Notes of PRML

Siyu Wang

August 2018

Chapter 1

Mathematical Foundations

Chapter 2

Probabilistic Models

2.1 Graphical Models

We shall find it highly advantageous to augment the analysis using diagrammatic representations of probability distributions, called *probabilistic graphical models*. These offer several useful properties:

1. They provide a simple way to visualize the structure of a probabilistic model and can be used to design and motivate new models.
2. Insights into the properties of the model, including conditional independence properties, can be obtained by inspection of the graph.
3. Complex computations, required to perform inference and learning in sophisticated models, can be expressed in terms of graphical manipulations, in which underlying mathematical expressions are carried along implicitly.

nodes: random variables. *links*: probabilistic relationships between the variables.
Bayesan networks (directed graphical models), *undirected graphical models*, *factor graph*.

2.1.1 Bayesian Networks

Decomposition of a joint probability distribution:

$$p(a, b, c) = p(c|a, b)p(b|a)p(a) \tag{8.1}$$

and the corresponding graphical model can be seen in figure2.1.
Rule: for each conditional distribution we add directed links (arrows) to the graph from the nodes corresponding to the variables on which the distribution is conditioned.
we can extend the decomposition to K variables:

$$p(x_1, ..., x_K) = p(x_K|x_1, ..., x_{K-1})...p(x_2|x_1)p(x_1). \tag{2.1}$$

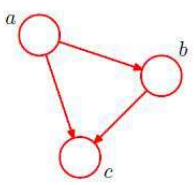

Figure 2.1:

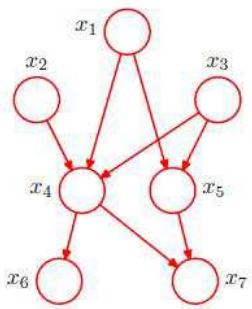

Figure 2.2:

And the corresponding directed graph is called *fully connected* because there is a link between every pair of nodes.
Case Given a directed graph as seen in 2.2, we can write the corresponding probability product:

$$p(x_1)p(x_2)p(x_3)p(x_4|x_1,x_2,x_3)p(x_5|x_2,x_3)p(x_6|x_4)p(x_7|x_4,x_5) \tag{2.2}$$

Rule:

$$p(\mathbf{x}) = \Pi_{k=1}^{K} p(x_k|\mathrm{pa}_k) \tag{2.3}$$

where pa_k denotes the set of parents of x_k, and $\mathbf{x} = \{x_1, .., x_K\}$.
The directed graphs that we are considering are subject to an important restriction namely that there must be no *directed cycles*, in other words there are no closed paths within the graph such that we can move from node to node along links following the direction of the arrows and end up back at the starting node. Such graphs are also called *directed acyclic graphs*, or *DAGs*. This is equivalent to the statement that there exists an ordering of the nodes such that there are no links that go from any node to any lower numbered node.

Example: polynomial regression

For more complex models, we shall adopt the convention that random variables will be denoted by open circles, and deterministic parameters will be denoted by smaller solid circles. some concepts: *observed variables, hidden variables, deterministic parameters.*

Generative models

sampling: given a joint distribution, we want to draw a sample $\hat{x}_1, \hat{x}_2, ..., \hat{x}_K$ from it. **ancestral sampling**
We start with the lowest-numbered node and draw a sample from the distribution $p(x_1)$, which we all $\hat{x}_1$. Then we work through each of the nodes in order, so that for node n we draw a sample from the conditional distribution p(xn—pan) in which the parent variables have been set to their sampled values. Note that at each stage, these parent values will always be available because they correspond to lower-numbered nodes that have already been sampled. The graphical model captures the *causal* process by which the observed data was generated. For this reason, such models are often called *generative* models. But the regression problem is not generative because there is no probability distribution associated with the input variable x, and it is not possible to generate synthetic data points from this model. But we can make it generative by introducing a suitable prior distribution $p(x)$, at the expense of a more complex model.

Discrete variables

parent-child pair in a directed graph. Two cases are particularly worthy of note, namely when the parent and child node each correspond to discrete variables and when they each correspond to Gaussian variables, because in these two cases the relationship can be extended hierarchically to construct arbitrarily complex directed acyclic graphs.
Case: discrete variable $\mathbf{x}$ having K possible states is given by

$$p(\mathbf{x}|\mathbf{mu}) = \Pi_{k=1}^{K} \mu_k^{x_k} \tag{2.4}$$

governed by μ. Suppose we have to discrete variables $\mathbf{x}_1, \mathbf{x}_2$, joint distribution can be written as

$$p(\mathbf{x}_1, \mathbf{x}_2|\mu) = \Pi_{k=1}^{K} \Pi_{k=1}^{K} \mu_{kl}^{x_{1k}x_{2l}}. \tag{2.5}$$

x_{1k} denotes the k^{th} component of $\mathbf{x}_1$.
$K^2 - 1$ parameters. If we have M variables, there is $K^M - 1$ parameters, exponential with the num M.
Suppose $\mathbf{x}_1$ and $\mathbf{x}_2$ are independent. Each variable is then described by a separate multinomial distribution, and total number of parameters will be $2(K-1)$. If extend to M variables, there will be $M(K-1)$ variables, linear growth.
Generally case: more links than independent variables but less links than a fully connected graph.
We can turn a graph over discrete variables into a Bayesian model by introducing Dirichlet priors

for the parameters.
Another way of controlling the exponential growth in the number of parameters in models of discrete variables is to use parameterized models for the conditional distributions instead of complete tables of conditional probability values.

Linear-Gaussian models

2.1.2 Conditional Independence

If

$$p(a|b,c) = p(a|c), \tag{2.6}$$

we say that a is conditionally independent of b given c, and we denote this by

$$a \perp\!\!\!\perp b|c.$$

And we will have

$$p(a,b|c) = p(a|c)p(b|c). \tag{2.7}$$

Examples

Three examples related to conditional independence: *tail-to-tail, head-to-tail, head-to-head.*

D-separation

We wish to ascertain whether a particular conditional independence statement

$$A \perp\!\!\!\perp B|C$$

is implied by a given directed acyclic graph.
i.i.d data points. Given μ, we can say the data points $x_1, x_2, ..., x_N$ are conditionally independent, but we can not say the data points are independent. because given x_1, the probability of μ will be affected and then x_2 will be affected.
naive Bayes Model Observation of $\mathbf{z}$ will block the path between x_i and x_j for $j \neq i$. So, if we are given a training set will, comprising inputs $\{x_1, ..., x_N\}$ together with their labels, then we can fit the naive Bayes model to the training data using maximum likelihood assuming that the data are drawn independently from the model.
We can view a directed graph as a filter and all probability distribution $p(\mathbf{x})$ that will be allowed through can make a set $\mathcal{DF}$.
Markov blanket.

$$p(\mathbf{x}_i|\mathbf{x}_{\{j\neq i\}}) = \frac{p(\mathbf{x}_1,...\mathbf{x}_D)}{\int p(\mathbf{x}_1,...,\mathbf{x}_D)\mathrm{d}\mathbf{x}_\mathrm{i}} = \frac{\prod_k p(\mathbf{x}_k|\mathrm{pa}_\mathrm{k})}{\int \prod_k p(\mathbf{x}_k|\mathrm{pa}_\mathrm{k})\mathrm{d}\mathbf{x}_\mathrm{i}} \tag{2.8}$$

2.1.3 Markov Random Fields

2.1.4 Inference in Graphical Models