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.

18 Confronting the Partition Function

- The partition function used to learn undirected models by maximum likelihood depends on the parameters. The gradient of the log-likelihood w.r.t. the parameters has a term corresponding to the gradient of the partition function: $\nabla \log p(\mathrm{x};\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \log \tilde{p}(\mathbf{x};\boldsymbol{\theta}) - \nabla_{\boldsymbol{\theta}} \log Z(\boldsymbol{\theta}$. This is a decomposition into the positive phase and negative phase of learning.
- The Monte Carlo approach to learning undirected models provides an intuitive framework to think about the positive phase and the negative phase. In the positive phase, we increase $\log \tilde{p}(\mathbf{x})$ for $\boldsymbol{x}$ drawn from the data. In the negative phase, we decrease the partition function by decreasing $\log \tilde{p}(\mathbf{x})$ drawn from the model distribution.
- The MCMC approach to maximum likelihood tries to balance between two forces, one pushing up (maximizing $\log \tilde{p}$) on the model distribution where the data occurs, and another pushing down (minimizing $\log Z$) on the model distribution where the model samples occur.
- The contrastive divergence (CD, or CD-k to indicate CD with Gibbs steps) algorithm initializes the Markov chain at each step with samples from the data distribution. See algorithm 18.2. This algorithm can fail to suppress spurious modes.
- A spurious mode is a mode that is present in the model distribution but absent in the data distribution.
- Stochastic maximum likelihood (SML) or Persistent contrastive divergence (PCD/PCD-k): solves many of the problems with CD by initializing the Markov chains at each gradient step with their states from the previous gradient step. See algorithm 18.3.
- The pseudolikelihood is based on the observation that conditional probabilities take this ratio-based form and thus can be computed without knowledge of the partition function. Suppose that we partition $\mathbf{x}$ into $\mathbf{a}$, $\mathbf{b}$ and $\mathbf{c}$, where $\mathbf{a}$ contains the variables we want to find the conditional distribution over, $\mathbf{b}$ contains the variables we want to condition on, and $\mathbf{c}$ contains the variables that are not part of our query:
$$p(\mathtt{a}|) = \frac{p(\mathtt{a},\mathtt{b})}{p(\mathtt{b})} = \frac{p(\mathtt{a},\mathtt{b})}{\sum_{\mathtt{a},\mathtt{c}} p(\mathtt{a},\mathtt{b},\mathtt{c})} = \frac{\tilde{p}(\mathtt{a},\mathtt{b})}{\sum_{\mathtt{a},\mathtt{c}} \tilde{p}(\mathtt{a},\mathtt{b},\mathtt{c})}$$
- Score matching provides another means of training a model without estimating Z or its derivatives. Its strategy is to minimize the expected squared difference between the derivatives of the model's log density w.r.t. the input and the derivatives of the data's log density w.r.t. the input.
- Like pseudolikelihood, score matching only works when it is possible to evaluate $\log \tilde{p}(\mathbf{x})$ and its derivatives directly.
- Denoising score matching is useful because in practice, we usually do not have access to the true p_{data} but rather only an empirical distribution defined by samples from it.
- In Noise-contrastive estimation (NCE), the probability distribution estimated by the model is represented explicitly as $\log p_{model}(\mathbf{x}) = \log \tilde{p}_{model}(\mathbf{x};\theta) + c$, where c is explicitly introduced as an approximation of $-\log z(\theta)$. Rather than estimating only θ, the NCE procedure treats c as just another parameter and estimates θ and c simultaneously, using the same algorithm for both.

References

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