Introduction to Graphical Models

lecture 1 - introduction

Marc Toussaint Ulf Brefeld
TU Berlin

overview:

– why Graphical Models and Bayesian methods?

– overview of the course

– some basics: random variables & probabilties

Why Graphical Models and Bayesian methods? – 1

- formalization of *Information Processing*
  – data is information
  – sensors give information
  – outputs/actions/decisions are *missing* information (to be 'inferred')
  – coupling between sources/points of information

$\Rightarrow$ Graphical Models formalize "networks of coupled information"

$\Rightarrow$ Information Processing can be viewed as inference or message passing in Graphical Models

Why Graphical Models and Bayesian methods? – 2

- neuroscientific motivation:
- Neural *Information Processing* Systems (NIPS)

  Bayesian Brain: Probabilistic Approaches to Neural Coding K. Doya, S. Ishii, A. Pouget, RPN. Rao (editors), MIT Press (2007)

  The Neurodynamics of Belief Propagation on Binary Markov Random Fields T. Ott, R. Stoop (NIPS 2006)

- Bayesian information processing is a possible *abstraction* of neuronal functions

  (not a model of *how* neurons work, but what their function from an information processing point of view is.)

Why Graphical Models and Bayesian methods? – 3

- many concrete algorithms can be derived/explained in terms of graphical models:
  – in speech & text processing (HMMs, CRFs, ..)
  – in computer vision (MRFs, sensor fusion, ..)
  – clustering (Dirichlet processes, LDA)
  – regression (GPs)
  – reinforcement learning (3rd part of lecture)
  – robotics (AICO)
  – dimension reduction (GPLVM, GTM)

- Graphical Models/Bayesian methods help to explain/unterstand many things in one coherent framework
- generic methodology to derive specialized algorithms in your own domain

plan

21.04.	Introduction
28.04.	Diskrete Wahrscheinlichkeitsverteilungen, Bedingte und Gesamtwahrscheinlichkeiten, Graphische Modelle, Faktorgraphen
05.05.	Inferenz, Eliminierungsalgorithmus, Evidenzen
12.05.	Summen-Produkt-Algorithmus, Junction Tree Alg. (JTA)
19.05.	Beliefpropagierung in zyklischen Graphen (Loopy BP), Mean-Field Alg.
26.05.	Hidden Markov Models (HMMs), Forward-Backward Alg., Viterbi, Expectation-Maximization (EM)
02.06.	Bedingte Wahrscheinlichkeiten, (Kernel-) Conditional Random Fields (k-CRFs), Features, Optimierung
09.06.	Optimierung (Fortsetzung), Strukturiertes Perzeptron
16.06.	Strukturierte Support-Vektor-Maschinen (SSVMs)
23.06.	Influence Diagramme
30.06.	Markov Decision Processes (MDPs)
07.07.	Inferenz zur Planung, Optimale Handlungsstrategien (Policies)
14.07.	Zusammenfassung und Fragestunde

Questions? ...

probability theory

- why do we need probabilities?
  – of course, in case of random events, stochasticity...

- but also in a deterministic world!:
  – lack of knowledge!
  – hidden (latent) variables
  – expressing *uncertainty*
  – expressing *information*

- probabilities are a generic tool to express uncertainty, information, and coupling

random variables

(for simplicity, this course mainly consideres *discrete* random variables)

- intuitively: a *random variable* takes on *values* with a certain probability
  Example: a dice can have values $\{1, .., 6\}$
  a bit more formally: a random variable introduces a probability measure on the domain (sample space) ("assigns a probability to each possible value")

- the *domain* dom(X) of a variable X is the set possible values of a random varible (mutually exclusive and collectively exhaustive)

- we use capital letters X to denote random variables and lower case letters x to denote values that they take

- we use the P to denote the mapping to probabilites

  $$P(X = x) \in \mathbb{R}$$

random variables (in terms of sets)

Let X be a random variable with domain $\Omega = \text{dom}(X)$
Let $A, B \subset \Omega$ be subsets of the domain and $x \in \Omega$ a value in the domain.

- $X \in A$ or $X \in B$ or $X = x$ are called *events*

- we use the P to denote the mapping assigning events to real numbers:
  – $P(X \in A) \in \mathbb{R}$

- we require
  – $P(X \in \emptyset) = 0$
  – $P(X \in \Omega) = 1$
  – if $A \cap B = \emptyset$ then $P(X \in A \cup B) = P(X \in A) + P(X \in B)$

  if the domain is discrete this implies
  – $\sum_{x \in \Omega} P(X = x) = 1$

probabilty distribution & tables

- for continuous domains: “probability distribution” is the integral of a “probability density function”
- for discrete domains: “probability distribution” and “probability mass function” are used synonymously

- a RV assigns a probability to each possible value
  $\rightarrow$ think of the probability distribution as a *table* of numbers:
  Example: A fair dice X, $\text{dom}(X) = \{1, 2, 3, 4, 5, 6\}$, with

$$\forall_{x \in \text{dom}(X)} : \; P(X = x) = \frac{1}{6}$$

  corresponds to the table

$$[\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}]$$

- in implementations we typically represent random variables by tables (arrays/vectors) of numbers

Probability: Frequentist and Bayesian

- Frequentist probabilities are defined in the limit of an infinite number of trials
- *Example:* The probability of a particular coin landing heads up is 0.43

- Bayesian (subjective) probabilities quantify degrees of belief
- *Example:* The probability of it raining tomorrow is 0.3
- Not possible to repeat tomorrow many times

joint distributions

- assume we have two random variable X and Y. The *joint probability distribution*

  $$P(X = x, Y = y)$$

  gives the probability that $X = x$ *and* $Y = y$.
  (In logic one would perhaps write something like $X = x \wedge Y = y$. But not so in joint probability distributions.)

- *Example:* Suppose Toothache and Cavity are the variables:

	Toothache = true	Toothache = false
Cavity = true	0.04	0.06
Cavity = false	0.01	0.89

  we write

  $$P(Toothache = true, Cavity = false) = 0.01$$

joint distributions

- note, the whole lecture will be about JOINT PROBABILITY DISTRIBUTIONS
  – graphical models are nothing but descriptions of joint probability distributions!
  – correlations, interdependence, coupling are all expressed in terms of joint probability distributions
  – whenever you're confused about the "model", the "approach", the "assumptions", etc, reconsider explicitly what the joint probability distribution over all involved variables is!

joint distributions

- *definitions:*
  – the *marginal* (probability) of X given $P(X, Y)$ is

$$P(X) = \sum_{Y} P(X, Y)$$

  – the *conditional* (probability) of X given Y and $P(X, Y)$ is

$$P(X|Y) = \frac{P(X, Y)}{P(Y)}$$

  defs also hold for tuples of variables, e.g., $X = (X_1, .., X_n)$, $Y = (Y_1, .., Y_m)$

- *implications:*
  – the *product rule* $P(X, Y) = P(X|Y)\ P(Y) = P(Y|X)\ P(X)$
  – the *chain rule* $P(X_1, .., X_n) = \prod_{i=1}^{n} P(X_i|X_1, .., X_{i\text{-}1})$
  – *Bayes Rule*

$$P(X|Y) = \frac{P(Y|X)}{P(Y)} P(X)$$

Bayes Rule

- Thomas Bayes (1702–1761)
- Bayes Rule is a trivial implication of the definitions of marginal and conditional probability!

- importance lies in its interpretation and use:

$$P(X|Y) = \frac{P(Y|X)}{P(Y)}\, P(X)\,, \quad \text{posterior} = \frac{\text{likelihood}}{\text{evidence}}\, \text{prior}$$

$$P(\text{cause}|\text{effect}) = \frac{P(\text{effect}|\text{cause})}{P(\text{effect})}\, P(\text{cause})$$

Example: let M be meningitis, S be stiff neck

$$P(M|S) = \frac{P(S|M)}{P(S)}\, P(M) = \frac{0.8}{0.1}\, 0.0001 = 0.0008$$

Note: posterior probability of meningitis still very small

Shows *importance of the prior*

inference

- we will deal with many variables $X = (H_1, ..H_n,\ E_1, .., E_m,\ Y_1, .., Y_k)$
  – we are given the joint probability distribution
  $P(H_1, ..H_n,\ E_1, .., E_m,\ Y_1, .., Y_k)$
  – some variables $E_1, .., E_m$ are observed (we have evidence)
  for the other variables $H_1, ..H_n\ , Y_1, .., Y_k$ we have no evidence
  we want to know the *posterior* over some variables $Y_1, .., Y_K$

$$P(Y_{1:k} \mid E_{1:m}) = \frac{P(Y_{1:k}, E_{1:m})}{P(E_{1:m})} \propto \sum_{H_{1:n}} P(Y_{1:k}, E_{1:m}, H_{1:n}) \tag{1}$$

- this is the problem of inference

- obvious problem: size of table $P(Y_{1:k}, E_{1:m}, H_{1:n})$ is d^{k+m+n}

summary

- focus of this lecture:
  – graphical models as a generic tool for inference with coupled random variables
  – probability theory as calculus for uncertainty, information, evidence
  – learning graphical models from data
  – using graphical models for decision making & RL

- next time:
  – naive Bayes
  – graphical models
  – inference using the elimination algorithm

- web links:

  Bayes Rule:

  `http://www.cs.ubc.ca/~murphyk/Bayes/bayesrule.html`

  Kevin's lecture:

  `http://www.cs.ubc.ca/~murphyk/Teaching/CS532c_Fall04/Lectures/index.html`

  `http://www.cs.ubc.ca/~murphyk/Bayes/bnsoft.html`

  google `site:http://www.cs.ubc.ca/~murphyk/Bayes`