582636 Probabilistic Models

Spring 2013

Petri Myllymäki
Department of Computer Science
University of Helsinki, Finland

http://www.cs.helsinki.fi/courses/582636/2013/k/k/1
Overview

• In reality many tasks require us to reason and act under uncertainty.

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

• Probabilistic modeling is a systematic approach to address these problems.

• Our focus will be on graphical models that impose "qualitative structures" on distributions, facilitating the elicitation and interpretation of the models, as well as efficient computation with them.

• Our goal: to study some basic principles and techniques in probabilistic graphical modeling.
Probabilistic *Graphical* Models

- Nothing "graphical" per se
- An elegant theoretical (and practical!) framework for probabilistic modeling
- Independence assumptions can be visualized as directed or undirected graphs
- The graph-theoretic framework has led to several computationally efficient algorithms
- Successfully used in several real-world applications
Example I: Spam filtering

- Most spam filters based on the "Naive Bayes classifier"
Example II: Speech Recognition

- State-of-the-art: Hidden Markov models
- Google's Voice Search app [...] predicts what you're probably saying
- Apple's Siri: "Something like a Hidden Markov Model".
Example III: Robotics

- E.g. Autonomous cars
Example IV: "Mind Reading"

- See e.g. The Gallant Lab at UC Berkeley
  
  "...a Bayesian decoder by combining estimated encoding models with a sampled natural movie prior."
Example V: NLP

- Typical (probabilistic graphical) models:
  - Latent Semantic Analysis (LSA), Independent Component Analysis (ICA), Probabilistic Latent Semantic Indexing (PLSI), Latent Dirichlet Allocation (LDA), Multinomial Principal Component Analysis (MPCA)

- Watson
Syllabus

1. Introduction: motivation and background
2. Bayesian inference
3. The Bayesian network model family
4. Inference in Bayesian networks
5. Learning Bayesian networks
6. Miscellaneous topics:
   - Missing data
   - Undirected models
A quest for a calculus for plausible reasoning
Reasoning under uncertainty

- The world is a very uncertain place
- Thirty years of Artificial Intelligence and Database research danced around this fact
- And then a few AI researchers decided to use some ideas from the eighteenth century
- *Uncertainty in Artificial Intelligence* conference series 1985-
- Probabilistic reasoning now mainstream AI
Yet another probability course?

- **Computer science** point of view
- **Artificial Intelligence** point of view
  - Agent point of view (single agent in this course)
  - Knowledge representation
  - Reasoning
  - Rationality
  - Uncertainty
  - Decision making
- **Machine learning** point of view
  - Computational methods for data analysis
  - Big Data
But first there was logic

- Historically too it was first - syllogisms
  - a model of rationality
- Certainty, correctness, modularity, monotonicity
- BUT limited applicability since

  Agents almost never have access to the whole truth about their environment!
Acting by certain knowledge only?

- Is it enough to leave home 90 minutes before the flight departure?
  - Anything can happen!
- How about X minutes before departure?
  - Are you bound to stay home?

**Qualification problem:**
What are the things that have to be taken into account?
Knowledge representation in FOPL

- Let us try to use FOPL for dental diagnosis

\[ \forall p \ Symptom(p, \textit{Toothache}) \Rightarrow Disease(p, \textit{Cavity}) \]

\[ \forall p \ Symptom(p, \textit{Toothache}) \Rightarrow \begin{align*}
& Disease(p, \textit{Cavity}) \\
& \lor Disease(p, \textit{GumDisease}) \\
& \lor Disease(p, \textit{Abscess}) \\
& \ldots
\end{align*} \]

\[ \forall p \ Disease(p, \textit{Cavity}) \Rightarrow Symptom(p, \textit{Toothache}) \]

Wrong again, not all cavities cause pain!
Problems with the logical representation

- **Laziness**
  - It is too much work to list all the factors to ensure exceptionless rules

- **Theoretical ignorance**
  - We do not know all the factors that play role in the phenomenon

- **Practical Ignorance**
  - Even if we know all the factors in general, we do not know them for each particular case
Bidirectional inference?

- 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 C with certainty x" and "If B then C with certainty y". Finding C and A being both true does not make B less credible - does not explain away the cause B.
  - Example: "Fire implies smoke" and "Fog machine implies smoke"
Correlated evidence?

- From "If A then C with certainty x" and "If B then C with certainty y", how do we deduce the certainty of C when both A and B are true?
  - For example, if "Adding strawberries in the food makes it taste better with 10% certainty" and "Adding mustard in the food makes it taste better with 5% certainty", what can we say about x in "Adding both strawberries and mustard in the food makes it taste better with x% certainty"?
Probability to rescue

• Probabilities provide a disciplined way to summarize the uncertainty that comes from our laziness and ignorance:
  
  - Logic: Dropping the plate breaks it except when the plate is made of steel or such, or the floor is very soft, or somebody catches the plate before it hits the ground, or we are not in the gravity field, or ...
  
  - Probability: Dropping the plate breaks it 95% of the time.

• Probabilistic reasoning handles both bidirectional inference and correlated evidence.
Probabilistic reasoning is plausible reasoning

“The actual science of logic is conversant at present only with things either certain, impossible, or entirely doubtful, non of which (fortunately) we have to reason on. Therefore the true logic for this world is the calculus of Probabilities, which takes account of the magnitude of the probability which is, or ought to be, in a reasonable man’s mind” (James Clerk Maxwell)

“Inside every non-Bayesian there is a Bayesian struggling to get out” (Dennis V. Lindley)
Reasons for using probability theory

- **Cox/Jaynes argument:** probability is an appealing choice as a consistent calculus for plausible inference
- **Berger argument:** Decision theory offers a theoretical framework for optimal decision making, and decision theory needs probabilities
- **Pragmatic argument:** it is a very general framework and it works
Real questions

- Q1: Given plausibilities $\text{Plaus}(A)$ and $\text{Plaus}(B)$, what is $\text{Plaus}(AB)$?
- Q2: How is $\text{Plaus}(\neg A)$ related to $\text{Plaus}(A)$?
Qualitative properties of p.r.

- D1. Degrees of plausibility are represented by real numbers
- D2. Direction of inference has a qualitative correspondence with common sense
  - For example: if Plaus(A | C’) > Plaus(A | C) and Plaus(B | C’) = Plaus(B | C), then Plaus(AB | C’) > Plaus(AB | C)
  - Ensures consistency in the limit (with perfect certainty) with deductive logic
- D3. If a conclusion can be inferred in more than one way, every possible way should lead to the same result
- D4. All relevant information is always taken into account
- D5. Equivalent states of knowledge must be represented by equivalent plausibility assignments
Cox/Jaynes/Cheeseman argument

• Every allowed extension of Aristotelian logic to plausibility theory is isomorphic to Bayesian probability theory

• **Product rule** (answers question Q1)
  - \( P(AB \mid C) = P(A \mid BC) \cdot P(B \mid C) \)

• **Sum rule** (answers question Q2)
  - \( P(A \mid C) + P(\bar{A} \mid C) = 1 \)
Introduction to probabilistic graphical models
On modeling

A model is a simplified representation of the world: incomplete, but potentially useful.
Machine learning as modeling

- Based on observational data, the goal is to automatically construct models that capture useful properties of the domain.
Modeling framework

- Decision making
- Prediction
- Modeling
- Problem
What does this mean?

- **Problem**: there is a need to model some part of the universe and make decisions based on the model
- **Modeling**: build the best model possible from a priori knowledge and data available
- **Prediction**: use the model to predict properties of interest
- **Decision making**: decide actions based on the predictions
For example

- **Problem**: online troubleshooting of software/hardware
- **Modeling**: build a latent variable (Bayes) model of the problems user encounters based on knowledge about the software and symptom data
- **Prediction**: use the model to predict the underlying problem given symptoms
- **Decision making**: propose actions to remove the problem (or to find more symptoms)
Simple Printer Troubleshooter

26 variables
Instead of $2^{26}$ parameters we get
$99 = 17x1 + 1x2^1 + 2x2^2 + 3x2^3 + 3x2^4$
Probabilistic graphical models: representation

- There is an associated graph whose vertices correspond to random variables and the edges encode conditional independence relations.
  - The encoding is by convention different for different types of graphs: directed, undirected, mixed, etc.)
- All distributions in the model family obey the independence relations specified by the graph.
- Allows (often) compact encoding of distributions.
- Graphical models are human interpretable.
A simple probabilistic graphical model

- **Alarm On?**
  - yes: 0.9
  - no: 0.1

- **Over-slept?**
  - alarm:
    - yes: 0.1
    - no: 0.9

- **Bus Late?**
  - yes: 0.2
  - no: 0.8

- **In Time?**
  - bus, overs.:
    - yes, yes: 0.1
    - yes, no: 0.2
    - no, yes: 0.3
    - no, no: 0.9
Probabilistic graphical models: inference

- Answering queries using a probability distribution as our model of the world.
  - For inference, one computes the posterior distribution of some variables of interest, given evidence.
  - Inference algorithms that work directly on the graph structure are usually orders of magnitude faster than the ones which manipulate the joint distribution explicitly.
Probabilistic graphical models: learning

- Graphical models support the data-driven approach
  - Learn from data a model that provides a good approximation of our past experience.
- With graphical models, modeling can be divided into two stages:
  - Qualitative modeling stage, specifying the graph
  - Quantitative modeling stage, specifying the parameters, i.e., the numerical attributes of the model
Probabilistic models in practise

- In our department alone, probabilistic models have been used to solve for example the following problems:
  - Positioning of mobile devices
  - Classification of molecules in drug discovery
  - Determining the structure of the stemmatic tree of hand-copied texts
  - Haplotype reconstruction
  - Locating disease genes
  - Analyzing factors related to depression
  - Semantic analysis of Twitter feeds
Real questions are ... 

• Infinite number of models - what models do we consider?  
  – Model is always chosen from a set of possible models!

• How do we compare models (i.e., measure if one model is better than another one) given some data?

• How do we find good models?
...and more

- How do we **use the models to predict** unobserved quantities of interest?
- **What actions do we choose** given the predictions?
General “rational agent” framework

- Problem domain
- Learning
  - Model family
  - Data
- Inference
  - Domain model
  - Predictive distribution
- Actions
  - Decision making
  - Utilities

Problem

Data
Choice of models

- Simple models vs. complex models
  - What is complex is a totally nontrivial question
  - One intuition: a complex model has more effective parameters
- Linear models vs. non-linear models
- Parametric models vs. non-parametric models
Over-fitting

- there is a trade-off between the model complexity and fit to the data
Simpler models are "better" than complex models

- interpretation: they are easier to understand
- computation: predictions are typically easier to compute (not necessarily!)
- universality: they can be applied in more domains (more accurate predictions)
- "models should be only as complex as the data justifies"
- BUT: simpler models are NOT more probable a priori!
- Bayesian model selection: automatic Occam’s razor for model complexity regularization
Two types of modeling

- **Descriptive modeling**
  - the goal is to construct exploratory structures that help us **understand** the problem domain better
  - generative/unsupervised models

- **Predictive modeling**
  - the goal is to construct models that are able to **predict** some particular aspect of the problem domain
  - discriminative/supervised models

- ...but probabilistic generative models can be used for focused (discriminative) predictions too!
The Occam’s razor principle

- The problem:
  - You are given the following sequence: -1, 3, 7, 11
  - Question: What are the two next numbers?

- Solution 1:
  - Answer: 15 and 19
  - Explanation: add 4 to the previous number

- Solution 2:
  - Answer: -19.9 and 1043.8
  - Explanation: if the previous number is \( x \), the next one is \( -\frac{x^3}{11} + \frac{9}{11}x^2 + \frac{23}{11} \)

- “Of two competing hypotheses both conforming to our observations, choose the simpler one.”
Some viewpoints

- “prediction is our business”
- why the best fit to data is not the best predictor
  - data can be erroneous - perfect fit is too “specialized” and models the errors also!
  - a sample can only “identify” up to a certain level of complexity
- intuitive goal: minimize model complexity + prediction error - it keeps you honest!
Many modeling frameworks

- Probabilistic models
  - Statistical inference
  - Bayesian inference
- support vector machines
- fuzzy logic
- Dempster-Shafer inference
- non-monotonic logic
- (deep) neural networks
- case-based reasoning
- ...

Petri Myllymäki, University of Helsinki
Bayesian inference: preliminaries
Some early history

- Bernoulli (1654-1705)
- Bayes (1701-1761)
- Laplace (1749-1827)
- Prediction problem (“forward probability”):
  - If the probability of an outcome in a single trial is $p$, what is the relative frequency of occurrence of this outcome in a series of trials?
- Learning problem (“inverse probability”):
  - Given a number of observations in a series of trials, what are the probabilities of the different possible outcomes?
The Bayes rule

- Axioms of probability theory:
  - The sum rule:
    - \( P(A \mid C) + P(\bar{A} \mid C) = 1 \)
  - The product rule:
    - \( P(AB \mid C) = P(A \mid BC) P(B \mid C) \)
- The Bayes rule:
  - \( P(A \mid BC) = P(A \mid C) P(B \mid AC) / P(B \mid C) \)
- A rule for updating our beliefs after obtaining new information
  - \( H = \) hypothesis (model), \( I = \) background information, \( D = \) data (observations):
    - \( P(H \mid D I) = P(H \mid I) P(D \mid H I) / P(D \mid I) \)
Example: do I have a good test?

- A new home HIV test is assumed to have “95% sensitivity and 98% specificity”
- a population has HIV prevalence of 1/1000. If you use the test, what is the chance that someone testing positive actually has HIV?
Test continued ...

- $P(\text{HIV} + \mid \text{test HIV} +) = ?$
- We know that
  - $P(\text{test HIV} + \mid \text{HIV} +) = 0.95$
  - $P(\text{test HIV} + \mid \text{HIV} -) = 0.02$
- from Bayes we have learned that we can calculate the probability of having HIV given a positive test result by
  
  $$
  P(\text{test HIV} + \mid \text{HIV} +) P(\text{HIV} +) = 0.95 \times 0.001
  $$

  
  $$
  P(\text{test HIV} + \mid \text{HIV} -) P(\text{HIV} -) = 0.02 \times 0.99
  $$

  
  $$
  = \frac{0.95 \times 0.001}{0.95 \times 0.001 + 0.02 \times 0.99} = 0.045
  $$
Thus finally

- thus over 95% of those testing positive will, in fact, not have HIV
- the right question is:

How does the test result change our belief that we are HIV positive?
Bayesian?

- Probabilities can be interpreted in various ways:
  - Frequentist interpretation (Fisher, Neyman, Cramer)
  - “Degree of belief” interpretation (Bernoulli, Bayes, Laplace, Jeffreys, Lindley, Jaynes)
Frequentist says ...

- The long-run frequency of an event is the proportion of the time it occurs in a long sequence of trials - probability is this frequency
- probability can only be attached to “random variables” - not to individual events
Bayesian says ...

- an event $x$ = state of some part of the universe
- probability of $x$ is the degree of belief that event $x$ will occur
- probability will always depend on the state of knowledge
- $p(x|y,C)$ means probability of event $x$ given that event $y$ is true and background knowledge $C$
Frequentist language for solving problems

- $P(\text{data} \mid \text{model})$
- sampling distributions

Model

Potential data

Observed data
Bayesian language for solving problems

- Bayesian: \( P(\text{data} \mid \text{model}) \) & \( P(\text{model} \mid \text{data}) \)

Prior knowledge

Data

?
Isn’t this what I already do? No.

Hypothesis testing

“Sampling distribution of the estimator”

Estimator (function of data)

Data
"The Bayesian way"

Data
Likelihood
Prior distribution of the models
Posterior distribution of the models
Reasons for using probability theory

- **Cox/Jaynes argument**: probability is an appealing choice as a consistent calculus for plausible inference.
- **Berger argument**: Decision theory offers a theoretical framework for optimal decision making, and decision theory needs probabilities.
- **Pragmatic argument**: it is a very general framework and it works.
Bayesian inference: How to update beliefs?

- Select the model space
- Use Bayes theorem to obtain the posterior probability of models (given data)

$$P(\text{Model} | \text{Data}) = \frac{P(\text{Data} | \text{Model}) P(\text{Model})}{P(\text{Data})}$$

Posterior distribution is “the result” of the inference; what one needs from the posterior depends on what decisions are to be made.
The Bayesian modeling viewpoint

- Explicitly include prediction (and intervention) in modeling

Models are a means (a language) to describe interesting properties of the phenomenon to be studied, but they are not intrinsic to the phenomenon itself.

“All models are false, but some are useful.”
(Being predictive …)

- True prediction performance is a function of future data, not a model fit to current data

Good *predictive* models describe useful regularities of the data generating mechanism, while models that give a high probability to the observed data have possibly only learnt to memorize it.
Bayesian decision making for kids

- assign a benefit for every possible outcome (for every possible decision)
- assign a probability to every possible outcome given every possible decision
- what is the best decision?
Decision theory argument

• Decision theory offers a theoretical framework for optimal decision making

- P($100)=0.1, P($50)=0.9
  Expected utility: 0.1*$100+0.9*$50=$55

- P($200)=0.2, P($5)=0.8
  Expected utility: 0.2*$200+0.8*$5=$44

- P($80)=0.5, P($50)=0.5
  Expected utility: 0.5*$80+0.5*$50=$65
Optimal actions

- Optimal policy: choose the action with maximal expected utility
- The Dutch book argument: betting agencies must be Bayesians
- Where to get the utilities? (decision theory)
“Pragmatic” reasons for using probability theory

- The predictor and predicted variables (the inference task) do not have to be determined in advance
  - probabilistic models can be used for solving both classification (discriminative tasks), and configuration problems and prediction (regression problems)
  - predictions can also be used as a criteria for Data mining (explorative structures)
More pragmatic reasons for using probability theory

- consistent calculus
  - creating a consistent calculus for uncertain inference is not easy (the Cox theorem)
  - cf. fuzzy logic
- Probabilistic models can handle both discrete and continuous variables at the same time
- Various approaches for handling missing data (both in model building and in reasoning)
Nice theory, but...

- “isn’t probabilistic reasoning counter-intuitive, something totally different from human reasoning?”
- Cause for confusion: the old frequentist interpretation. But probabilities do NOT have to be thought of as frequencies, but as measures of belief.
- The so called paradoxes are often misleading
  - A: $P(\text{€1.000.000})=1.0$
  - B: $P(\text{€1.000.000})=0.25, P(\text{€4.000.000})=0.25, P(\text{€0})=0.5$
- Even if that were true, maybe that would be a good thing!
Nice theory, but...

- "Where do all the numbers come from?"
  - Bayesian networks: small number of parameters
  - the numbers do not have to be accurate
  - probability theory offers a framework for constructing models from sample data, from domain knowledge, or from their combination
We can learn from Bayesians :-) 

- Bayesian approaches never overfit (in principle)
- Bayesian approaches infer only from observed data (not possible data)
- Bayesian inference is always relative to a model family
- Does all this semi-philosophical debate really matter in practice?
  - YES!!
  - see e.g. “The great health hoax” by Robert Matthews, The Sunday Telegraph, September 13, 1998, or “Why Most Published Research Findings are False” by John Ioannidis, PLOS Medicine 2 (2005) 8.

“I rest my case”
Fundamental questions

- What is the model space?
- How do we search?
- How do we compare models?
Bayesian answers

- Model family (space) is made explicit
- Comparison criteria is a probability
- No restrictions on the search algorithm

Classical statistics answers

- Model family is implicit (normal distributions)
- Comparison criteria is fit to data, deviation from “random” behavior, “model index”
- Simple deterministic “greedy” algorithms