm{T}} \right) \\ &= \frac{1}{\xi_0} \left(\xi_2 - \frac{\xi_1 \xi_1^{\mathrm{T}}}{\xi_0} - \frac{\xi_1 \xi_1^{\mathrm{T}}}{\xi_0} + \frac{\xi_1 \xi_1^{\mathrm{T}}}{\xi_0} \right). \end{aligned}$$

So, to summarise, the solution in terms of these statistics is

$$\hat{\mu} = \frac{\xi_1}{\xi_0}, \qquad \hat{C} = \frac{\xi_0 \xi_2 - \xi_1 \xi_1^{\mathrm{T}}}{\xi_0^2}. \tag{43}$$

Because they capture all the information in the data about the parameters being estimated, ξ_0, ξ_1 and ξ_2 are *sufficient statistics* for the multdimensional Gaussian.

A Kullback-Leibler divergence and entropy

A.1 Entropy and differential entropy

The entropy of a discrete random variable $U : \Omega \to \mathcal{U}$ is defined as

$$H(U) = \sum_{u \in \mathcal{U}} -P(U = u) \log P(U = u). \tag{44}$$

If we introduce a function $l_U(u) = -\log P(U = u)$, then we can write this as an expectation

$$H(U) = \sum_{u \in \mathcal{U}} P(U{=}u) l_U(u) = \mathrm{E}\ l_U(U) \tag{45}$$

The entropy is non-negative and, for a given domain of values $\mathcal{U}$, reaches a maximum of $\log|\mathcal{U}|$ for a uniform distribution, where $|\mathcal{U}|$ is the number of elements in the domain.

Now, if $X : \Omega \to \mathcal{X}$ is a continuous random variable, the definition of entropy does not apply. Even if we quantise X and compute the discrete entropy of the quantised variable, its entropy does not converge to anything as the quantisation grid gets finer and finer. Instead, the *differential* entropy of a continuous random variable with pdf p_X is defined as

$$H(X) = \int_{\mathcal{X}} -p_X(x) \log p_X(x)\ \mathrm{d}x. \tag{46}$$

If we define $l_X(x) = -\log p_X(x)$, then once again, the entropy is equivalent to an expectation:

$$H(X) = \int_{\mathcal{X}} p_X(x) l_X(x)\ \mathrm{d}x = \mathrm{E}\ l_X(X). \tag{47}$$

Sometimes it is convenient to consider the entropy as a function of the the probability density/mass function rather than of the random variable itself, so we can define an alternative entropy function as follows:

$$\mathcal{H}(p) = \begin{cases} \sum_{x \in \mathcal{X}} -p(x) \log p(x) & \text{if } p \text{ is discrete} \\ \int_{\mathcal{X}} -p(x) \log p(x)\ \mathrm{d}x & \text{if } p \text{ is continuous} \end{cases} \tag{48}$$

where $\mathcal{X}$ stands for whatever the domain of p happens to be.

Both the entropy and the differential entropy are measures of how spread out the probability distribution is, and are commonly thought of as measures of *uncertainty*: if an agent represents its beliefs about an unknown value with a probability distribution, the entropy of the distribution is a measure of the agent's uncertainty about the value.

In addition, both $l_U(u)$ and $l_X(x)$ can be thought of as the 'surprisingness' of their respective arguments, since they vary inversely with the probability of the corresponding observation, so the entropy can also be thought of as the 'average surprisingness' of samples of the random variable.

Yet another way of conceptualising entropy comes from thermodynamics and statistical mechanics, which equate entropy with the logarithm of the number microscopic arrangements of a system that are compatible with a given macrostate. Consider drawing a large number of samples from a discrete random variable $X : \Omega \to \mathcal{X}$ and arranging them into a histogram, which will tend to have roughly the same shape as the underlying distribution. The entropy is essentially the log of the number of possible sequences of draws that would lead to that histrogram. Let the sequence of samples be $(x_1, \ldots, x_N)$, and the resulting histogram be Q such that $Q(x) = \sum_{x \in \mathcal{X}} \delta_{x,x_i}$, the number of samples that equal x. The number of ways of rearranging the sequence is $N!$, but this includes all permutations of samples with identical values. If there are $Q(x)$ observations of the value x, these give rise to $Q(x)!$ indistinguishable permutations. Hence, the total number of *distinct* rearrangements of $(x_1, \ldots, x_N)$ is

$$W(Q) = \frac{N!}{\prod_{x \in \mathcal{X}} Q(x)!}. \tag{49}$$

Using Stirlings approximation for the factorial function, and defining the empirical distribution q as $q(x) = Q(x)/N$, we obtain

$$\begin{aligned} \log W(\mathbf{x}) &\approx - \sum_{x \in \mathcal{X}} Q(x) \log \frac{Q(x)}{N} && (50) \\ &= -N \sum_{x \in \mathcal{X}} q(x) \log q(x) = N\mathcal{H}(q) && (51) \end{aligned}$$

A.2 Kullback-Leibler divergence

The Kullback-Leibler (KL) divergence is a measure of the difference between two probability distributions. If $p : \mathcal{U} \to [0, 1]$ and $q : \mathcal{U} \to [0, 1]$ are two probability *mass* functions (not densities) defined on the same *discrete* domain $\mathcal{U}$, then the KL divergence from q to p is

$$\mathcal{D}(q||p) = \sum_{u \in \mathcal{U}} q(u) \log \frac{q(u)}{p(u)}. \tag{52}$$

If p and q are two probability *density* functions defined on the same *continuous* domain $\mathcal{X}$, then the KL divergence is

$$\mathcal{D}(q||p) = \int_{\mathcal{X}} q(x) \log \frac{q(x)}{p(x)} \, \mathrm{d}x. \tag{53}$$

In both cases $\mathcal{D}(q||p) \geq 0$. In the discrete case, $\mathcal{D}(q||p) = 0$ if and only if $q(u) = p(u)$ everywhere. In the continuous case, $\mathcal{D}(q||p) = 0$ if and only if $q(x) = p(x)$ 'almost' everyhwere, which in this context means that the probability of drawing a sample from the q distribution such that $q(x) \neq p(x)$ is zero. It means that there can be isolated points where p and q differ, but their total probability mass according to q must be zero.

The KL divergence is *not* a metric in the strict mathematical sense: it doesn't satisfy the triangle inequality and, more importantly, it is not symmetric: it is quite possible that $\mathcal{D}(q||p) \neq \mathcal{D}(p||q)$. Despite this, it crops up in statistical inference and information theory all the time because because of its close relationship with probability theory. One way of conceptualising it is to consider drawing a large number of samples $\mathbf{x} = (x_1, \ldots, x_N)$ from distribution p and arranging them into a histogram which turns out to look like distribution q: the KL divergence behaves more or less like the negative log probability of this event, namely that of drawing a sample from p that looks like the empirical distribution q. If Q is the histogram of $\mathbf{x}$, such that $Q(x)$ is the number of samples with value x, (i.e. $Q(x) = \sum_{i=1}^{N} \delta_{x,x_i}$) then $q(x) = Q(x)/N$ and

$$\begin{aligned} \log p(\mathbf{x}) &= \sum_{i=1}^{N} \log p(x_i) = \sum_{x \in \mathcal{X}} Q(x) \log p(x) \\ &= N \sum_{x \in \mathcal{X}} q(x) \log p(x). \end{aligned}$$

The probability of observing the histogram Q is $p(\mathbf{x})$ times $W(Q)$, the number of distinct sequences that would produce Q, which we've already shown is approximately $\exp N\mathcal{H}(q)$. Hence,

$$\begin{aligned} -\log W(Q)p(\mathbf{x}) &\approx N \sum_{x \in \mathcal{X}} q(x) \log q(x) - N \sum_{x \in \mathcal{X}} q(x) \log p(x) \\ &= N \sum_{x \in \mathcal{X}} q(x) \log \frac{q(x)}{p(x)} = N\mathcal{D}(q||p). \end{aligned}$$

The fact that this is minimised when $q = p$ confirms the notion that the most probable histogram is one that matches the true underling distribution, and also that, if we are trying to infer the underlying distribution from the data, the maximum likelihood estimate of p is simply q, the empirical distribution.

Another interpretation of the KL divergence that crops up in information theory is as a measure of *information gain*. If an agent's initial beliefs about X are represented by a probability distribution p, and it receives new data causing it to update it's beliefs to a new distribution q, then $\mathcal{D}(q||p)$ is a measure of the information contained in the evidence about the variable X. If the KL divergence is computed using the natural logarithm (i.e. to the base e), then information is expressed in 'nats', where $1\,\text{nat} = 1/\log 2 = \log_2 e\,\text{bits}$. If the KL divergence is computed using logs to the base 2, then the information is in bits. This idea is directly applicable to Bayesian inference using Bayes' rule: if O is the event that certain data were observed, then

$$p(x|O) = \frac{p(O|x)p(x)}{p(O)}.$$

and the information in O about X is $\mathcal{D}(p_{X|O}||p_X)$.

B Delta distributions

The Dirac δ ('delta') distribution is something which looks like a function and has the following properties:

$$\forall x \neq 0, \quad \delta(x) = 0 \tag{54}$$

$$\int_{-\infty}^{\infty} \delta(x) \, \mathrm{d}x = 1. \tag{55}$$

The behaviour of the δ-distribution at $x = 0$ means that it doesn't strictly qualify as a function, though you can think of it as being 'infinite' at zero in such a way as to make the integral come to one. The δ-distribution can also be thought of as the limit of a series of functions consisting of a tall, narrow spike at zero; that is, $\delta = \lim_{\epsilon \to 0} \delta_a$ where δ_a can be defined in a number of ways, e.g.

$$\delta_\epsilon(x) = \frac{e^{-x^2/2\epsilon^2}}{\epsilon\sqrt{2\pi}} \quad \text{(Gaussian distribution)} \tag{56}$$

$$\text{or} \quad \delta_\epsilon(x) = \begin{cases} 1/2\epsilon & \text{if } |x| < \epsilon \\ 0 & \text{if } |x| \geq \epsilon \end{cases} \quad \text{(rectangular distribution)} \tag{57}$$

The values of these functions at zero do not converge to a limit, but the object as a whole converges in the sense of distributions or measures. In particular, a very useful property of the δ-distribution emerges in the limit: for any function f with finite value at zero,

$$\lim_{\epsilon \to 0} \int_{-\infty}^{\infty} f(x)\delta_\epsilon(x) \, \mathrm{d}x = f(0) \quad \Longrightarrow \quad \int_{-\infty}^{\infty} f(x)\delta(x) \, \mathrm{d}x = f(0) \tag{58}$$

This is called the *sifting* property.

Delta distributions can be useful in probability theory because, in some situations, a discrete random variable can be considered to have probability density function composed of a weighted sum of δ-distributions. For example, if $X : \Omega \to \{a_1, \ldots, a_N\}$ and $w_i = P(X = a_i)$, then the notional pdf is

$$f_X(x) = \sum_{i=1}^{N} w_i \delta(x - a_i). \tag{59}$$

Though this isn't strictly a function, it can be integrated, and in particular, expectation can be computed in the normal way:

$$\begin{aligned} \int_{-\infty}^{\infty} g(x) f_X(x) \, \mathrm{d}x &= \sum_{i=1}^{N} w_i \int_{-\infty}^{\infty} g(x)\delta(x - a_i) \, \mathrm{d}x \\ &= \sum_{i=1}^{N} w_i g(a_i) = \mathrm{E} \; g(X). \end{aligned}$$

Note, however, that some probabilistic concepts distinguish between discrete and continuous variables in a fundamental way. In particular the entropy of a discrete random variable is *not* equal to the differential entropy computed using a pdf composed of δ-distributions. The entropy of X as defined above is

$$H(X) = \sum_{i=1}^{N} -w_i \log w_i, \tag{60}$$

and is bounded above and below: $0 \leq H(X) \leq \log N$, whereas the differential entropy computed with the pdf is

$$\mathcal{H}(f_X) = \int_{-\infty}^{\infty} -f_X(x) \log f_X(x) \, \mathrm{d}x, \tag{61}$$

which is not defined since f_X is infinite at the points picked out by the sifting property. Stated more carefully, if we define $f_X^\epsilon(x) = \sum_{i=1}^{N} w_i \delta_\epsilon(x - a_i)$, we find that $\mathcal{H}(f_X^\epsilon) \to -\infty$ as $\epsilon \to 0$. If fact this differential entropy diverges like $\log \epsilon$

Unlike the entropy, the KL divergence between two discrete distribution *can* be represented as the KL divergence between two density functions composed of Dirac δ-distributions; there is no conceptual distinction between discrete and continuous distributions as far as the KL divergence is concerned.