Probabilistic Graphical Models

Summary
Fabian Damken
February 18, 2022

TECHNISCHE
UNIVERSITÄT
DARMSTADT

Contents

List of Figures

List of Tables

List of Algorithms

1 Introduction

1.1 Examples

1.2 Fundamental Questions

2 Foundations

This chapter covers fundamental concepts of probability theory and machine learning that are required for the later chapters. Note that not all relevant concepts of probability theory are covered.

2.1 Probability Theory

This section covers some very important concepts of probability theory, however, one should already be familiar with some basics like probability measures, density functions, joint distributions, marginalization, etc.[1]

One note on notation: whenever a sum represents a marginalization over some random variable X, it is written as

$$P(Y) = \overset{\text{marg.}}{\sum_{X}} P(X, Y) \coloneqq \sum_{x \in \mathrm{val}(X)} P(X = x, Y)$$

for brevity.

2.1.1 (Conditional) Independence

The most important concept leveraged in probabilistic graphical models is (conditional) independence of random variables. Two random variables X and Y are *statistically independent* if knowing either does not change the belief/probability of the other, i.e.,

$$P(X \,|\, Y) = P(X) \quad \text{and} \quad P(Y \,|\, X) = P(Y).$$

This is equivalent to the definition of independence, $P(X, Y) = P(X)\,P(Y)$. Independence is denoted $X \perp Y$ and is a symmetric properties. A milder property is *conditional* independence, i.e., two random variables X and Y are independent if Z is given:

$$P(X \,|\, Y, Z) = P(X \,|\, Z) \quad \text{and} \quad P(Y \,|\, X, Z) = P(Y \,|\, Z).$$

Again, this property can be written as $P(X, Y \,|\, Z) = P(X \,|\, Z)\,P(Y \,|\, Z)$ by the chain rule. Conditional independency is denoted $X \perp Y \,|\, Z$.

The following properties hold and can be useful for some proofs later on:

$$\begin{aligned}
X \perp Y \,|\, Z &\iff Y \perp X \,|\, Z && \text{(Symmetry)} \\
X \perp (Y, W) \,|\, Z &\iff \big(X \perp Y \,|\, Z\big) \land \big(X \perp W \,|\, Z\big) && \text{(Decomposition)} \\
X \perp (Y, W) \,|\, Z &\implies X \perp Y \,|\, (Z, W) && \text{(Weak Union)} \\
\big(X \perp W \,|\, (Y, Z)\big) \land \big(X \perp Y \,|\, Z\big) &\implies X \perp (Y, W) \,|\, Z && \text{(Contraction)} \\
\big(X \perp Y \,|\, (W, Z)\big) \land \big(X \perp W \,|\, (Y, Z)\big) &\implies X \perp (Y, W) \,|\, Z && \text{(Intersection)}
\end{aligned}$$

[1] Take a look the chapter of statistics fundamentals of `https://fabian.damken.net/summaries/cs/elective/vc/statml/statml-summary.pdf`.

Monty Hall Problem

2.1.2 Inference

Information Theory

Information theory is trying to quantify how much information is encoded in some distribution $P(X)$. The central measure is *entropy*:

$$H_P(X) = \mathbb{E}\big[\log(1/P(X))\big] = \sum_{x \in \text{val}(X)} P(X) \log \frac{1}{\log P(X)} = -\sum_{x \in \text{val}(X)} P(X) \log P(X)$$

If the logarithm is of base two, the entropy encodes how much bits are required *on average* to encode X when X follows the distribution $P(X)$. Similarly, *conditional entropy* can be defined as

$$H_P(X \mid Y) = \mathbb{E}\big[\log(1/P(X \mid Y))\big] = H_P(X, Y) - H_P(X)$$

where $H_P(X, Y)$ is the joint entropy over X and Y. Like for probabilities, a *chain rule of entropies* is derivable:

$$H_P(X, Y, Z) = H_P(X) + H_P(Y \mid X) + H_P(Z \mid X, Y).$$

To quantify (in)dependency between two variables X and Y, the *mutual information*

$$I_P(X; Y) = H_P(X) - H_P(X \mid Y)$$

can be used. This quantity is symmetric and is zero if and only if X and Y are independent.

2.1.3 Potentials

A *potential* is an alternative way of representing (conditional) probabilities aside from conditional probability tables (CPTs). A potential $\phi_{X,Y,Z}$ is is a function that maps each configuration $(x, y, z) \in \text{val}(X) \times \text{val}(Y) \times \text{val}(Z)$ to a non-negative real number. The set of random variables targeted by a potential is its *domain*, i.e., dom $f_{X,Y,Z} = \{X, Y, Z\}$. Note that a (conditional) probability distribution is a special case of potentials where the potential is normalized. Vice versa, a potential can always be normalized into a CPT. This is illustrated in Figure 2.1.

Similar to CPTs, potentials can be multiplied by pairing up the entries and marginalized by summing up the corresponding entries. Compared to CPTs, it is not necessary to normalize a potential into a probability distribution afterwards, easing some calculations.

Potentials will come in handy later on when covering inference in junction trees (section 5.4).

2.2 Machine Learning

2.2.1 (Document) Classification

$P(X = x \mid Y = y, Z = z)$	$x = \mathfrak{f}$	$x = \mathfrak{t}$
$(Y, Z) = (\mathfrak{t}, \mathfrak{t})$	0.8	0.2
$(Y, Z) = (\mathfrak{t}, \mathfrak{f})$	0.5	0.5
$(Y, Z) = (\mathfrak{t}, \mathfrak{t})$	0.2	0.8
$(Y, Z) = (\mathfrak{t}, \mathfrak{f})$	0.7	0.3

$\longleftrightarrow$

X	Y	Z	$\phi_{X,Y,Z}$	$\tilde{\phi}_{X,Y,Z}$
$\mathfrak{t}$	$\mathfrak{t}$	$\mathfrak{t}$	0.8	8
$\mathfrak{t}$	$\mathfrak{t}$	$\mathfrak{f}$	0.2	2
$\mathfrak{t}$	$\mathfrak{f}$	$\mathfrak{t}$	0.5	5
$\mathfrak{t}$	$\mathfrak{f}$	$\mathfrak{f}$	0.5	5
$\mathfrak{f}$	$\mathfrak{t}$	$\mathfrak{t}$	0.2	2
$\mathfrak{f}$	$\mathfrak{t}$	$\mathfrak{f}$	0.8	8
$\mathfrak{f}$	$\mathfrak{f}$	$\mathfrak{t}$	0.7	7
$\mathfrak{f}$	$\mathfrak{f}$	$\mathfrak{f}$	0.3	3

Figure 2.1: Comparison of a CPT (left) and the corresponding potential (right). The rightmost column in the potential $\tilde{\phi}$ is equivalent to ϕ as it can be normalized accordingly.

3 Bayesian Networks

3.1 The Naive Bayes Model

3.1.1 Classification

3.1.2 Maximum Likelihood Parameter Estimation

3.1.3 Application

3.2 Definition and Independence Assumptions

3.2.1 Local Markov Assumption

3.2.2 "Explaining Away" / Berkson's Paradox

3.2.3 Representation Theorem

3.2.4 Building a Bayesian Network

3.3 Encoded Independencies

3.3.1 Dependency Structures

3.3.2 d-Separation

(Active) Trails

Independencies

Soundness

Completeness

3.3.3 Faithful Distributions

3.3.4 Context-Specific Independence (CSI)

Tree CPD

Determinism

3.3.5 The Bayes' Ball Algorithm

4 Inference

4.1 Chain Models

4.2 Variable Elimination

4.2.1 Evidence

4.2.2 Complexity

4.2.3 VE for Potentials

4.3 Abductive Inference

4.3.1 Consistency

4.3.2 Finding Most Probable Explanations (MPEs)

4.4 Complexity of Conditional Queries

4.5 Moralizing

4.6 Variable Elimination in Moral Graphs

4.6.1 Perfect Elimination Sequences

4.6.2 Complexity

4.6.3 Induced Graph

4.6.4 Induced Treewidth

4.6.5 Elimination on Trees

Polytrees

4.6.6 General Networks

5 Markov Random Fields

5.1 Bayesian Networks as MRFs

5.2 Triangulated Graphs

5.3 Join Trees

5.4 Junction Trees

5.4.1 Collecting Evidence

5.4.2 Distributing Evidence

5.5 Non-Triangulated Graphs

6 Learning

6.1 Complete and Incomplete Data Sets

6.1.1 Hidden Variables

6.2 Parameter Estimation

6.2.1 Known Structure, Complete Data

Maximum Likelihood

Decomposability of the Likelihood

Likelihood for (Conditional) Bi- and Multinomials

6.2.2 Known Structure, Incomplete Data (Expectation-Maximization)

EM Idea

Complete-Data Likelihood

EM for (Conditional) Multinomials

Monotonicity

6.2.3 Gradient Ascent

6.2.4 Bayesian Parameter Estimation

Laplace Estimation

Bayesian Prediction

Conjugate Priors

Binomial Prior

Dirichlet Prior

Bayesian Networks and Bayesian Prediction

6.2.5 Summary

6.3 Structure Learning / Model Selection

6.3.1 Minimal I-Maps

6.3.2 Perfect Maps (P-Maps)

6.3.3 I-Equivalence

Skeleton and Immoralities

6.3.4 Obtaining a P-Map

Identifying the Skeleton

Identifying Immoralities

From Immoralities to Structures

6.3.5 Accurate Structures

6.3.6 Learning

Constrained-Based

Score-Based

Likelihood Score

Bayesian Score and Bayesian Information Criterion

6.3.7 Structure Search as Optimization

Learning Trees (Complete Data)

Heuristic (Local) Search

6.3.8 Structural EM

6.3.9 Summary

7 Dynamic Bayesian Networks

7.1 Hidden Markov Models

7.2 Inference

7.2.1 Decoding

Forward Pass

Backward Pass

7.2.2 Best State Sequence

Viterbi Algorithm

7.2.3 Parameter Estimation

7.3 State Estimation (Kalman Filter)

7.3.1 Recursive Bayesian Updating

7.3.2 (Modeling) Actions

7.3.3 Bayes Filter

7.3.4 Discrete-Time Kalman Filter

Dynamics and Observations

Belief Update: Prediction

Belief Update: Correction

7.4 General Dynamic Bayesian Networks

7.4.1 Exact Inference

7.4.2 Tractable, Approximate Inference

Assumed Density Filtering

8 Approximate Inference

8.1 Message Passing

8.1.1 Sum-Product Belief Propagation

8.1.2 (Acyclic) Belief Propagation as Dynamic Programming

8.1.3 Loopy Belief Propagation

8.2 Sampling

8.2.1 Forward Sampling (Without Evidence)

8.2.2 Forward Sampling (With Evidence)

8.2.3 Gibbs Sampling

Burn-In

Irreducibility, Aperiodicity, and Ergodicity

Convergence

Performance

Speeding Convergence

Skipping Samples

Randomized Variable Order

Blocking

Rao-Blackwellization

Multiple Chains

8.2.4 Likelihood Weighting

9 Tractable Probabilistic Models

9.1 Deep Learning

9.2 Probabilistic Circuits

9.3 Sum-Product Networks

9.3.1 Inference

9.3.2 Learning

Directly Learning SPNs

9.3.3 Inference on Devices

10 Deep Generative Models

10.1 Likelihood-Based

10.1.1 Autoregressive Generatie Models

Learning and Inference

Parametrization

10.1.2 Variational Auto-Encoders

Inference as Optimization

Variational Bayes

Learning and Inference

Open Questions

10.1.3 Normalizing Flows

Learning and Inference

10.2 Likelihood-Free

10.2.1 Generative Adversarial Networks

Inference

10.3 Applications in Scientific Discovery