STCS 6701: Foundations of Graphical Models: Reading 6

Maxime TCHIBOZO (MT3390)

October 2020

1 Rezende et al. (2013) – Stochastic Back-Propagation and Variational Inference in Deep Latent Gaussian Models

This paper explores deep directed graphical models (in this context, deep=hierarchical), which capture non-linearities in data, are scalable and can efficiently generate posterior samples.

In a deep graphical model, observed data is assumed to be generated from layers of latent variables, with the directed transition from one layer to the next being either linear or non-linear. Priors are assumed to follow exponential family distributions, and the graph structure of the data generating process defines conditional dependencies between latent variables.

As is the case with other graphical models, deep latent model posterior parameters are inferred through optimization, with the difference here being the introduction of stochastic back-propagation for efficient gradient computations. The paper focuses on the case where all latent variables are assumed to follow Gaussian distributions, a case for which gradients have been well studied (notably using reparametrization tricks - see Kingma and Welling 2013 reading).

Unfortunately, the elegant mathematical derivations from back-propagation do not change the reality that marginal likelihoods are intractable to compute due to the inability to integrate over complex parameter spaces, and this paper uses similar method to Variational Inference (i.e optimization of variational objective function over a parameter space) for approximations. To perform optimization, the paper introduces assumptions similar to those of Mean Field VI, combines them with back-propagation derivations and uses gradient descent.

The authors of this paper apply a deep latent model with 3 x 200-unit latent layers to model the generation of the MNIST dataset. They find the model successfully captures the properties of MNIST digits, and generates realistic-looking digits. They believe these models could be used for missing data imputation for recommender systems and in bioinformatics.

They also highlight the value of latent parameters as tools for visualization of high-dimensional datasets.

2 Bengio et al. (2013) - Representation Learning: A Review and New Perspectives

In Machine Learning, modelling is generally considered separately of feature engineering (usually human-driven or reliant on expert domain knowledge). Representation Learning studies the automatic creation of representations which capture data-specific properties. Representation Learning models span but are not limited to Neural Networks (autoencoders), manifolds and probabilistic graphical models (directed and undirected).

The quality of a representation for data can be evaluated based on its smoothness (if $x \approx y, f(x) \approx f(y)$), interpretability (recovering underlying factors which ideally are conditionally independent of each other and can be ordered hierarchically), transferability (representations which work well for one task should work well for similar tasks), generalizability (underlying latent architectures hold across tasks), compression capabilities, and coherence.

While low-dimensional embeddings (e.g. PCA) or manifolds which capture properties such as neighborhood (e.g. t-SNE which utilizes neighborhood graphs) are well-established Representation Learning methods, this paper focuses on deep architectures, which learn increasingly abstract yet conditionally independent ("disentangled") features as the number of layers grows.

Restricted Boltzman Machines (RBM) are a type of undirected probabilistic graphical model (energy-based model) with strong assumptions on the joint distribution of observed and latent variables, and have been extended (mean-covariance RBM - mcRBM; spike and slab RBM - ssRBM) and adapted (mean-product of Student's T-distributions - mPoT) to produce convincing generative representation models. Deep Boltzmann Machines (DBM) also have multiple hidden layers, and are designed such that layers of odd-number indexed layers conditionally independent given even-number indexed layers (and vice-versa).

A key obstacle in Representation Learning is the curse of dimensionality. The aforementioned methods do not scale well, and the distributions at play become intractable for more than a couple of interconnected layers. Exact inference on DBMs is impossible, so we resort to approximation inference methods such as Mean-Field Variational Inference. Unfortunately, this often leads to convergence to unsatisfactory local minima which do not learn a behavior for certain latent variables. The probabilistic approach is appealing for its interpretability, but is constrained by approximate inference and is not well-suited to multimodal data.

Beyond latent variable models, autoencoders offer a non-probabilistic approach to the encoding of inputs. The authors of this paper reference autoencoders with sparsity constraints, denoising autoencoders (training data is passed through a corruption process ahead of reconstruction - DAE) and contractive autoencoders (CAE), among others. These architectures are unfortunately also challenging to train, with optimization functions which may get stuck in local optima or may be prone to vanishing gradients.

In spite of these novel methods, in many cases the best course of action is still to incorporate domain knowledge in feature engineering, and utilize generic techniques such as dataset augmentation with appropriate deformations (e.g. using pooling and convolutions).

Interestingly, the authors claim that Representation Learning, (a.k.a. unsupervised pre-training) can be conceptualized as a prior when the representation is used for a downstream supervised task. However, when the representation is used for a supervised task and the available quantities of labelled data are large, unsupervised pre-training does not generally improve performance.

Also interestingly, the authors of this paper mention that there is such a thing as over-parametrization for latent variable models.

3 Kingma and Welling (2013) - Auto-Encoding Variational Bayes

Inference on graphical models usually involves intractable posterior distributions of latent variables, especially when latent variables are taken to be continuous. While Variational Inference usually approximates these posteriors under the Mean-Field assumption, this assumption requires analytical estimates of parameter expectations, which can also be intractable. The Mean-Field assumptions are strict and not always appropriate. With Auto-Encoding Variational Bayes (AEVB), on the other hand, the distribution over latent variables $q_\phi(z \mid x)$ is learnt jointly, and is not factorized.

AEVB uses Stochastic Gradient methods to find unbiased estimators of the

ELBO (variational objective function), under the assumption that the dataset at hand is i.i.d with continuous latent variables.

Instead of computing a gradient on the $\mathbf{E}_{q_\phi(z|x)}[\log p_\theta(x \mid z)]$ non-negative term of the ELBO which requires knowledge of the latent distribution q_ϕ, Kingma and Welling approximate this distribution using a probabilistic differentiable parametrization $g_\phi(\epsilon, x)$ such that random variables $\hat{z} \sim q_\phi(z \mid x)$ verify $\hat{z} = g_\phi(\epsilon, x)$ and $\epsilon \sim p(\epsilon)$. Stochastic optimization is then performed on the reparametrized (and differentiable) form of the ELBO.

For example, under the model $z \sim p(z \mid x) = \mathcal{N}(\mu, \sigma^2)$, a parametrization could be $z = \mu + \sigma\epsilon$ and $\epsilon \sim \mathcal{N}(0, 1)$.

The paper views $q_\phi(z \mid x)$ as an encoder for x, and $p_\theta(x \mid z)$ as a decoder which gives a distribution over the possible values of x. Under the reparametrization, the inferred conditional distribution which generated x using a latent $\tilde{z}$ drawn from g can be seen as the reconstruction error, with z drawn from ϕ the true latent encoded embedding.