MAC6916 PROBABILISTIC GRAPHICAL MODELS
LECTURE 1: PROBABILITY THEORY

DENIS D. MAUÁ

1. Course Overview

1.1. **Objective.** To *guide* graduate students in their study of the *algorithmic* aspects of (discrete) probabilistic graphical models, in particular Bayesian networks, and their application in machine learning tasks, especially in classification. At the end of the course, students are expected to have built their own programming library implementing the basic tasks of graph-based probabilistic reasoning such as independence testing, inference and learning.

1.2. **Organization.** Roughly one lecture (mostly Tuesdays) and one assignment per week.

1.3. **Assignments.**

- Each assignment is composed of reading of recommended material (chapters of books and papers), answering of exercises, implementation of related code, and short report summarizing the topic.
- The report can be written either in Portuguese or in English, and must be saved in pdf; It is strongly recommended to use LaTeX.
- You can write the solutions to the exercises in the report.
- Documentation of your code can be in either a text file (e.g., README, HOWTO, man pages) or pdf (which may also appear together with your report).
- The description of the assignment will be made available on the website (paca.ime.usp.br).
- You should submit report/exercise/code in electronic form via website; best option is to create a compressed package containing everything.

1.4. **Grading.** Each assignment is graded 0,1,2,3, standing for missing, insufficient, sufficient, and excellent. The grading considers the quality of your report (clarity, soundness, depth, breadth, style) and of your code (correctness, cleverness, documentation).

Your final grade will be based on your average assignment grade (not submitting an assignment counts as 0). There is also the possibility of being asked to write a term paper on a specific topic to complement the assessment.

Date: February 22, 2016.

1.5. **Required skills.**

- Proficiency in some programming language.
- Basic (undergrad) mathematical skills (calculus, combinatorics, etc)
- Commitment and autonomy (attending lectures is optional; handing in assignments on time is *not*).

1.6. **Course textbooks.**

- Adnan Darwiche, Modeling and Reasoning with Bayesian networks, Cambridge Press, 2009.
- Daphne Koller and Nir Friedman, Probabilistic Graphical Models: Principles and Techniques, MIT Press, 2009.

2. Fundamentals of Probability Theory

2.1. **Outline.**

1. Probabilistic models
2. Structured Probabilistic models

2.2. **Probabilistic models.** We denote the power set (i.e., the set of all subsets of) of a set Ω as 2^Ω.

Definition 1. A *field of sets* is a pair $(\Omega, \mathcal{F})$ where $\mathcal{F}$ is a non-empty collection of subsets of Ω closed under intersection, union and complement. The elements of $\mathcal{F}$ are here called *events* (and Ω is called the possibility space, sample space or the universe).

If $\mathcal{F}$ is closed under countable intersections and unions, then it is a sigma-algebra, and the field of sets is called a measurable space. Measurable spaces are required to formalize probability theory on infinite spaces, but are not necessary for finite domains. As we will mostly deal with finite domains in this course, we do not require countable additivity.

Example 1. *For any set Ω, $(\Omega, \{\emptyset, \Omega\})$ and $(\Omega, 2^\Omega)$ are field of sets. The former is called the trivial algebra and the latter is called the discrete algebra. For example, if $\Omega = \mathbb{Z}_2 = \{0, 1\}$, then $(\Omega, \{\emptyset, \Omega\})$ and $(\Omega, \{\emptyset, \{0\}, \{1\}, \Omega\})$ are fields of sets.*

Definition 2. A *probabilistic model* (also called probability space) is a tuple $(\Omega, \mathcal{F}, \mathbb{P})$, where Ω is a finite set of *atomic events* (also called atoms or worlds), $(\Omega, \mathcal{F})$ is a field of sets and $\mathbb{P}$ is a real-valued (probability) function on $\mathcal{F}$ such that

(A1) $\mathbb{P}(\alpha) \geq 0$ [probabilities are positive];
(A2) $\mathbb{P}(\Omega) = 1$ [some event will occur with probability one];
(A3) $\mathbb{P}(\alpha \cup \beta) = \mathbb{P}(\alpha) + \mathbb{P}(\beta)$ for any disjoint events α and β [probability is additive].

Let $\mathbb{Z}_2 = \{0, 1\}$ and $\mathbb{Z}_2^n = \mathbb{Z}_2 \times \cdots \times \mathbb{Z}_2$, with n products.

Example 2. *A simple probabilistic model of the toss of a loaded coin is $(\mathbb{Z}_2, 2^{\mathbb{Z}_2}, \mathbb{P})$, where 0 represent heads, 1 represents tails, and $\mathbb{P}(\emptyset) = 0, \mathbb{P}(\{0\}) = 1/5, \mathbb{P}(\{1\}) = 4/5, \mathbb{P}(\mathbb{Z}_2) = 1$.*

Proposition 1. *The following facts are true:*

(1) For any event α, it follows that $\mathbb{P}(\alpha) = 1 - \mathbb{P}(\alpha^c)$, where α^c is the complement of α.
(2) $\mathbb{P}(\emptyset) = 0$.

(3) *If $\alpha \subseteq \beta$ are events then $\mathbb{P}(\alpha) \leq \mathbb{P}(\beta)$. This is called* monotonicity*.*
(4) *For any event α, we have that $0 \leq \mathbb{P}(\alpha) \leq 1$.*
(5) *For any events α and β (not necessarily disjoint), we have that $\mathbb{P}(\alpha \cup \beta) = \mathbb{P}(\alpha) + \mathbb{P}(\beta) - \mathbb{P}(\alpha \cap \beta)$.*
(6) *If $\mathbb{P}(\alpha) = 1$ for some event α then $\mathbb{P}(\beta) = \mathbb{P}(\alpha \cap \beta)$ for any event β.*

Definition 3. A *distribution* on Ω is a non-negative real-valued function p on Ω such that $\sum_{\omega \in \Omega} p(\omega) = 1$.

Proposition 2. *Let $(\Omega, \mathcal{F})$ be a field of sets and p be a distribution on Ω. Then $\mathbb{P}(\alpha) = \sum_{\omega \in \alpha} p(\omega)$ is a probability function.*

Proof. It is straightforward to see that $\mathbb{P}$ satisfies (A1) and (A2). To show that it also satisfies (A3), take two disjoint events α and β. Then $\mathbb{P}(\alpha \cup \beta) = \sum_{\omega \in \alpha \cup \beta} p(\omega) = \sum_{\omega \in \alpha} p(\omega) + \sum_{\omega \in \beta} p(\omega) = \mathbb{P}(\alpha) + \mathbb{P}(\beta)$. $\square$

The result above shows that one can specify a probability function by specifying (a much more succinct) distribution.

Definition 4. The *conditional probability* $\mathbb{P}(\alpha|\beta)$ of an event α given an event β is

$$\mathbb{P}(\alpha|\beta) = \frac{\mathbb{P}(\alpha \cap \beta)}{\mathbb{P}(\beta)}.$$

Note that the conditional probability when $\mathbb{P}(\beta) = 0$ is undefined.

Proposition 3 (Chain Rule)**.** *For any sequence of events $\alpha_1, \ldots, \alpha_n$ we have that*

$$\mathbb{P}(\alpha_1 \cap \cdots \cap \alpha_n) = \mathbb{P}(\alpha_1) \prod_{i=1}^{n} \mathbb{P}(\alpha_i | \alpha_1 \cap \cdots \cap \alpha_{i-1}).$$

Proposition 4 (Total Probability Rule)**.** *Let $\alpha_1, \ldots, \alpha_n$ be a partition of Ω (i.e., $\cup_i \alpha_i = \Omega$ and $\alpha_i \cap \alpha_j = \emptyset$ for $i \neq j$, $\alpha_i \neq \emptyset$). Then for any event β it follows that*

$$\mathbb{P}(\beta) = \sum_{i=1}^{n} \mathbb{P}(\beta|\alpha_i)\mathbb{P}(\alpha_i).$$

Proposition 5 (Bayes' rule)**.** *Let α and β be events of positive probability. Then,*

$$\mathbb{P}(\beta|\alpha) = \frac{\mathbb{P}(\alpha|\beta)\mathbb{P}(\beta)}{\mathbb{P}(\alpha)}.$$

Bayes' rule is a simple corollary of the Chain rule and the definition of probability theory. Its power lies in a process called *conditioning*. Let $\mathbb{P}$ and $\mathbb{P}'$ be probability functions before and after receiving information α. Then Bayes' rule suggests that $\mathbb{P}'(\beta) = \mathbb{P}(\beta|\alpha)$. In other words, Bayes' rule prescribes a way to update our knowledge about the world (represented as a probabilistic model) in the light of new information. This is strictly connected to a subjective view of probabilities and to what is called *Bayesianism*. For our matters here, Bayes rule is a simple artifact to produce *inferences*.

2.2.1. *Simpson's paradox.* There are many counter-intuitive situations with probability theory. A famous is known as Simpson's paradox, after the mathematician that made it popular. The paradox relies on the fact that probability theory allows $\mathbb{P}(\alpha) > \mathbb{P}(\beta)$ while $\mathbb{P}(\alpha|\gamma) < \mathbb{P}(\beta|\gamma)$. This is a mathematically sound system of inequalities that is often seen as unnatural when probabilities are related to real-world phenomena.

Example 3. *Consider a group of patients who have been given the option of taking a new drug for treating a disease. We can split the patients into eight groups according to whether each person (i) took the drug, (ii) it is female, and (iii) recovered from the illness. We can specify a probabilistic model such that the probability of recovery is given as follows:*

	Drug	*No Drug*
Men	*0.93*	*0.87*
Women	*0.73*	*0.69*
Either	*0.78*	*0.83*

Define the events:

$$\begin{aligned} R &: \text{ the patient recovered} \\ F &: \text{ the patient is female} \\ D &: \text{ the patient took the drug} \end{aligned}$$

The model specifies that $\mathbb{P}(R|F \cap D) > \mathbb{P}(R^{\mathsf{c}}|F \cap D)$, $\mathbb{P}(R|F^{\mathsf{c}} \cap D) > \mathbb{P}(R^{\mathsf{c}}|F^{\mathsf{c}} \cap D)$, *and (maybe surprisingly) that* $\mathbb{P}(R|D) < \mathbb{P}(R^{\mathsf{c}}|D)$.

2.2.2. *Independence.*

Definition 5. We say that events α and β are independent if $\mathbb{P}(\alpha \cap \beta) = \mathbb{P}(\alpha)\mathbb{P}(\beta)$.

Note that any event is defined to be independent of an event with zero probability. A remarkable feature of independence is that it allows the decomposition of a probabilistic model in terms of smaller models.

Example 4. *Consider the independent tosses of five coins. Represent each coin by an integer* $i \in \{1, \ldots, 5\}$, *and define for each coin a probabilistic model with universe* $\Omega_i = \{0, 1\}$ *and probability function* $\mathbb{P}_i$. *Now consider a* joint *probabilistic model that considers all the coins. Its universe is* $\Omega = \Omega_1 \times \cdots \times \Omega_5$. *Consider the events* $\alpha_i^{k_i}$ *denoting "the outcome of the ith coin is k", where* $k \in \{0, 1\}$. *As we have assumed all tosses to be independent, it follows that* $\mathbb{P}(\alpha_1^{k_1} \cap \cdots \cap \alpha_5^{k_5}) = \prod_{i=1}^5 \mathbb{P}(\alpha_i^{k_i}) = \prod_{i=1}^5 \mathbb{P}_i(\alpha_i^{k_i})$. *Note that for any choice of* $k_1, \ldots, j_5$, $\alpha_1^{k_1} \cap \cdots \cap \alpha_5^{k_5}$ *is a singleton whose element is in* Ω, *and that each* $\alpha_i^{k_i}$ *is a singleton whose element is in* Ω_i. *Hence, the probability function of the joint model is completely specified by the probability function of the individual models.*

Independence extends to conditional events:

Definition 6. We say that events α and β are conditionally independent given event γ if $\mathbb{P}(\alpha \cap \beta|\gamma) = \mathbb{P}(\alpha|\gamma)\mathbb{P}(\beta|\gamma)$, as long as these probabilities are defined.

Note that the definitions allow for α and β being unconditionally dependent (i.e., $\mathbb{P}(\alpha \cap \beta) \neq \mathbb{P}(\alpha)\mathbb{P}(\beta)$ while being independent conditional on γ; the converse is also possible: we might have that α and β are unconditionally independent and conditionally dependent given some γ (and they may even be conditionally independent given another event γ').

Example 5. *Consider a probabilistic model describing the connection among smoking, bronchitis, lung cancer and dyspnoea. We can build a probabilistic model with universe* $\Omega = \{0, 1\}^4$, *where each dimension represents the presence/absence of a habit/disorder. A probability function reflecting common medical knowledge about* *the domain might encode that bronchitis and lung cancer are unconditionally dependent, conditionally dependent given smoking and conditionally independent given dyspnoea. Also, dyspnoea and smoking are unconditionally dependent, conditionally dependent given either bronchitis or lung cancer alone and conditionally independent given the bronchitis* and *lung cancer.*

2.3. **Direct inference.** Any probabilistic model can be characterized by a distribution on Ω (which in turn can be characterized by simpler functions if we *assume* independences). Thus any (conditional) probability can be obtained by specifying a distribution. While this is always the case, computing a probability induced by a distribution can be computationally challenging. The probablem is better formalized as follows.

Definition 7 (Inference Problem)**.** Given an encoding of a probability distribution p on Ω and two events α and β compute $\mathbb{P}(\alpha|\beta)$, where $\mathbb{P}$ is the probability function induced by p.

We can such a computation a *query*. The event α is *target* while the event β is the *evidence*. Computing the probability is often called "querying the model".

2.4. **Structured Probabilistic Models.** As many of the previous examples have shown, the set-theoretic definition of events is inconvenient for specifying and relating complex events. For example, consider a universe which describes the performance of a student in a certain course. An important aspect of the performance is the student's final grade. For instance, we might reason about the probability of the student having grade "A" or grade "B". We could come up with *ad hoc* notations γ_A and γ_B to represent such events, but any such ad hoc notation would have to be defined and explained for each event. Instead, it is more useful to define a convenient *language* in which complex events can be described. We adopt a generalization of propositional logic that accepts multivalued variables.[1]

The basic ingredients of our language are variables which stand for properties of the domain being described. For example, in the example of the drug treatment, we can variables describing the patient gender, the recovery status, the ingestion of the drug and the patient age. Each variable has a domain: for example, the domain of the variable gender is female or male and the domain of the variable age are the integers between 1 and 120.

We often represent variables with upper case Latin letters: A, B, C, X, Y, Z. When convenient we also use words and expressions, e.g. Smoker, Age and Lung Cancer. We denote the domain of a variable X by either Ω_X or $\mathrm{dom}(X)$. The elements of Ω_X are called values (of X) and often represented with lower case letters: a, b, c, x_1, x_2, true, false. We consider a fixed *vocabulary* of variable symbols $X_1, \ldots, X_n$ and corresponding domains $\mathrm{dom}(X_1), \ldots, \mathrm{dom}(X_n)$.

Definition 8. A valid sentence is defined inductively as follows.

(1) For any variable X and value $x \in \Omega_X$, $(X = x)$ is a sentence.
(2) If ϕ is a sentence, then so is $\neg(\phi)$.
(3) If ϕ_1 and ϕ_2 are sentences, then so is $(\phi_1 \wedge \phi_2)$ and $(\phi_1 \vee \phi_2)$.
(4) Nothing else is a sentence.

---

[1] Many textbooks in probabilistic graphical models (e.g. Koller and Friedman's book) adopt a simpler language of random variables and joint instantiations. The language we adopt here is (perhaps unnecessarily) more expressive.

The symbols $\neg, \wedge, \vee$ are called logical connectives and stand, respectively, for negation, conjunction and disjunction. The language defined requires the use of parenthesis to avoid ambiguity. It is standard to simplify sentences by adopting a precedence order over connectives and discarding parentheses. The standard precedence order is $=$, $\neg$, $\wedge$, $\vee$. For example, $\neg X = a \wedge X = b \vee Y = c$ is understood as $((\neg(X = a) \wedge (X = b)) \vee (Y = c))$.

The meaning of complex sentences is based on the concepts of valuations and satisfaction. A *valuation* is a function ν that maps each variable X to a value $x \in \Omega_X$. A valuation describes a context or state of affairs of the domain. The meaning of an arbitrary sentence is given by the *satisfaction* relation $\models$ over pairs of valuation and sentence. It is standard to write $\nu \models \phi$ to indicate that the pair (ν, phi) is in the relation. We then say that ν satisfies ϕ (ϕ is satisfied by ν). We also write $\nu \not\models \phi$ to indicate that (ν, ϕ) is *not* in the relation. In this case, we say that ν does not satisfy ϕ (ϕ is not satisfied by ν).

Definition 9. The satisfaction relation is inductively defined by:

(1) $\nu \models X = x$ if and only if $\nu(X) = x$.
(2) $\nu \models \neg\phi$ if and only if $\nu \not\models \phi$.
(3) $\nu \models \phi \wedge \psi$ if and only if $\nu \models \phi$ and $\nu \models \psi$.
(4) $\nu \models \phi \vee \psi$ if and only if $\nu \models \phi$ or $\nu \models \psi$.

Note that $\nu \models \phi \vee \neg\phi$ for any valuation ν and sentence ϕ. A sentence $\phi \vee \neg\phi$ is called a *tautology*. Similarly, $\nu \models \phi \wedge \neg\phi$ for any valuation and sentence. A sentence $\phi \wedge \neg\phi$ is called a *contradiction*.

Example 6. *A domain describing the connection between smoking and cancer might adopt the following vocabulary:*

Variable	*Meaning*	*Domain*
S	*Smoker?*	*yes, no*
A	*Age*	$1, \ldots, 80, > 80$
G	*Gender*	*male, female*
H	*Family history of cancer*	*yes, no*
Y	*How long begin a smoker*	$0, 1, \ldots, 80, > 80$

Definition 10. A set of sentences $\phi_1, \ldots, \phi_n$ is *mutually exclusive* if there is no valuation that satisfies more than one model, that is, there is no ν such that $\nu \models \phi_i \wedge \phi_j$ for $i \neq j$. They are *exhaustive* if each valuation satisfies at least one model, that is, for each ν there is i such that $\nu \models \phi_i$.

Probabilistic models can be defined over the language defined.

Definition 11. Let $\mathcal{L}$ be a generalized propositional language over variables $X_1, \ldots, X_n$ and domains $\mathrm{dom}(X_1), \ldots, \mathrm{dom}(X_n)$. A (structured) probabilistic model is a pair $(\mathcal{L}, \mathbb{P})$ where $\mathbb{P}$ is a real-valued function on $\mathcal{L}$ such that

(A'1) $\mathbb{P}(\phi) \geq 0$.
(A'2) $\mathbb{P}(\phi \vee \neg\phi) = 1$.
(A'3) If ϕ and ψ are mutually exclusive then $\mathbb{P}(\phi \vee \psi) = \mathbb{P}(\phi) + \mathbb{P}(\psi)$.

It is customary to call the a value $\mathbb{P}(X = x)$ a marginal probability and a value $\mathbb{P}(Y_1 = Y_1 \wedge \cdots \wedge Y_k = y_k)$ a joint probability.

The correspondence with the (unstructured) set-based definition is achieved by the following translation.

Proposition 6. *Let Ω be the set of all valuations and $\mathcal{F}$ be the set of events*

$$\alpha_\phi \coloneqq \{\nu \in \Omega : \nu \models \phi\},$$

where ϕ is a sentence. Then $(\Omega, \mathcal{F})$ is a field of sets.

The result above can be shown by noticing that disjunction corresponds to union, conjunction to intersection and negation to complement. Thus, all properties from the standard (Kolmogorov's) definition of probabilities extends to structured models, as well any related concept (independence, conditional probability, distribution, etc). In particular we have:

Proposition 7. *For any sentence ϕ,*

$$\mathbb{P}(\phi) = \sum_{\nu \in \Omega : \nu \models \phi} \mathbb{P}(\{\nu\}).$$

Hence, a distribution over sentences $X_1 = x_1 \wedge \ldots \wedge X_n = x_n$ fully characterizes a probabilistic function. This is better notated using distributions of variables.

Definition 12. The *distribution* of a set of variables $Y_1, \ldots, Y_k$ is the real-valued function p such that

$$p(y_1, \ldots, y_k) = \mathbb{P}(Y_1 = y_1 \wedge \cdots \wedge Y_k = y_k),$$

where each $y_j \in \text{dom}(Y_j)$. We often write $p(Y_1, \ldots, Y_k)$ to denote such that distribution.

In particular, the distribution over all variables in the vocabulary satisfies

$$p(x_1, \ldots, x_n) = \mathbb{P}(\{\nu\})$$

where ν is the valuation such that $\nu(X_i) = x_i$. We can look at this result as a set of sufficient conditions to specify a probabilistic model:

Proposition 8. *For any function*

$$p : dom(X_1) \times \cdots \times dom(X_n) \to [0, 1]$$

there is a single probabilistic model $(\mathcal{L}, \mathbb{P})$ for which p is the distribution of $X_1, \ldots, X_n$. Moreover, every probability $\mathbb{P}(\phi)$ can be specified in terms of p.

The result above is a simple corollary of 7. In fact, it is possible to axiomatize the theory of probabilistic models in terms of the proposition above (and then derive the axioms (A'1)–(A'3)).

It is often convenient to define distribution over conditional probabilities:

Definition 13. The *conditional distribution* of a set of variables $Y_1, \ldots, Y_k$ given variables $Z_1, \ldots, Z_m$ is the real-valued function p such that

$$p(y_1, \ldots, y_k, z_1, \ldots, z_m) = \mathbb{P}(Y_1 = y_1 \wedge \cdots \wedge Y_k = y_k | Z_1 = z_1 \wedge \cdots \wedge Z_m = z_m),$$

where $y_i \in \text{dom}(Y_i)$ and $z_j \in \text{dom}(Z_j)$. We often write $p(Y_1, \ldots, Y_k | Z_1, \ldots, Z_m)$ to denote that distribution and $p(y_1, \ldots, y_k | z_1, \ldots, z_m)$ to denote one of its values.

We can re-state basic properties of probability functions in the language of distributions:

Proposition 9 (Chain Rule for variables)**.** *For any ordered set of variables $Y_1, \dots, Y_k$ it follows that*

$$p(Y_1, \dots, Y_k) = p(Y_1) \prod_{j=2}^{k} p(Y_j | Y_1, \dots, Y_{j-1}) \,,$$

where the equality and product of functions are taken element-wise.

The equation above can be interpreted as quantified over all combinations of values $y_1, \dots, y_k$ of the variables, but is often more convenient to interpret it as an algebra over functions.

Proposition 10 (Total Probability Rule for Variables)**.** *Let $Y_1, \dots, Y_k$ and $Z_1, \dots, Z_m$ be disjoint sets of variables. Then,*

$$p(Y_1, \dots, Y_k) = \sum_{Z_1, \dots, Z_m} p(Y_1, \dots, Y_k | Z_1, \dots, Z_m) p(Z_1, \dots, Z_m) \,.$$

Proposition 11 (Bayes' Rule for Variables)**.** *Let $Y_1, \dots, Y_k$ and $Z_1, \dots, Z_m$ be (ordered) sets of variables such that $p(Z_1, \dots, Z_m) > 0$. Then,*

$$p(Y_1, \dots, Y_k | Z_1, \dots, Z_m) = \frac{p(Z_1, \dots, Z_m | Y_1, \dots, Y_k) p(Y_1, \dots, Y_k)}{p(Z_1, \dots, Z_m)} \,.$$

Propositions 9, 10 and 11 can be generalized to condition all distributions on a set $W_1, \dots, W_\ell$ of variables. We can also define independence in terms of variables.

Definition 14. We say that variables $\mathbf{Y} = (Y_1, \dots, Y_k)$ and $\mathbf{Z} = (Z_1, \dots, Z_m)$ are conditionally independent given variables $\mathbf{W} = (W_1, \dots, W_\ell)$ if

$$p(\mathbf{Y}, \mathbf{Z} | \mathbf{W}) = p(\mathbf{Y} | \mathbf{W}) p(\mathbf{Z} | \mathbf{W}) \,.$$

Note that in the definition above, we have used boldface to indicate an ordered set of variables. One can shown that $\mathbf{Y}$ and $\mathbf{Z}$ are conditionally independent if either $p(\mathbf{Y} | \mathbf{Z}, \mathbf{W}) = p(\mathbf{Y} | \mathbf{W})$ or $\mathbb{P}(\mathbf{Z}) = 0$ somewhere.

Example 7. *Consider again Example 4, which involves five tosses of loaded coins. Denote by $C_1, \dots, C_5$ the variables of the problem with domains $dom(C_i) = \{heads, tails\}$. The joint probability factorizes as $p(C_1, \dots, C_5) = \prod_{i=1}^{5} p(C_i)$. The probability of any sentence ϕ can be obtained from these (marginal) distributions as $\mathbb{P}(\phi) = \sum_{\nu \models \phi} \prod_{i=1}^{5} (\nu(C_i))$.*

3. Exercises

Exercise 1. *Prove Proposition 1.*

Exercise 2. *Let $\mathbb{P}$ be a probability function (on some probability space). Show that for any event β with $\mathbb{P}(\beta) > 0$ the function $\mathbb{P}'(\alpha) = \mathbb{P}(\alpha | \beta)$ is also a probability function.*

Exercise 3. *Prove Proposition 3.*

Exercise 4. *Prove Proposition 4.*

Exercise 5. *Prove the extended Bayes' rule: For any events α, β and γ with positive probability, it follows that*

$$\mathbb{P}(\beta | \alpha \cap \gamma) = \frac{\mathbb{P}(\alpha | \beta \cap \gamma) \mathbb{P}(\beta | \gamma)}{\mathbb{P}(\alpha | \gamma)} \,.$$

Exercise 6. *Show that for any two events α and β with positive probability the following assertions are equivalent:*

(1) α *and* β *are independent.*
(2) $\mathbb{P}(\alpha|\beta) = \mathbb{P}(\beta)$.
(3) $\mathbb{P}(\beta|\alpha) = \mathbb{P}(\alpha)$.

Exercise 7. *Prove Proposition 6.*

4. Reading

- Chapter 1 of Pearl's book.
- Chapters 1.1–1.4, 2.1–2.6 and 3.1–3.5 of Darwiche's book.
- Chapters 1.1–1.2, 2.1.1-2.1.5.1 of Koller and Friedman's book.

5. Assignment

- Writing code is *optional* for this task; an example C code representing and querying a structured probabilistic model is provided (you might modify the code or write you own code from scratch if you prefer).
- Use the code (or write your own) to solve Exercises 3.1, 3.2 and 3.9 in Darwiche's book.
- Deliverables (via PACA):
  - any scripts, input files or supporting files used to solve the exercises
  - A *short* report summarizing the theory learned (include anything you deem relevant)
  - documentation describing how to use your code (if you wrote any)
  - the solutions of the exercises (practical and theoretical).

  An example of a report is provided. This is to be used only as a guide on the format (and not on the content).
- Deadline: next lecture

5.1. Bibliography.

- Adnan Darwiche, Modeling and Reasoning with Bayesian networks, Cambridge Press, 2009.
- Daphne Koller and Nir Friedman, Probabilistic Graphical Models: Principles and Techniques, MIT Press, 2009.
- Judea Pearl, Probabilistic Reasoning in Intelligent Systems, Morgan Kaufman, 1988.