A Statistical View of Deep Learning (II): Auto-encoders and Free Energy [1]

15 Mar 2015 | Machine Learning and Statistics
Tags: auto-encoders · deep learning · density estimation · latent variable models · variational inference

With the success of discriminative modelling using deep feedforward neural networks (or using an alternative statistical lens, recursive generalised linear models) in numerous industrial applications, there is an increased drive to produce similar outcomes with unsupervised learning. In this post, I'd like to explore the connections between denoising auto-encoders as a leading approach for unsupervised learning in deep learning, and density estimation in statistics. The statistical view I'll explore casts learning in denoising auto-encoders as that of inference in latent factor (density) models. Such a connection has a number of useful benefits and implications for our machine learning practice.

Generalised Denoising Auto-encoders

Denoising auto-encoders are an important advancement in unsupervised deep learning, especially in moving towards scalable and robust representations of data. For every data point $\boldsymbol{y}$, denoising auto-encoders begin by creating a perturbed version of it $\boldsymbol{y'}$, using a known corruption process $C(\mathbf{y}'|\mathbf{y})$. We then create a network that given the perturbed data $\boldsymbol{y'}$, reconstructs the original data $\boldsymbol{y}$. The network is grouped into two parts, an encoder and a decoder, such that the output of the encoder $\boldsymbol{z}$ can be used as a representation/features of the data. The objective function is [1]:

$$\text{Perturbation:} \quad \mathbf{y}' \sim C(\mathbf{y}'|\mathbf{y})$$

$$\text{Encoder:} \quad \mathbf{z}(\mathbf{y}') = f_\phi(\mathbf{y}') \qquad \text{Decoder:} \quad \mathbf{y} \approx g_\theta(\mathbf{z})$$

$$\text{Objective:} \quad \mathcal{L}_{DAE} = \log p(\mathbf{y}|\mathbf{z})$$

where $\log p(\cdot)$ is an appropriate likelihood function for the data, and the objective function is averaged over all observations. Generalised denoising auto-encoders (GDAEs) realise that this formulation may be limited due to finite training data, and introduce an additional penalty term $\mathcal{R}(\cdot)$ for added regularisation [2]:

$$\mathcal{L}_{GDAE} = \log p(\mathbf{y}|\mathbf{z}) - \lambda \mathcal{R}(\mathbf{y}, \mathbf{y}')$$

GDAEs exploit the insight that perturbations in the observation space give rise to robustness and insensitivity in the representation $\mathbf{z}$. Two key questions that arise when we use GDAEs are: how to choose a realistic corruption process, and what are appropriate regularisation functions.

Separating Model and Inference

The difficulty in reasoning statistically about auto-encoders is that they do not maintain or encourage a distinction between a model of the data (statistical assumptions about the properties and structure we expect) and the approach for inference/estimation in that model (the ways in which we link the observed data to our modelling assumptions). The auto-encoder framework provides a computational pipeline, but not a statistical explanation, since to explain the data (which must be an outcome of our model), you must know it beforehand and use it as an input. Not maintaining the distinction between model and inference impedes our ability to correctly evaluate and compare competing approaches for a problem, leaves us unaware of relevant approaches in related literatures that could provide useful insight, and makes it difficult for us to provide the guidance that allows our insights to be incorporated into our community's broader knowledge-base.

To ameliorate these concerns we typically re-interpret the auto-encoder by seeing the ***decoder as the statistical model of interest*** (and is indeed how many interpret and use auto-encoders in practice). A probabilistic decoder provides a generative description of the data, and our task is inference/learning in this model. For a given model, there are many competing approaches for inference, such as maximum likelihood (ML) and maximum *a posteriori* (MAP) estimation, noise-contrastive estimation, Markov chain Monte Carlo (MCMC), variational inference, cavity methods, integrated nested Laplace approximations (INLA), etc. The role of the encoder is now clear: the ***encoder is one mechanism for inference*** in the model described by the decoder. Its structure is not tied to the model (decoder), and it is just one from the smorgasbord of available approaches with its own advantages and tradeoffs.

Approximate Inference in Latent Variable Models

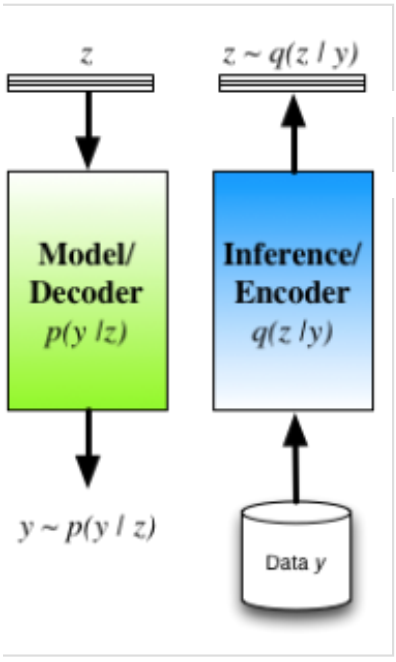
Another difficulty with DAEs is that robustness is obtained by considering perturbations in the data space — such a corruption process will, in general, not be easy to design. Furthermore, by carefully reasoning about the induced probabilities, we can show [1] that the DAE objective function $\mathcal{L}_{DAE}$ corresponds to a lower bound obtained by applying the variational principle to the log-density of the *corrupted data* $\log p(\mathbf{y}')$ — this though, is ***not*** a quantity we are interested in reasoning about.

A way forward would be to instead apply the variational principle to the quantity we are interested in, the log-marginal probability of the *observed data* $\log p(\mathbf{y})$ [3][4]. The objective function obtained by applying the variational principle to the generative model (probabilistic decoder) is known as the **variational free energy**:

$$\mathcal{L}_{VFE} = \mathbb{E}_{q(\mathbf{z})}[\log p(\mathbf{y}|\mathbf{z})] - KL[q(\mathbf{z})\|p(\mathbf{z})]$$

By inspection, we can see that this matches the form of the GDAE objective. There are notable differences though:

Instead of considering perturbations in the observation space, we consider perturbations in the hidden space, obtained by



Encoder-decoder view of nference in latent variable models.

using a prior $p(\mathbf{z})$. The hidden variables are now random, latent variables. Auto-encoders are now generative models that are straightforward to sample from.

The encoder $q(\mathbf{z}|\mathbf{y})$ is a mechanism for approximating the true posterior distribution of the latent/hidden variables $p(\mathbf{z}|\mathbf{y})$.

We are now able to explain the introduction of the penalty function in the GDAE objective in a principled manner. Rather than designing the penalty by hand, we are able to derive the form this penalty should take, appearing as the KL divergence between the the prior and the encoder distribution.

Auto-encoders reformulated in this way, thus provide an efficient way of implementing approximate Bayesian inference. Using an encoder-decoder structure, we gain the ability to jointly optimise all parameters using the single computational graph; and we obtain an efficient way of doing inference at test time, since we only need a single forward pass through the encoder. The cost of taking this approach is that we have now obtained a potentially harder optimisation, since we have coupled the inferences for the latent variables together through the parameters of the encoder. Approaches that do not implement the q-distribution as an encoder have the ability to deal with arbitrary missingness patterns in the observed data and we lose this ability, since the encoder must be trained *knowing the missingness pattern* it will encounter. One way we explored these connections is in a model we called Deep Latent Gaussian Models (DLGM) with inference based on stochastic variational inference (and implemented using an encoder) [3], and is now the basis of a number of extensions [5][6].

Summary

Auto-encoders address the problem of statistical inference and provide a powerful mechanism for inference that plays a central role in our search for more powerful unsupervised learning. A statistical view, and variational reformulation, of auto-encoders allows us to maintain a clear distinction between the assumed statistical model and our approach for inference, gives us one efficient way of implementing inference, gives us an easy-to-sample generative model, allows us to reason about the statistical quantity we are actually interested in, and gives us a principled loss function that includes the important regularisation terms. This is just one perspective that is becoming increasingly popular, and is worthwhile to reflect upon as we continue to explore the frontiers of unsupervised learning.

Some References

[1] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, Pierre-Antoine Manzagol, *Extracting and composing robust features with denoising autoencoders*, Proceedings of the 25th international conference on Machine learning, 2008

[2] Yoshua Bengio, Li Yao, Guillaume Alain, Pascal Vincent, *Generalized denoising auto-encoders as generative models*, Advances in Neural Information Processing Systems, 2013

[3] Danilo Jimenez Rezende, Shakir Mohamed, Daan Wierstra, *Stochastic Backpropagation and Approximate Inference in Deep Generative Models*, Proceedings of The 31st International Conference on Machine Learning, 2014

[4] Diederik P Kingma, Max Welling, *Auto-encoding variational bayes*, arXiv preprint arXiv:1312.6114, 2014

[5] Diederik P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, Max Welling, *Semi-supervised learning with deep generative models*, Advances in Neural Information Processing Systems, 2014

[6] Karol Gregor, Ivo Danihelka, Alex Graves, Daan Wierstra, *DRAW: A Recurrent Neural Network For Image Generation*, arXiv preprint arXiv:1502.04623, 2015