Deep Learning

AlejandroMllo

This document serves as a very brief summary of the topics covered in each chapter of the book Deep Learning [1].

Disclaimer: This document is completely extracted from [1], the author does not attribute any ownership over the material.

19 Approximate Inference

- In the context of deep learning, we usually have a set of visible variables $\boldsymbol{v}$ and a set of latent variables $\boldsymbol{h}$. The challenge of inference usually refers to the difficult problem of computing $p(\boldsymbol{h}|\boldsymbol{v})$ or taking expectations with respect to it.
- Most graphical models with multiple layers of hidden variables have intractable posterior distributions. Exact inference requires an exponential amount of time in these models.
- Exact inference can be described as an optimization problem. To construct the optimization problem, assume we have a probabilistic model consisting of observed variables $\boldsymbol{v}$ and latent variables $\boldsymbol{h}$. We would like to compute the log-probability of the observed data, $\log p(\boldsymbol{v};\boldsymbol{\theta})$. Sometimes it is difficult to compute $\log p(\boldsymbol{v};\boldsymbol{\theta})$ if it is costly to marginalize out $\boldsymbol{h}$. Instead, we can compute a lower bound $\mathcal{L}(\boldsymbol{v},\theta,q) = \log p(\boldsymbol{v};\boldsymbol{\theta}) - D_{KL}(q(\boldsymbol{h}|\boldsymbol{v})||p(\boldsymbol{h}|v;\theta))$, where q is an arbitrary PD over $\boldsymbol{h}$.
- Expectation Maximization (EM) algorithm: maximizes a lower bound $\mathcal{L}$. It is an approach to learning with an approximate posterior. It consists of alternating between two steps until convergence: (1) The E-step (expectation step), (2) The M-step (maximization step).
- Maximum a posteriori inference (MAP): an alternative form of inference that computes the single most likely value of the missing variables, rather than to infer the entire distribution over their possible values.
- The core idea behind variational learning is that we can maximize $\mathcal{L}$ over a restricted family of distributions q, which should be chosen so that it is easy to compute $\mathbb{E}_q \log p(\boldsymbol{h},\boldsymbol{v})$. A typical way to do this is to introduce assumptions about how q factorizes.
- Using approximate inference as part of a learning algorithm affects the learning process, and this in turn affects the accuracy of the inference algorithm.
  Specifically, the training algorithm tends to adapt the model in a way that makes the approximating assumptions underlying the approximate inference algorithm become more true.
- The wake-sleep algorithm draws samples of both $\boldsymbol{h}$ and $\boldsymbol{v}$ from the model distribution. The inference network can then be trained to perform the reverse mapping: predicting which $\boldsymbol{h}$ caused the present $\boldsymbol{v}$.

References

[1] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.