Structured Probabilistic Models for Deep Learning

Lecture slides for Chapter 16 of *Deep Learning*
www.deeplearningbook.org
Ian Goodfellow
2016-10-04
Roadmap

- Challenges of Unstructured Modeling
- Using Graphs to Describe Model Structure
- Sampling from Graphical Models
- Advantages of Structured Modeling
- Structure Learning and Latent Variables
- Inference and Approximate Inference
- The Deep Learning Approach to Structured Probabilistic Modeling
Tasks for Generative Models

- Density estimation
- Denoising
- Sample generation
- Missing value imputation
  - Conditional sample generation
  - Conditional density estimation
Samples from a BEGAN

(Berthelot et al, 2017)

Images are 128 pixels wide, 128 pixels tall
R, G, and B pixel at each location.
Cost of Tabular Approach

\[ k^n \]

- Number of variables
  - For BEGAN faces: \( 128 \times 128 = 16384 \)

- Number of values per variable
  - For BEGAN faces: 256

There are roughly ten to the power of forty thousand times more points in the discretized domain of the BEGAN face model than there are atoms in the universe.
Tabular Approach is Infeasible

- Memory: cannot store that many parameters
- Runtime: inference and sampling are both slow
- Statistical efficiency: extremely high number of parameters requires extremely high number of training examples
Roadmap

• Challenges of Unstructured Modeling

• Using Graphs to Describe Model Structure

• Sampling from Graphical Models

• Advantages of Structured Modeling

• Structure Learning and Latent Variables

• Inference and Approximate Inference

• The Deep Learning Approach to Structured Probabilistic Modeling
Insight of Model Structure

- Most variables influence each other
- Most variables do not influence each other \textit{directly}
- Describe influence with a graph
  - Edges represent \textit{direct influence}
  - Paths represent \textit{indirect influence}
- Computational and statistical savings come from \textit{omissions of edges}
As we saw earlier, our estimate of Alice finishes. Likewise, Carol only gets to start running after Bob finishes, so Bob's values. This means that using the directed graphical model reduced our number of values, and so does the table defining parameters by a factor of more than 50!

In our relay race example, this means that, using the graph drawn in figure 16.2, illustrated in figure 16.2:

\[
p(x) = \prod_i p(x_i \mid Pa_G(x_i)).
\]

\[
p(t_0, t_1, t_2) = p(t_0)p(t_1 \mid t_0)p(t_2 \mid t_1).
\]

Directed models work best when influence clearly flows in one direction.

(Goodfellow 2017)
Undirected Models

Undirected models work best when influence has no clear direction or is best modeled as flowing in both directions.

\[ \tilde{p}(x) = \prod_{C} \theta(C) \]

Figure 16.3: An undirected graph representing how your roommate's health, your health, and your work colleague's health affect each other. You and your roommate might infect each other with a cold, and you and your work colleague might do the same, but assuming that your roommate and your colleague do not know each other, they can only infect each other indirectly via you.

Our example of the cold spreading between you, your roommate, and your colleague contains two cliques. One clique contains \( h_y \) and \( h_c \). The factor for this clique can be defined by a table and might have values resembling these:

<table>
<thead>
<tr>
<th>( h_y )</th>
<th>( h_c )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>21</td>
</tr>
<tr>
<td>1</td>
<td>10</td>
</tr>
</tbody>
</table>

A state of 1 indicates good health, while a state of 0 indicates poor health (having been infected with a cold). Both of you are usually healthy, so the corresponding state has the highest affinity. The state where only one of you is sick has the lowest affinity, because this is a rare state. The state where both of you are sick (because one of you has infected the other) is a higher affinity state, though still not as common as the state where both are healthy.
Undirected Models

Unnormalized probability

\[ \tilde{p}(\mathbf{x}) = \prod_{C \in G} \phi(C). \]  

(16.3)

Partition function

\[ Z = \int \tilde{p}(\mathbf{x}) d\mathbf{x}. \]  

(16.5)

\[ p(\mathbf{x}) = \frac{1}{Z} \tilde{p}(\mathbf{x}), \]  

(16.4)
Separation

When $s$ is not observed, influence can flow from $a$ to $b$ and vice versa through $s$.

When $s$ is observed, it blocks the flow of influence between $a$ and $b$: they are separated.
Separation example

The nodes a and c are separated

One path between a and d is still active, though the other path is blocked, so these two nodes are not separated.
d-separation

The flow of influence is more complicated for directed models. The path between a and b is active for all of these graphs:

- a → s → b
- a ← s ← b
- a → s ← b
- a ← s → b
- a ← s ← b
- a → s ← b
- a ← s → b
- a ← s ← b
- a ← s ← b
d-separation example

Figure 16.9: From this graph, we can read out several d-separation properties. Examples include:

- $a$ and $b$ are d-separated given the empty set
- $a$ and $e$ are d-separated given $c$
- $d$ and $e$ are d-separated given $c$

Observing variables can activate paths!

- $a$ and $b$ are not d-separated given $c$
- $a$ and $b$ are not d-separated given $d$
A complete graph can represent any probability distribution.

In the undirected case, the complete graph is unique.

In the directed case, there is not a unique complete graph. We choose an ordering of the variables and draw an arc from each variable to every variable that comes after it in the ordering. There are thus a factorial number of complete graphs for every set of random variables. In this example, we order the variables from left to right, top to bottom.

Directed models and undirected models both have their advantages and disadvantages. Neither approach is clearly superior and universally preferred. Instead, we should choose which language to use for each task. This choice will partially depend on which probability distribution we wish to describe. We may choose to use either directed modeling or undirected modeling based on which approach can capture the most independences in the probability distribution or which approach uses the fewest edges to describe the distribution. Other factors can affect the decision of which language to use. Even while working with a single probability distribution, we may sometimes switch between different modeling languages. Sometimes a different language becomes more appropriate if we observe a certain subset of variables, or if we wish to perform a different computational task. For example, the directed model description often provides a straightforward approach to efficiently draw samples from the model (described in section 16.3), while the undirected model formulation is often useful for deriving approximate inference procedures (as we will see in chapter 19, where the role of undirected models is highlighted in equation 19.56).

Every probability distribution can be represented by either a directed model or an undirected model. In the worst case, one can always represent any distribution by using a “complete graph.” For a directed model, the complete graph is any directed acyclic graph in which we impose some ordering on the random variables, and each variable has all other variables that precede it in the ordering as its ancestors in the graph. For an undirected model, the complete graph is simply a graph containing a single clique encompassing all the variables. See figure 16.10 for an example.

The benefits of graphical models come from omitting edges.
Converting between graphs

- Any specific probability distribution can be represented by either an undirected or a directed graph.

- Some probability distributions have conditional independences that one kind of graph fails to imply (the distribution is simpler than the graph describes; need to know the conditional probability distributions to see the independences).
Converting directed to undirected

Figure 16.11: Examples of converting directed models (top row) to undirected models (bottom row) by constructing moralized graphs.

Must add an edge between unconnected coparents
Converting undirected to directed

No loops of length greater than three allowed!

Add edges to triangulate long loops

Assign directions to edges. No directed cycles allowed.
Factor graphs are less ambiguous

Undirected graph: is this three pairwise potentials or one potential over three variables?

Factor graphs disambiguate by placing each potential in the graph

Figure 16.13: An example of how a factor graph can resolve ambiguity in the interpretation of undirected networks.


Graphical models also facilitate the task of drawing samples from a model. 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. In other words, we first sample \( x_1 \sim P(x_1) \), then sample \( x_2 \sim P(x_2 | Pa_G(x_2)) \), and so on, until finally we sample \( x_n \sim P(x_n | Pa_G(x_n)) \).

So long as each conditional distribution \( p(x_i | Pa_G(x_i)) \) is easy to sample from, then the whole model is easy to sample from. The topological sorting operation guarantees that we can read the conditional distributions in equation 16.1 and sample from them in order. Without the topological sorting, we might attempt to sample a variable before its parents are available.
Roadmap

- Challenges of Unstructured Modeling
- Using Graphs to Describe Model Structure
- Sampling from Graphical Models
- Advantages of Structured Modeling
- Structure Learning and Latent Variables
- Inference and Approximate Inference
- The Deep Learning Approach to Structured Probabilistic Modeling
Sampling from directed models

• Easy and fast to draw fair samples from the whole model

• *Ancestral sampling:* pass through the graph in topological order. Sample each node given its parents.

• Harder to sample some nodes given other nodes, unless the observed nodes are at the start of the topology
Sampling from undirected models

- Usually requires Markov chains
- Usually cannot be done exactly
- Usually requires multiple iterations even to approximate
- Described in Chapter 17
Roadmap

• Challenges of Unstructured Modeling

• Using Graphs to Describe Model Structure

• Sampling from Graphical Models

• Advantages of Structured Modeling

• Structure Learning and Latent Variables

• Inference and Approximate Inference

• The Deep Learning Approach to Structured Probabilistic Modeling
Tabular Case

• Assume each node has a tabular distribution given its parents

• Memory, sampling, inference are now exponential in *number of variables in factor with largest scope*

  • For many interesting models, this is very small

  • e.g., RBMs: all factor scopes are size 2 or 1

• Previously, these costs were exponential in *total number of nodes*

• Statistically, much easier to estimate this manageable number of parameters
Roadmap

• Challenges of Unstructured Modeling
• Using Graphs to Describe Model Structure
• Sampling from Graphical Models
• Advantages of Structured Modeling
  • Structure Learning and Latent Variables
• Inference and Approximate Inference
• The Deep Learning Approach to Structured Probabilistic Modeling
Learning about dependencies

• Suppose we have thousands of variables
  • Maybe gene expression data
• Some interact
• Some do not
• We do not know which ahead of time
Structure learning strategy

• Try out several graphs

• See which graph does best job of some criterion
  • Fitting training set with small model complexity
  • Fitting validation set

• Iterative search, propose new graphs similar to best graph so far (remove edge / add edge / flip edge)
Latent variable strategy

• Use one graph structure

• Many latent variables

• Dense connections of latent variables to observed variables

• Parameters learn that each latent variable interacts strongly with only a small subset of observed variables

• Trainable just with gradient descent; no discrete search over graphs
Roadmap

• Challenges of Unstructured Modeling

• Using Graphs to Describe Model Structure

• Sampling from Graphical Models

• Advantages of Structured Modeling

• Structure Learning and Latent Variables

• Inference and Approximate Inference

• The Deep Learning Approach to Structured Probabilistic Modeling
Inference and Approximate Inference

- Inferring marginal distribution over some nodes or conditional distribution of some nodes given other nodes is $\#P$ hard

- NP-hardness describes decision problems. $\#P$-hardness describes counting problems, e.g., how many solutions are there to a problem where finding one solution is NP-hard

- We usually rely on approximate inference, described in chapter 19
Roadmap

- Challenges of Unstructured Modeling
- Using Graphs to Describe Model Structure
- Sampling from Graphical Models
- Advantages of Structured Modeling
- Structure Learning and Latent Variables
- Inference and Approximate Inference
- The Deep Learning Approach to Structured Probabilistic Modeling
Deep Learning Stylistic Tendencies

- Nodes organized into layers
- High amount of connectivity between layers
- Examples: RBMs, DBMs, GANs, VAEs

![Diagram of an RBM](image)

Figure 16.14: An RBM drawn as a Markov network.
For more information...