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.

16 Structured Probabilistic Models for Deep Learning

- A structured probabilistic model or graphical model is a way of describing a probability distribution, using a graph to describe which random variables in the probability distribution interact with each other **directly** (this allows the model to have significantly less parameters and therefore be estimated reliably from less data).
- Deep learning's goal is to be able to understand high-dimensional data with rich structure.
- In a graphical model, each node represents a random variable, and each edge represents a direct interaction. These direct interactions imply other, indirect interactions, but only the direct interactions need to be explicitly modeled.
- Graphical models can be divided into: models based on direct acyclic graphs, and models based on undirected graphs.
- Directed models:
  - A Directed graphical model, is also known as a belief network or Bayesian network.
  - If node a has an arrow pointing to node b, then the distribution over b depends on the value of a (that is, the PD of b is defined via a conditional distribution).
  - A directed graphical model defined on variables $\mathbf{x}$ is defined by a directed acyclic graph $\mathcal{G}$ whose vertices are the random variables in the model, and a set of local conditional probability distributions $p(\mathrm{x}_i|Pa_{\mathcal{G}}(\mathrm{x}_i))$, where $Pa_{\mathcal{G}}(\mathrm{x}_i)$ gives the parents of x_i in $\mathcal{G}$. The PD over $\mathbf{x}$ is given by $p(\mathbf{x}) = \prod_i p(\mathrm{x}_i|Pa_{\mathcal{G}}(\mathrm{x}_i))$.
  - As long as each variable has few parents in the graph, the distribution can be represented with very few parameters.
  - The graph encodes only simplifying assumptions about which variables are conditionally independent from each other. Information that cannot be encoded in the graph, is encoded in the definition of the conditional distribution itself.
- Undirected models:
  - An undirected model is also known as Markov random field (MRF) or Markov network.
  - They are used when the interactions seem to have no intrinsic direction, or to operate in both directions.
  - If two nodes are connected by an edge, then the random variables corresponding to those nodes interact with each other directly.
  - An undirected graphical model is a structured probabilistic model defined on an undirected graph $\mathcal{G}$. For each clique $\mathcal{C}$ in the graph, a factor (constrained to be nonnegative) $\phi(\mathcal{C})$ (also called a clique potential) measures the affinity of the variables in that clique for being in each of their possible joint states. Together they define an unnormalized PD: $\tilde{p}(\mathbf{x}) = \prod_{\mathcal{C}\in\mathcal{G}} \phi(\mathcal{C})$.
  - There is nothing to guarantee that multiplying the cliques together will yield a valid PD.
- To obtain a valid PD from the unnormalized PD, use the corresponding normalized PD (Gibbs distribution): $p(\mathbf{x}) = \frac{1}{Z}\tilde{p}(\mathbf{x})$, where Z (the partition function) is the value that results in the PD summing or integrating to 1: $Z = \int \tilde{p}(\mathbf{x})\, d\mathbf{x}$.
- In undirected models, it is possible to specify the factors in such a way that Z does not exist. This happens if some of the variables in the model are continuous and the integral of $\tilde{p}$ over their domain diverges.

- A convenient way to enforce the undirected model's assumption that: $\forall \mathbf{x}$, $\tilde{p}(\mathbf{x}) > 0$, is to use an energy-based model (EBM) where $\tilde{p}(\mathbf{x}) = \exp(-E(\mathbf{x}))$, and $E(\mathbf{x})$ is known as the energy function.
- Any distribution of the form given by the previous equation is an example of a Boltzmann distribution.
- Many algorithms that operate on probabilistic models need to compute not $p_{model}(\boldsymbol{x})$ but only $\log \tilde{p}_{model}(\boldsymbol{x})$.
- Separation: conditional independence implied by a graph, there is no requirement that the graph imply all independences that are present. It is said that a set of variables $\mathbb{A}$ is separated from another set of variables $\mathbb{B}$ given a third set of variables $\mathbb{S}$ if the graph structure implies that $\mathbb{A}$ is independent from $\mathbb{B}$ given $\mathbb{S}$. Similar concepts apply to directed models, but in that context they are referred to as d-separation.
- We may choose to use either directed modeling or undirected modeling based on which approach can capture the most independences in the PD or which uses the fewest edges to describe the distribution.
- Only directed models can represent a structure called immortality. It occurs when two random variables a and b are both parents of a third random variable c, and there is no edge directly connecting a and b in either direction.
- To convert a directed model with graph $\mathcal{D}$ into an undirected model, we need to create a new graph $\mathcal{U}$. For every pair of variables x and y, we add an undirected edge connecting x and y to $\mathcal{U}$ if there is a directed edge (in either direction) connecting x and y in $\mathcal{D}$ or if x and y are both parents in $\mathcal{D}$ of a third variable z. The resulting $\mathcal{U}$ is known as a moralized graph.
- A directed graph $\mathcal{D}$ cannot capture all the conditional independences implied by an undirected graph $\mathcal{U}$ if $\mathcal{U}$ contains a loop of length greater than three, unless that loop contains a chord (a connection between any nonconsecutive variables in the sequence defining a loop).
- A factor graph is a graphical representation of an undirected model that consists of a bipartite undirected graph.
- One advantage of directed graphical models is that a simple and efficient procedure called ancestral sampling can produce a sample from the joint distribution represented by the model. The basic idea is to sort the variables x_i in the graph into a topological ordering so that for all i and j, j is greater than i if x_i is a parent of x_j. The variables can then be sampled in this order.
- A good generative model needs to accurately capture the distribution over the observed, or "visible", variables $\mathbf{v}$. Often the different elements of $\mathbf{v}$ are highly dependent on each other. In the context of deep learning, the approach most commonly used to model these dependencies is to introduce several latent or "hidden" variables, $\mathbf{h}$. The model can then capture dependencies between any pair of variables v_i and v_j indirectly, via direct dependencies between v_i and $\mathbf{h}$, and direct dependencies between $\mathbf{h}$ and v_j.
- Roughly, structure learning, is to connect those variables that are tightly coupled and omit edges between other variables.
- Inference problems: we must predict the value of some variables given other variables, or predict the PD over some variables given the value of other variables.
- Deep learning essentially always makes use of the idea of distributed representations. Even shallow models nearly always have a single layer of latent variables. Deep learning models typically have more latent variables than observed variables.
- Models used in deep learning tend to connect each visible unit v_i to many hidden units h_j, so that $\mathbf{h}$ can provide a distributed representation of v_i (and probably several other observed variables too).
- The deep learning approach to graphical modeling is characterized by a marked tolerance of the unknown. The power of the model is increased until it is just barely possible to train or use.

References

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