Variational autoencoders

December 19, 2020

1 Variational autoencoders

This reading is a review of the variational autoencoder (VAE) algorithm, that we will be working with this week.

It is split into four sections:

- Section **??** This section motivates and describes the structure of a VAE, which is that of a latent variable model. A VAE is a latent variable model in which the 'encoder' and 'decoder' are neural networks
- Section **??** The second section derives the evidence lower bound (ELBO). The ELBO is the objective function that is used to fit a VAE
- Section **??** In practice, a reparameterisation trick, detailed in the third section, is used when estimating the ELBO
- Section **??** The reading concludes by highlighting that the ELBO can be estimated more precisely if part of it is evaluated analytically, and by providing a recipe for fitting a VAE.

The sections of this reading summarise the main results of the following reference:

- D. P. Kingma and M. Welling. "Auto-Encoding Variational Bayes", 2014. https://arxiv.org/abs/1312.6114

Variational autoencoders have been used for anomaly detection, data compression, image denoising, and for reducing dimensionality in preparation for some other algorithm or model. These applications vary in their use of a trained VAE's encoder and decoder: some use both, while others use only one.

The key point of similarity between a VAE and an autoencoder is that they both use neural networks for tasks that can be interpreted as compression and reconstruction. Additionally, a term in the ELBO resembles the reconstruction error of an autoencoder. Apart from these similarities, VAEs are quite different from autoencoders. Crucially, a VAE is an unsupervised generative model, whereas an autoencoder is not. An autoencoder is sometimes described as being 'self-supervised'. A VAE on the other hand describes the variability in the observations, and can be used to synthesise observations.

1.1 Latent variables and the latent variable model

A latent variable is a random variable that cannot be conditioned on for inference because its value is not known. 'Latent' means hidden. Latent variables do not need to correspond to real quantities. Sometimes models that outwardly do not involve latent quantities are more conveniently expressed by imagining that they do. A perfect example of this is the mixture of Gaussians model: observations can be generated by sampling a label from a categorical distribution, then drawing from the Gaussian in the mixture that has that label.

A latent variable model underlies the variational autoencoder: some latent random variable Z is assumed to have distribution p_θ, and the observation X is assumed to be distributed according to the conditional distribution $p_\theta(x|z)$. X may be either continuous or discrete.

Given some data, our objective is to obtain a maximum likelihood estimate for θ, denoted θ_{ML}. Once θ_{ML} is available, then the distribution of the observable given the latent variable, $p_{\theta_{ML}}(x|z)$, and the marginal likelihood of an observation, $p_{\theta_{ML}}(x)$, can be used.

This model could be fit by maximising the marginal likelihood,

$$p_\theta(x) = \int p_\theta(x|z)p_\theta(z)dz,$$

If this likelihood or its gradient can be efficiently evaluated or approximated, then maximising it with respect to θ is straightforward. Alternatively, the marginal likelihood may be intractable while the posterior $p_\theta(z|x)$ is known or can be efficiently approximated, in which case the EM algorithm could be used.

A simple approach to estimating $p_\theta(x)$ is to take samples z_i ($i \in I$) from $p_\theta(z)$, then take the average of their $p_\theta(x|z_i)$ values. The problem with this method is that if z is high-dimensional, then a very large sample is required to estimate $p_\theta(x)$ well.

Variational inference provides an alternative approach to fitting the model. The high-level idea is this: approximate $p_\theta(z|x)$, then use this approximation to estimate a lower bound on $\log p_\theta(x)$. θ can then be updated based on this lower bound.

The first step in this variational approach is to introduce an approximating distribution for $p_\theta(z|x)$. Call this approximating distribution $q_\phi(z|x)$, where ϕ is its parameter. q_ϕ is fit to p_θ by minimising the Kullback- Leibler divergence

$$D_{KL}(q_\phi(z|x)||p_\theta(z|x))$$

The reasons for this choice of objective function are discussed in more detail in a reading exclusively on the KL divergence later in the week. Its most important properties for now are that it is non-negative, and is zero if and only if q_ϕ and p_θ are equal almost everywhere.

1.2 A bound on the marginal log-likelihood

The marginal log-likelihood of a single observation x can be written

$$\log p_\theta(x) = \log p_\theta(z|x) + \log p_\theta(x,z)$$

Adding and subtracting $\log q_\phi(z|x)$ to the right-hand side of this equation, rearranging the logs, then taking the expectation of both sides under $q_\phi(z|x)$, results in

$$\log p_\theta(x) = \underbrace{E_{zq_\phi}\left[\log \frac{q_\phi(z|x)}{p_\theta(z|x)}\right]}_{D_{KL}(q_\phi||p_\theta)} + E_{Zq_\phi}\left[\log p_\theta(x,z) \log q_\phi(z|x)\right], \qquad (1)$$

where the definition of the KL divergence $D_{KL}(f||g)$ for distributions f and g where $g(x) = 0 \Rightarrow f(x) = 0$ is given by

$$D_{KL}(f||g) = E_{xf}\left[\log \frac{f(X)}{g(X)}\right].$$

If θ is held fixed in (1), then $\log p_\theta(x)$ is fixed too. Because the Kullback-Leibler divergence is non-negative, increasing

$$E_{zq_\phi}\left[\log p_\theta(x,z) q(z|x)\right]$$

with respect to ϕ will reduce $D_{KL}(q_\phi||p_\theta)$, improving our approximating distribution. Additionally, we have the inequality

$$\log p_\theta(x) \geq E_{Zq_\phi}\left[\log p_\theta(x,z) q_\phi(z|x)\right] =: \mathcal{L}(\theta,\phi;x) \qquad (2)$$

This lower bound on the marginal log-likelihood, $\mathcal{L}(\theta,\phi;x)$, is the objective function maximised in variational inference. It is known as the evidence lower bound (ELBO), since the marginal likelihood is the Bayesian evidence of posterior inference in the latent variable model. Notice that $\mathcal{L}$ does not involve evaluating $p_\theta(z|x)$, which we assumed was intractable.

Usually, an analytic expression for the entire ELBO is unavailable. Instead, a Monte Carlo estimate of it can be made. Two estimators for the ELBO are described in Kingma and Welling's original paper. The simplest uses a samples $\{z_j\}_{j=1}^L$ from $q_\phi(z|x)$:

$$\hat{\mathcal{L}}^A(\theta,\phi;x) := \frac{1}{L}\sum_{j=1}^{L} \log p_\theta(x,z_j)\log q_\phi(z_j|x) \qquad (3)$$

In principle, θ and ϕ can now be updated via stochastic gradient ascent using the derivatives of $\mathcal{L}$. Unfortunately, there is a fly in the ointment: the z_j values are not differentiable functions of ϕ, since they are samples. To remove this obstacle to evaluating the gradients, a trick is used.

1.3 The reparameterisation trick

The reparameterisation trick enables derivatives to be propagated to the parameters of a distribution that is sampled from when computing the objective. The essence of the trick is to change how sampling is executed. Rather than sampling from $q_\phi(z|x)$ directly, we instead sample *auxiliary variables* ϵ_j from a distribution $p(\epsilon)$ that is not parameterised by ϕ, then pass them through a ϕ-dependent deterministic transformation $g_\phi(\epsilon,x)$.

We therefore need to choose the distribution $p(\epsilon)$ and transformation $g_\phi(\epsilon,x)$ so that $q_\phi(z|x)$ has the same distribution as $g_\phi(\epsilon;x)$, where $\epsilon p(\epsilon)$, i.e. our sampling procedure is equivalent to sampling from $q_\phi(z|x)$.

For the time being, assume that we know of a $g_\phi(\epsilon,x)$ and $p(\epsilon)$ that satisfy this criterion.

We can then re-write an estimate from the $\hat{\mathcal{L}}^A$ as expressed in (3) in terms of the auxiliary samples:

$$\hat{\mathcal{L}}^A(\theta,\phi;x) := \frac{1}{L}\sum_{j=1}^{L} \log p_\theta(x,z_j)\log q_\phi(z_j|x), \qquad \text{where } z_j = g_\phi(\epsilon_j,x)$$

and the ϵ_j have been sampled from $p(\epsilon)$. This quantity is differentiable with respect to both ϕ and θ, so it can be used for parameter updates in a minibatch gradient ascent algorithm.

For some distributions $q_\phi(z|x)$, an obvious choice of $p(\epsilon)$ and g_ϕ is available. For instance, if $q_\phi(z|x)$ is the density of the multivariate normal $N(\mu,\Sigma)$ with $\phi = (\mu,\Sigma)$, then

$$p(\epsilon) = N(\mathbf{0}, \mathbf{I}), \quad g_\phi(\epsilon, x) = \mu + L\epsilon, \quad \text{where } LL^T = \Sigma$$

results in $q_\phi(z|x)$ and $g_\phi(\epsilon; x)$ being equal in distribution, and g_ϕ being differentiable with respect to ϕ.

1.4 A lower-variance estimator for the ELBO

Referring back to equation (2), we can see that the ELBO can be re-written as

$$\mathcal{L}(\theta, \phi; x) = D_{KL}(q_\phi(z|x)||p_\theta(z)) + E_{zq_\phi}[\log p_\theta(x|z)] \tag{4}$$

If the KL divergence term, an integral, in this expression can be evaluated analytically, then we might expect the estimator

$$\hat{\mathcal{L}}^B(\theta, \phi; x) := D_{KL}(q_\phi(z|x)||p_\theta(z)) + \frac{1}{L}\sum_{j=1}^{L} \log p_\theta(x|z_j)$$

where $z_j = g_\phi(\epsilon_j, x)$, $\epsilon_j p(\epsilon)$, to have lower variance than $\hat{\mathcal{L}}^A$. This is the second ELBO estimator introduced in Kingma and Welling's paper. Usually $q_\phi(z|x)$ and $p_\theta(z)$ are chosen to be Gaussians, meaning that an analytic expression for the divergence can be computed.

Equation (4) helps us to understand the components of the ELBO. The negative of the KL divergence between $q_\phi(z|x)$ and $p_\theta(z)$ penalises $q_\phi(z|x)$ for placing probability mass in locations where $p_\theta(z)$ does not. This has the effect of regularizing $q_\phi(z|x)$.

The second term favours parameter values for which the reconstruction error is small. Given an input $\overline{x}$, an encoding z is sampled from $q_\phi(z|\overline{x})$, then the probability density of a perfect reconstruction is $p_\theta(\overline{x}|z)$. Averaging over the encodings via E_{zq_ϕ} results in utility being placed on parameters that yield probable reconstruction of the input $\overline{x}$.

1.5 Conclusion

To specify and fit a variational autoencoder, choose $p_\theta(z)$, $p_\theta(x|z)$, and $q_\phi(x|z)$, then repeat:

1. Sample a minibatch of observations $x_1, x_2, \ldots, x_n$ and evaluate an estimate of its ELBO,

$$\sum_{j=1}^{n} \hat{\mathcal{L}}(\theta, \phi; x_j)$$

where $\hat{\mathcal{L}}$ is either $\hat{\mathcal{L}}^A$ or $\hat{\mathcal{L}}^B$. In either case, for each x_j, L samples from $q_\phi(z|x_j)$ will be required. These samples should be taken using the reparameterisation trick.

2. Use the gradients of the ELBO estimate to update the parameters θ and ϕ.

Often $q_\phi(z|x)$ is chosen to be a multivariate normal distribution, with a neural network mapping x to its mean and covariance matrix, and ϕ is the parameter vector of the neural network. For continuous data, the distribution $p_\theta(x|z)$ is often also a multivariate normal distribution. Again, a neural network maps z to a mean and covariance matrix, and this neural network is parameterised by θ. For discrete data, often Bernoulli or categorical distributions are used. Typically $p_\theta(z)$ is fixed as a standard multivariate normal distribution.

The model can be sampled from by drawing z from $p_\theta(z)$, then sampling x from $p_\theta(x|z)$. Encodings associated with an observation x can be retrieved by sampling from $q_\phi(z|x)$.

1.6 Further reading and resources

In addition to the Kingma and Welling paper cited above, the following is a general introduction to variational inference, which you may find is useful context for VAEs.

- D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. "Variational Inference: A Review for Statisticians", 2016. https://arxiv.org/abs/1601.00670