Introduction to Graphical Models
lecture 12 - summary & open problems

Marc Toussaint
TU Berlin

Information

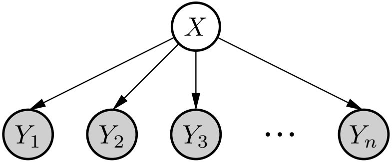
- inference in graphical models is about information processing...
- what is information?
  – Shannon Entropy

$$H(X) = -\sum_X P(X) \log P(X)$$

  – information $\leftrightarrow$ neg entropy, (non-uniform) probability distribution
- we use probability distribution as an *information calculus*
  (Bayesian vs. frequentist (description of repeatable experiments) view on probabilities)
  David MacKay: *Information Theory, Inference, and Learning Algorithms*, Cambridge University Press, 2003
- Graphical models
  = joint probability distribution over multiple variables
  $\rightarrow$ allow information processing between multiple variables

Core operations on information

1. **summing/marginalizing**
   – marginalizes a joint distribution $P(X) = \sum_Y P(X, Y)$
   – "eliminate Y" "subsume information on Y" "resolve coupling to Y"
2. **product**
   – fusing (independent) information
   – Bayes rule $P(X|Y) \propto P(Y|X) P(X)$, posterior $\propto$ likelihood $\cdot$ prior
   – Naive Bayes



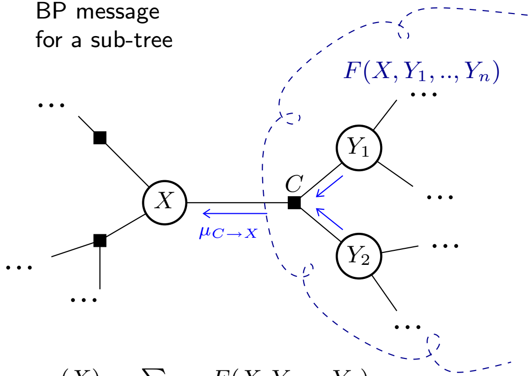
$$P(X|Y_{1:n}) \propto P(X) \prod_{i=1}^n \mu_{Y_i \to X}(X) \quad \text{with} \quad \mu_{Y_i \to X}(X) := P(Y_i = y_i \,|\, X)$$

   – message propagation: $\quad b_i(X_i) := \prod_{C \in \partial i} \mu_{C \to i}(X_i)$

Message propagation

- on tree structures: (see also Bishop: *Pattern Regocnition*)

BP message
for a sub-tree



$$\mu_{C\to X}(X) := \sum_{Y_{1:n}} F(X, Y_1, .., Y_n)$$
$$= \sum_{Y_1,Y_2} C(X, Y_1, Y_2)\ \mu_{Y_1\to C}\ \mu_{Y_2\to C}$$

messages subsume the information from a whole sub-tree

such that (as in Naive Bayes) the belief is the *product* of independent informations:

$$b_i(X_i) = \prod_{C\in\partial i} \mu_{C\to i}(X_i)$$

Message propagation

- BP can also be implemented on loopy graphs:

  1) we can't resolve recursion of msg. eqns $\rightarrow$ update eqns

  2) marginal consistency is a fixed poing of BP update eqns

$$\sum_{X_C \setminus X_i} b(X_C) = \sum_{X_D \setminus X_i} b(X_D) = b(X_i)$$

  3) problem: we multiply/fuse *dependent* information

  4) may diverge

  5) ongoing theory: Bethe approx., loop correction, generalized BP, etc

Learning & inference

LEARNING a model	likelihood maximization structured output Expectation Maximization
USING a model	inference information processing planning

Learning

- Maximum Likelihood:
  learn parameters θ of $P(X;\theta)$ such that complete data log-likelihood for data $D = \{x_i\}_{i=1}^n$ is maximal:

$$L(\theta) = \sum_{i=1}^{n} \log P(x_i\ ;\theta)$$

- structured output (Ulf Brefeld): given "external" inputs x
  – learn a mapping $x \mapsto P(y|x;\boldsymbol{w})$ from x to a distribution over outputs y
  – learn a "conditional" distribution, typically in the form

$$P(y|x;\boldsymbol{w}) \propto \exp\{\langle \boldsymbol{w}, \boldsymbol{\Phi}(x,y)\rangle\}$$

  – $\boldsymbol{w}$ parameterizes how the distribution over y depends on the input x

- Expectation Maximization: learning $P(X,Y)$ without observing X...

Summary

- we addressed the core of
  - information processing, in a literal sense, in terms of probabilistic inference, messages, multiplying, marginalizing, etc
  - learning, in the sense of learning how information/RVs are coupled (also to input) $\leftrightarrow$ learning parameters of joint (or conditional) distributions

- so, isn't that all we need for AI? Why not?
  – computational limits
  – representations...

Representations I

- Have you noticed:
  In every example so far we started with saying
  "Let there be n RVs $X_{1:n}$ with domain ... "
  – Let there be binary RVs "Toothacke, Cavity"
  – Let there be binary RVs "Battery, Gauge, Fuel, TurnOver, Start"
  – Let there be binary RVs D, X, E, B, L, T, S, A (Asia network)
  – Let there be binary RVs "Rain, Sprinkler, Holmes, Watson"

- *We always assume to know what are the relevant quantities (RVs) for which to represent information – also for the latent/unobservable information!*

Representations II

- Could we not have a system that invents its own internal variables?

  Develops own internal representations which allow it to concisely model the data?

  Don't humans invent/develop new concepts/categories/quantities exactly for that purpose?

Representations III

- These are very hard and open problems:
  – a related research field is called “structure learning”
    - easier part: given we know which RVs exist, learn which are coupled
    - medium part: we know there is a certain semantic RV, but don’t know how many values it can have (dom(X) unknown) (some buzzwords: Dirichlet allocation, Chinese Restaurant Process, infinite HMMs, etc)
    - harder part: we don’t know which RVs might even exist, are latent in the data, or which should be introduced to model the data

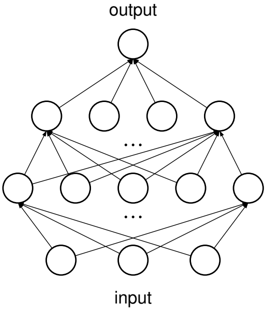
- Example: Imagine an artifical system watching tons of movies
  – it is tabula rasa, doesn’t know what exists, only sees video pixels
  – perhaps its intrinsic goal is to model (=“understand”?) what it sees
  – will/should it develop a RV for cows??

Representations IV

- Graphical Models:
  - one RV $\leftrightarrow$ one semantic quantity
  - usually explicitly defined by a human as part of the model definition
  - in many applications is is perfectly ok!
  - but very hard to address the above mentioned questions..

Other kinds of "networks"

- Neural Networks



activation of neurons $\sim$ representation of information

- More closely related to Graphical models:
  - Helmholtz machine
  - Boltzmann machine, restricted Boltzmann machine (RBM)
  - layers of RBMs (Hinton's deep networks)
  - auto-encoders
  - new ideas needed!

Conclusions

- Graphical models give a concise framework for
  – information processing, in terms of probabilistic inference, message propagation, etc
  – learning from data, in terms of learning how variables are coupled in a joint probability distribution

- current research:
  – on the one hand, graphical models become more and more a standard tool in applications and engineering
  – on the other hand, research in Machine Learning also seeks for alternative approaches to learn and develop representations

Bengio, Yoshua and LeCun, Yann: *Scaling learning algorithms towards AI*
Rodney Douglas et al.: *Future Challenges for the Science and Engineering of Learning*
Thomas G. Dietterich et al.: *Structured machine learning: the next ten years*