Overview of the Course and some Basic Concepts

Huizhen Yu

janey.yu@cs.helsinki.fi
Dept. Computer Science, Univ. of Helsinki

Course 582636: Probabilistic Models, Spring 2010

Lecture Outline

Overview

In reality many tasks require us to reason and act under *uncertainty*:

- disentangle systematic variation from random variation
- test hypothetical theories about unknown physical processes
- make decisions that will bear consequences in the unforeseen future, based on fragmentary information

How do we access uncertainty, pool information together, and make coherent reasoning and decisions?

Probabilistic modeling is a systematic approach to address these problems. It has a wide range of application areas: e.g.,

- natural/humanity sciences, medicine, signal processing/analysis

and some more recent ones:

- bioinformatics
- natural language processing
- data mining

also artificial intelligence systems, which are traditionally logic-based:

- knowledge-based expert systems

Overview

Two key features of the probabilistic approach are:

- treating observed data as realizations of random variables
- using probability distributions to represent variability

Our focus will be on graphical models; they are

- models that impose on distributions "qualitative structures," which can facilitate the elicitation and interpretation of the models, as well as efficient computation with them

Our goal: study some basic principles and techniques in probabilistic graphical modeling

Course Syllabus

Syllabus:

- Introduction to graphical models ($\sim$ 2 weeks)
  - Markov models
  - Markov random fields
  - Simple Bayesian networks
- Introduction to Bayesian inference ($\sim$ 1 week)
- Further study of graphical models ($\sim 2\frac{1}{2}$ weeks)
  - Markov properties in undirected and directed graphs
  - Junction tree algorithms
  - Case studies

Materials:

- Lecture slides and chapters from several books

References

Some class materials will be from the chapters of:

Jensen, Finn V. and Nielsen, Thomas D. *Bayesian Networks and Decision Graphs*. Springer, 2007.

Cowell, Robert G., Dawid, Philip A., Lauritzen, Steffen L., and Spiegelhalter, David J. *Probabilistic Networks and Expert Systems: Exact Computational Methods for Bayesian Networks*. Springer, 2007.

Davison, A. C. *Statistical Models*. Cambridge Univ. Press, 2003.

Remark: These books are quite advanced for our purpose, and we will not delve deeply into them. But they are worth reading and also good for further self-study.

A lot of helpful materials can be found from the previous course website. Besides, the following book offers a friendly guide to its subject:

Jensen, Finn V. *An Introduction to Bayesian Networks*. UCL Press, 1996.

Additional recommended readings will be mentioned at each class.

Course Information and Requirements

Exercises:

- Problem sets will be given weekly. They will be due near the end of the following week, at the exercise group meeting. This means there will be about $1\frac{1}{2}$ weeks to work on a problem set, and some overlap between the start of a new problem set and the end of the previous one.
- You may hand in your answers at the exercise meetings or email them to me before the meetings.
- Time and location for meetings: temporarily, Fri 14-16, B222
  We decide a suitable meeting time in this class.

Grades: exercises 40%, final exam 60%

Office hours: Mon 13-14; or contact me for an appointment

A hands-on project-work course in the next period:

582637 Project in Probabilistic Models/Todennäköisyysmallien harjoitustyö

Other information can be found on the course webpage:

http://www.cs.helsinki.fi/group/cosco/Teaching/Probability/2010/

Outline

Probability as Defined Mathematically

Probability theory is founded on set and measure theories.

Probability space $(\Omega, \mathcal{F}, P)$

- Ω – sample space
  possible outcomes of an experiment
  outcomes are mutually exclusive and collectively exhaustive
- $\mathcal{F}$ – σ-field (σ-algebra)
  a collection of subsets of Ω, closed under set operations: complement, union, and intersection;
  important mathematically, but beyond our scope
- P – probability (measure)
  a function that assigns mass – a number in $[0, 1]$ – to every *event*, which is a subset of Ω (and also in $\mathcal{F}$);
  $P(A)$ – probability of event A: mass associated with the set A

Examples of Sample Space

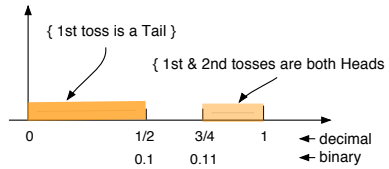
Simple examples

- Coin tossing, n times:

$$\Omega = \{a_1 a_2 \dots a_n \,|\, a_i \in \{H, T\}\};$$

  e.g., $n = 2$: $\Omega = \{HH, HT, TH, TT\}$
- Coin tossing, infinitely many times:
  H: 1, T: 0; view "$a_1 a_2 \dots$" as a binary point: $0.a_1 a_2 \dots$

$\Omega = [0, 1]$



- Tossing two coins infinitely many times: $\Omega = [0, 1]^2$

The abstract probability framework can handle almost arbitrarily complex outcomes.

Axioms of Probability

Axioms

- Nonnegativity: $P(A) \geq 0$ for all events A.
- Normalization: $P(\Omega) = 1$.
- Additivity:
  For any disjoint events A, B,

$$P(A \cup B) = P(A) + P(B).$$

  Countable additivity: for any countable collection of pairwise disjoint events A_i, $i = 1, 2, \dots$,

$$P\Big(\bigcup_{i \geq 1} A_i\Big) = \sum_{i \geq 1} P(A_i)$$

Additivity ensures consistency:
There are many ways to count the total mass $P(A)$ associated with the set A; at the end they all lead to the same number.

Probability Calculus

Conditional probability of A given B (definition):

$$P(A \,|\, B) = \frac{P(A \cap B)}{P(B)}, \quad \forall B \text{ with } P(B) > 0$$

(undefined if $P(B) = 0$)

Some properties of $P(A \,|\, B)$:

- Asymmetry: $P(A \cap B) = P(B \cap A)$, but generally,

$$P(A \,|\, B) \neq P(B \,|\, A).$$

- Meaning: given that the outcome is in B, how likely the outcome is also in A. E.g., if the number of outcomes is finite and all of them are equally likely,

$$P(A \,|\, B) = |A \cap B| / |B|.$$

- $P(\cdot \,|\, B)$ is also a probability on Ω (it concentrates on the "new universe" B), and it defines the probability of any event A given B.
- Conditional probabilities are derived from P and hence determined by P. Using the calculus, information is updated *consistently* with P.

Bayes' Theorem

Bayes' theorem

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}, \qquad \forall A, B, \text{ with } P(B) > 0.$$

- Follows from the definition of conditional probabilities:

$$P(A \cap B) = P(B|A)P(A) = P(A|B)P(B)$$

- $P(A)$: prior probability of A before observing evidence B
- $P(A|B)$: posterior probability of A after observing evidence B

Why is this formula useful?

- The direction $A \to B$ can be more intuitive, and related probabilities are more readily available: e.g.,
  $A =$ 'having disease x' and $B =$ 'showing symptom y;'
  $A =$ 'word x is spoken' and $B =$ 'sound signal y is detected.'
- Reasoning in the reverse direction, about A given B, can be complicated and counterintuitive, even if the relation $A \to B$ is simple. But the reverse reasoning is often what is needed in applications.

Outline

Interpretations of Probability

The mathematical framework of probability is clean and general.
But how do we interpret the meanings of probabilities in real world?

This is a serious issue because how we interpret probability relates closely to how we specify the numbers (probabilities) in the first place, reason about and react to the numbers obtained from calculation.

Intense interests in and debates on this subject have continued till today.
(Two related articles are given as recommended readings.)

We take a glimpse at this issue in the next few slides. First, does it make sense to say $P(A) = 0.7$ or 0.2 for the following A?

- $A = \{$A patient recovers from H1N1 flu$\}$
- $A = \{$It will snow tomorrow$\}$
- $A = \{$It snowed this day last year$\}$
- $A = \{$There is life beyond earth$\}$
- $A = \{$The Suez canal is longer than the Panama canal$\}$

Interpretations of Probability

Two commonly used, different interpretations of probability:

- frequency-based, objective:
  frequencies from repetitions of experiments (realizable or hypothetical)
- logic of partial belief, subjective:
  Here A is a proposition and $P(A)$ the degree of belief in A being true.
  - We may say "I believe to the extent of $P(A)$ that A is true," but not "I believe A is true to the extent of $P(A)$."
  - Contrast with the objective interpretation: there, $P(A)$ is the proportion of times that A occurs to be true.

Notes:

- In practice both interpretations may be involved in the same application, which complicates the overall picture even more.
- "Uncertainty" is a natural concept to use to interpret probability. However, it faces the same interpretation issue when uncertainty needs to be quantified.

In the preceding slide, which interpretation might be applied to $P(A)$?

Logic vs. Probability as Partial Belief

Semantic differences between probability and "if-then" logical statements:

- It is natural to do bidirectional inferences with $P(A|B)$ and $P(B|A)$:

  causes $\leftrightarrow$ effects, evidence $\leftrightarrow$ explanations.

- $P(A|B)$ is our uncertainty about A given we know B but nothing else. If we also know C, our uncertainty about A changes to $P(A|B \cap C)$. Correlation between the sources of evidences is accounted for.

With logical statements, it is not easy to handle

- Bidirectional inferences:
  - "If A then B with certainty x" does not entail that B being true makes A more credible. Example: "Fire implies smoke."
  - Suppose "If A then B with certainty x" and "If C then B with certainty y." Finding B and A being true does not make C less credible – does not *explain away* the cause C.
- Correlated evidences:
  - From "If B then A with certainty x" and "If C then A with certainty y," how do we deduce the certainty of A when both B and C are true?
- Exceptions: such as "Birds fly," "Penguins can't fly."

More about Interpretations of Probability

About subjective probability/partial belief:

- The purely subjective case is self-coherent (follows probability calculus), but unscientific: anything goes.
  Okay for personal use, but not much can be said beyond that.
- For applications with public impact, impersonal subjective probabilities need to be carefully sought.

About probabilistic modeling:

People differ in opinions on how uncertainty should be assessed. Choosing a model already involves subjectivity. The model can have no relation to the data generating process – how do we then interpret the probabilities even if they are frequencies-based? These are among the difficult conceptual issues.

About relation between two kinds of consistency:

- self-consistency in the way one reasons: rationality
- consistency with facts (not implied by the first)

Implications to us:

(i) A probabilistic approach does not guarantee our being correct. *Self-critisicm* is always important: checking whether assumptions and their implications are sound based on data.

(ii) A self-consistent reasoning method may not be "trouble-free" when confronting the truth. (This can be conceptually more difficult to understand when studying various statistical methods.)

Outline

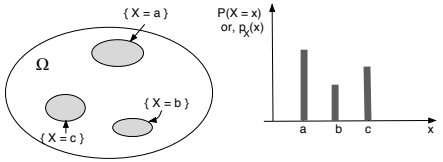
Random Variables

Random variables are functions on the sample space Ω.

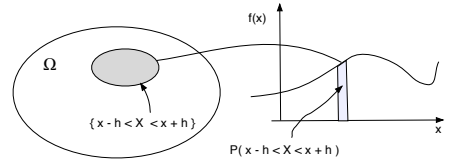
- They can take arbitrary values. They will be denoted by capital letters, and their values lower-case letters.

Examples

- Discrete random variables



- Continuous random variables



- Functions of random variables

  e.g., the indicator function $I(X \in A)$, $f(X)$, $E[f(X,Y)|Y]$

Remark: In practice we usually work with random variables and their distributions directly, and we do not specify explicitly Ω, which we may not know actually, (for instance, in scientific studies). It is also usual to construct Ω for the variables of interest and expand Ω if we want to take more variables into consideration.

Some Concepts Associated with Random Variables

Let X, Y be discrete random variables on $(\Omega, \mathcal{F}, P)$, with $x \in S_X, y \in S_Y$.

- Joint distribution P_{XY}: probability on $S_X \times S_Y$

$$P_{XY}\big((X,Y) \in A\big) = P\big((X,Y) \in A\big), \quad A \subset S_X \times S_Y.$$

  Probability mass function (PMF) $p_{XY} : S_X \times S_Y \to [0,1]$

$$p_{XY}(x,y) = P(X = x, Y = y), \quad x \in S_X, \ y \in S_Y.$$

- Marginal distribution P_X and its PMF p_X:

$$P_X(X \in A) = P(X \in A), \quad A \subset S_X; \quad p_X(x) = P(X = x), \quad x \in S_X.$$

- Relation between the joint and marginal

$$p_X(x) = \sum_{y \in S_Y} p_{XY}(x,y), \quad x \in S_X.$$

- Expectation: $E[f(X)]$, where f is a real-valued function on S_X.

Notes:

- For continuous random variables with P_{XY}, P_X having densities, we use density functions in place of PMFs – both are referred to generally as densities in some books.
- For general random variables, we either define densities in a general sense or work with distributions. These are beyond our scope; we will mostly focus on discrete random variables.

Some Concepts Associated with Random Variables

- Conditional probabilities of X given Y (defined as those for events):

$$P(X = x \,|\, Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)}, \quad \text{if } P(Y = y) > 0.$$

  Conditional distributions $P_{X|Y}$ and conditional PMFs $p_{X|Y}$:

$$P_{X|Y}(X \in A \,|\, Y = y) = P(X \in A \,|\, Y = y), \quad A \subset S_X, \ y \in S_Y;$$

$$p_{X|Y}(x|y) = P(X = x \,|\, Y = y) = P_{X|Y}(X = x \,|\, Y = y), \quad x \in S_X, \ y \in S_Y.$$

- Relation between $p_{X|Y}$, p_{XY} and p_X, p_Y:

$$p_{XY}(x,y) = p_Y(y) p_{X|Y}(x|y), \quad \forall x, y. \qquad \text{(by def. of conditional probability)}$$

$$p_X(x) = \sum_{y \in S_Y} p_Y(y) p_{X|Y}(x|y), \quad \forall x. \qquad \text{(marginalization)}$$

  (Define $0 \cdot \textit{undefined value} = 0$ in the above.)

- Conditional expectation: $E[f(X) \,|\, Y = y]$

Notation:

- We will often identify Ω with the space of all possible values of random variables of interest, so P will simply be the joint distribution.
- We will often omit the subscripts X, Y of p or P for simplicity, when the context is clear from the arguments of p or P.

Outline

Independence/Conditional Independence of Events

Let A, B, C be events.

- A and B are *independent* if

$$P(A \cap B) = P(A)P(B), \quad \text{equivalently,} \quad P(A \,|\, B) = P(A), \quad \text{if } P(B) > 0.$$

  This implies that A^c and B^c are independent:

$$\begin{aligned} P\big(A^c \cap B^c\big) &= P\big((A \cup B)^c\big) = 1 - P(A \cup B) && \text{(axioms)} \\ &= 1 - \Big(P(A) + P(B) - P(A \cap B)\Big) && \text{(axioms)} \\ &= 1 - P(A) - P(B) + P(A)P(B) && \text{(independence)} \\ &= \big(1 - P(A)\big)\big(1 - P(B)\big) = P(A^c)P(B^c), \end{aligned}$$

  and also that A and B^c are independent.

- Meaning: knowing B or B^c (*and nothing else*) does not change our uncertainty about A.

- A and B are *conditionally independent* given C, where $P(C) > 0$, if

$$P(A \cap B \,|\, C) = P(A \,|\, C)P(B \,|\, C), \quad \text{equivalently,} \quad P(A \,|\, B \cap C) = P(A \,|\, C).$$

- Meaning: if we already know C, knowing B or B^c (*and nothing else*) does not change our uncertainty about A.

Independence of Random Variables

Independence of random variables

- X and Y are (marginally) independent, written as $X \perp Y$, if

  $$P(X = x, Y = y) = P(X = x)P(Y = y), \quad \text{or } p(x, y) = p(x)p(y), \ \forall x, y.$$

- Equivalent definition, but in an asymmetric form:

  $$P(X = x \,|\, Y = y) = P(X = x), \quad \text{or } p(x|y) = p(x), \quad \forall\, x, y \text{ with } p(y) > 0.$$

- Meaning: knowing Y (and nothing else) does not change our uncertainty about X (vice versa).

- X_i, $i = 1, \ldots, n$, are *mutually* independent, if

  $$P(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^{n} P(X_i = x_i), \qquad \forall x_i, \ i = 1, \ldots, n.$$

- Pairwise independence does not imply mutual independence.
  Parity-checking example: $X_i, i = 1, \ldots, n-1$ are mutually independent with $x_i \in \{0, 1\}$; $X_n = \left(\sum_{i=1}^{n-1} X_i\right) \bmod 2$. Then any $n-1$ variables of X_is are mutually independent, but together X_is are not independent.

Conditional Independence of Random Variables

Conditional independence of random variables

- X and Y are conditionally independent given Z, written as $X \perp Y \,|\, Z$, if $\forall x, y, z$ with $P(Z = z) > 0$,

  $$P(X = x, Y = y \,|\, Z = z) = P(X = x \,|\, Z = z)\, P(Y = y \,|\, Z = z).$$

- Equivalent definition in an asymmetric form: $\forall x, y, z$ with $p(y, z) > 0$,

  $$P(X = x \,|\, Y = y, Z = z) = P(X = x \,|\, Z = z), \quad \text{or } p(x|y, z) = p(x|z).$$

- Meaning: If we already know Z, knowing Y (and nothing else) does not change our uncertainty about X (vice versa).

Notes:

- $X \perp Y \,|\, C$ for an event C is analogously defined; it is weaker than conditional independence between variables.
- Independence/conditional independence has more to do with the *form*/"structure" than the numerical values of distributions.
- In practice, independence/conditional independence are often related to the natural concept of "irrelevance;" and being able to identify irrelevant factors helps efficient decision making.

Some Conventions to Simplify Expressions

When expressing $X \perp Y \,|\, Z$ in terms of P or p, it is bothersome to write the positivity conditions on the conditioning variables, such as "$P(Z = z) > 0$."

We will adopt some conventions to omit writing these conditions:

- For *discrete* random variables, when we write

  $$P(X = x, Y = y \,|\, Z = z) = P(X = x \,|\, Z = z)\, P(Y = y \,|\, Z = z), \quad \forall x, y, z;$$
  $$P(X = x \,|\, Y = y, Z = z) = P(X = x \,|\, Z = z), \quad \forall x, y, z;$$
  $$P(X \,|\, Y, Z) = P(X \,|\, Z);$$

  or

  $$p(x, y \,|\, z) = p(x \,|\, z)\, p(y \,|\, z), \quad \forall x, y, z;$$
  $$p(x|y, z) = p(x|z), \quad \forall x, y, z,$$

  we mean that equality holds whenever all the quantities involved are well-defined under P.

Importance of Dependence/Independence Relations

Ignoring dependences can mislead us

- An example of *Simpson's paradox*:

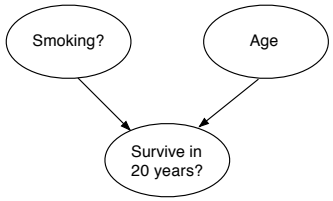
  20-year survival and smoking status for 1314 women; numbers refer to dead/total (% dead)

Smokers	Non-smokers
139/582 (24)	230/732 (31)

  Conclusion: smoking is better for health (?)

Full data on smoking and survival:

Age	Smokers	Non-smokers
overall	139/582 (24)	230/732 (31)
18-24	2/55 (4)	1/62 (2)
25-34	3/124 (2)	5/157 (3)
35-44	14/109 (13)	7/121 (6)
45-54	17/130 (21)	12/78 (15)
55-64	51/115 (44)	40/121 (33)
65-74	29/36 (81)	101/129 (78)
75+	13/13 (100)	64/64 (100)



We interpret an edge '$X \to Y$' loosely as "X has direct influence on Y" for the time being.

- Another example:
  association between a child's reading ability and shoe size

Importance of Dependence/Independence Relations

The naive structureless probabilistic approach:

a look-up table for $P(X_1, X_2, \ldots, X_n)$ + probability calculus for inference

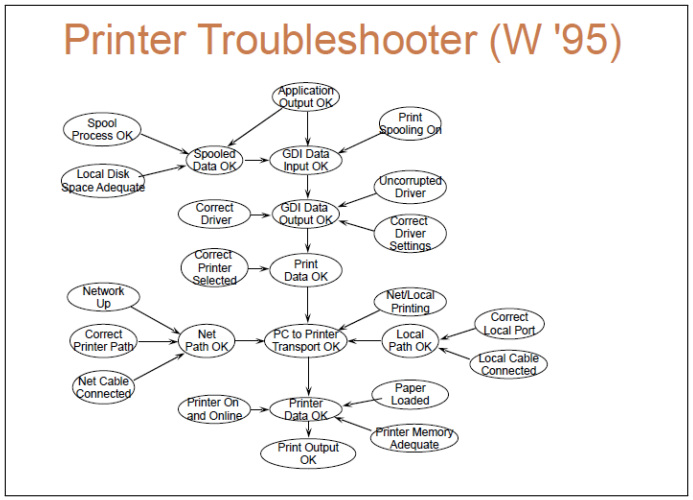
- Intractable computationally even for small n
  For instance, with $n = 100$ and $x_i \in \{0, 1\}$, the table size is 2^{100}. To calculate $p(x_1, \ldots, x_{80} \,|\, x_{81}, \ldots, x_{100})$, we need to add up 2^{80} numbers.
- Difficult to understand and interpret
  dependence structures are buried in a table of numbers
- Difficult to specify P in the first place
  for either experts, or automatic data-analysis programs (would require an enormous amount of data)

Key dependence relations, such as cause-effect, direct interaction, are often

- what we are interested to discover
- qualitative building blocks of a modular model
  easy to understand and manipulate computationally

Snapshots of Real Application Systems

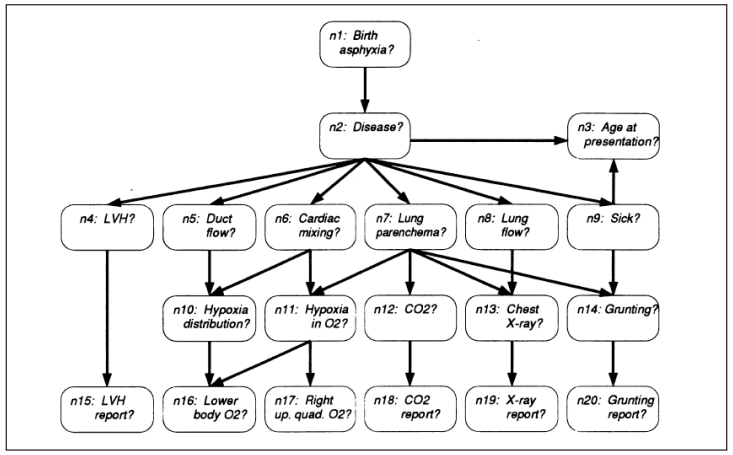
A trouble-shooting system based on Bayesian network (network structure can be specified before any quantitative modeling):



Snapshots of Real Application Systems

CHILD: a medical diagnosis system based on Bayesian network
(Spiegelhalter et al. Statistical Science, 8(3):219-283, 1993)

Network structure is elicited from experts before any quantitative modeling.



What is a Graphical Model?

A model, often denoted by Θ, is a collection of probability distributions for random variables of interest. It may not contain the true distribution of the random variables.

- $\theta \in \Theta$ refers to the parameters of a distribution or the distribution itself.
- The distributions in the model usually share some common features.
- Some people refer to an individual distribution $\theta \in \Theta$ as a model.

In a graphical model:

- There is an associated graph whose vertices correspond to random variables and edges represent "direct interactions" among variables.
- This graph is more than a pictorial representation: it "encodes" conditional independence relations. (The encoding is by convention different for different types of graphs: directed, undirected, and mixed, etc.)
- All distributions in the model obey the independence relations specified by the graph.

With graphical models, modeling can be divided into two stages:

- qualitative modeling stage – specifying the graph
- quantitative modeling stage – specifying numerical attributes of the model

Recommended Further Readings

On meanings of probability:

1. Alan Hájek. *Interpretations of probability*, 2002.
2. Frank P. Ramsey. *Truth and Probability*, 1926.

You can find the articles online from the links on the course webpage.

For comparison between probability and logic:

3. Judea Pearl. *Probabilistic Reasoning in Intelligent Systems*, Morgan Kaufmann, 1988. Chap. 1.

For a brief overview of probabilistic expert systems:

4. Robert G. Cowell et al. *Probabilistic Networks and Expert Systems*, Springer, 2007. Chap. 2.