Graphical Models

Lecture 1:
Motivation and Foundations

Andrew McCallum
mccallum@cs.umass.edu

Thanks to Noah Smith and Carlos Guestrin for some slide materials.
Board work

- Expert systems
  - the desire for probability and dependencies.
- Joint probability tables
- Exponential blow-up in size
- MYCIN & certainty factors
Judea Pearl  (1936-)

- First proposed Bayesian Networks, qualitative structure for encoding independence relations.  c.1988
- *Probabilistic Reasoning in Intelligent Systems*
- Now working on causality.
Probabilistic Graphical Models

- Framework for obtaining, representing, querying large probability distributions.

- A beautiful formalism that generalizes many ideas from CS and from Statistics.

- Carlos Guestrin: “one of the most exciting developments in machine learning (knowledge representation, AI, EE, Stats, ...) in the last two (or three, or more) decades...”
Speech Recognition
Robot Navigation
Evolutionary Biology
Sensor Data
Planning
Images
"This is not the sort of market to have a big position in," said David Smith, who heads trading in all non-U.S. stocks.
Relational, Structured Data
Applications of GMs

• Speech recognition (hidden Markov models)
• Tracking and robot localization (Kalman filters)
• Evolutionary biology (Bayesian networks)
• Modeling sensor data (undirected GMs)
• Planning under uncertainty (dynamic BNs, factored Markov decision problems)
• Images (hierarchical BNs)
• Natural language processing (probabilistic grammars)
• Structured data like social networks and linked documents (probabilistic relational models)
• ... (your additions?)
GMs as a *Lingua Franca*

• A major barrier to research is communication: people from different technical backgrounds speak different “languages.”

• PGMs are a language with a diverse following; they enable cross-fertilization that wasn’t possible before.

• Cases in point:
  – missiles and speech
  – chemistry and CS
  – computational social science
What We’re Going to Cover

1. Representation
   – Bayesian networks (directed GMs)
   – Markov networks (undirected GMs)

2. Inference
   – exact, approximate
   – variational, sampling

3. Learning
   – parameters, structure

4. Research Topics
Keyword Soup

D-separation
Bayes Ball algorithm
Tree-width
Factor Graphs
Context-specific Independence
Partition Function
Gaussian Random Field
Exponential Family
Markov Blanket
Moralization
Probabilistic Relational Models

Variable Elimination
Junction Tree
Mean Field Inference
Variational Inference
Generalized BP
Loopy BP
Free Energy
Expectation Propagation
Collapsed Gibbs Sampling
Markov-chain Monte Carlo
Metropolis-Hastings
Blocked Gibbs Sampling

MAP Inference
Particle filtering
Linear-Programming

Expectation Maximization
Maximum Likelihood
Gradient Optimization
Conditional Random Fields
Boltzmann Machines
Deep Belief Networks
Chow-Liu Algorithm
Topic Models
Variational Bayes
Non-parametric Models
Dirichlet Process
Foundations
Events & Random Variables

• **Space** of possible **outcomes**
• An **event** is a subset of the outcomes.
• Events are complicated!
  – We tend to **group** events by **attributes**
  – Person → Age, Grade, HairColor

• **Random variables** formalize attributes:
  – “Grade = A” is shorthand for the set of events:
    \[ \{ \omega \in \Omega : f_{\text{Grade}}(\omega) = A \} \]

• Properties of random variable \( X \):
  – \( \text{Val}(X) = \text{possible values of } X \)
  – For discrete (categorical):
  – For continuous:
    \[ \sum_{x \in \text{Val}(X)} P(X = x) = 1 \]
    \[ \int P(X = x)dx = 1 \]
    \[ \forall x \in \text{Val}(X), P(X = x) \geq 0 \]
Two Interpretations of Probability

• Frequentists
  – $P(\alpha)$ is the frequency of $\alpha$ in the limit
  – Many arguments against this interpretation
    • What is the frequency of the event “it will rain tomorrow”?

• Subjective (Bayesian) interpretation
  – $P(\alpha)$ is my degree of belief that $\alpha$ will happen
  – What does “degree of belief” mean? Ground in betting.
  – Vaguely expecting a horse, catching a glimpse of a donkey, and strongly believing you’ve seen a mule.

• For this class, we (mostly) don’t care which camp you are in.
Conditional Probabilities

• After learning that \( \alpha \) is true, how do we feel about \( \beta \)? \( P(\beta \mid \alpha) \)
Chain Rule

\[ P(\beta | \alpha) = \frac{P(\alpha \cap \beta)}{P(\alpha)} \quad \text{and} \quad P(\alpha \cap \beta) = P(\alpha)P(\beta | \alpha) \]

\[ P(\alpha_1 \cap \cdots \cap \alpha_k) = P(\alpha_1)P(\alpha_2 | \alpha_1) \cdots P(\alpha_k | \alpha_1 \cap \cdots \cap \alpha_{k-1}) \]

“Factorization”
Bayes Rule

\[ P(\alpha | \beta) P(\beta) = P(\alpha \cap \beta) = P(\beta | \alpha) P(\alpha) \]

\[ P(\alpha | \beta) = \frac{P(\beta | \alpha) P(\alpha)}{P(\beta)} \]

\[ P(\alpha | \beta \cap \gamma) = \frac{P(\beta | \alpha \cap \gamma) P(\alpha | \gamma)}{P(\beta | \gamma)} \]

where \( \gamma \) is an “external event”
Marginalization and Conditioning

• (Shown in tables on board)
Back to our Problem

Large number of variables.
Massive joint probability table.
Independence

• \( \alpha \) and \( \beta \) are **independent** if
  \[ P(\beta \mid \alpha) = P(\beta) \]

• Independence implied by joint table \( P(\alpha, \beta) \)
  \[ P : (\alpha \perp \beta) \]

• \( \alpha \) and \( \beta \) are independent, **written**: \( \alpha \perp \beta \)

• **Proposition**: \( \alpha \) and \( \beta \) are **independent**
  if and only if
  \[ P(\alpha \cap \beta) = P(\alpha) \ P(\beta) \]
Conditional Independence

• Independence is rarely true.

• $\alpha$ and $\beta$ are conditionally independent given $\gamma$ if

$$P(\beta \mid \alpha \cap \gamma) = P(\beta \mid \gamma)$$

$$P : (\alpha \perp \beta \mid \gamma)$$

**Proposition:** $P : (\alpha \perp \beta \mid \gamma)$ if and only if

$$P(\alpha \cap \beta \mid \gamma) = P(\alpha \mid \gamma) P(\beta \mid \gamma)$$
Board Work

• Conditional independence and compression
Conditional Independence of Random Variables

- **Sets** of variables \(X, Y, Z\)
- \(X\) is independent of \(Y\) given \(Z\) if
  - \(P : (X=x \perp Y=y \mid Z=z)\),
  - \(\forall x \in \text{Val}(X), y \in \text{Val}(Y), z \in \text{Val}(Z)\)

- Shorthand:
  - **Conditional independence**: \(P : (X \perp Y \mid Z)\)
  - For \(P : (X \perp Y \mid \emptyset)\), write \(P : (X \perp Y)\)

- **Proposition**: \(P\) satisfies \((X \perp Y \mid Z)\) if and only if
  \[P(X,Y \mid Z) = P(X \mid Z) \cdot P(Y \mid Z)\]
Properties of Independence

• Symmetry:
  – \((X \perp Y \mid Z) \Rightarrow (Y \perp X \mid Z)\)

• Decomposition:
  – \((X \perp Y, W \mid Z) \Rightarrow (X \perp Y \mid Z)\)

• Weak union:
  – \((X \perp Y, W \mid Z) \Rightarrow (X \perp Y \mid Z, W)\)

• Contraction:
  – \((X \perp W \mid Y, Z) \land (X \perp Y \mid Z) \Rightarrow (X \perp Y, W \mid Z)\)

• Intersection:
  – \((X \perp W \mid Y, Z) \land (X \perp Y \mid W, Z) \Rightarrow (X \perp Y, W \mid Z)\)
  – Only for positive distributions: \(P(\alpha) > 0, \forall \alpha, \alpha \neq \emptyset\)
Course Administration
Learning Philosophy

• Learn by doing!
  – Don’t really understand belief-propagation until you have implemented it yourself.
  – Therefore, 7 programming assignments.

• How to make this OK.
  – Simple, yet open-ended; you define.
  – Use whatever programming language you like. MatLab, Java, F#.
  – Write a ~2 page report about your experiences.
    Informal format.
  – We will give you data, but not code.
  – No long written HW assignments
  – No midterm. No final project.
  – Mini-quizzes. “pen & pencil quizzes”. (Sometimes take home)
  – Drop lowest programming assignment grade.
Grading

• 50% homework (programming assignments)
  – We will read your reports lovingly
  – Coarse-grained grading: check, ++, +, -, --.

• 10% quizzes

• 25% final exam

• 15% class participation
  – Might include brief presentation of your HW
Prerequisites

• Helpful to have some working knowledge of...
  – Probability (distributions, densities, marginalization)
  – Basic statistics (moments, typical distributions, regression)
  – Algorithms (dynamic programming, basic data structures, a little complexity)

• Programming
  – Facility with some programming language of your choice
  – Helpful to have some experience programming for machine learning (e.g. NB, HMM)

• Ability to deal with “abstract mathematical concepts”
About the Instructor

• Main research focus:
  – Information extraction ↔ Databases ↔ Data mining
  – Natural language processing, Machine Learning
  – Joint Inference
  – Structured prediction
  – Semi-supervised learning

• Why I’m teaching this course
  – promote the topic
  – perceived need & interest in the department
About the TAs

• Michael Wick
  – parameter estimation in undirected graphical models
  – probabilistic databases

• Sameer Singh
  – joint inference
  – parallel & distributed graphical models
More Info & Contacting Us

• Course web site:
  http://www.cs.umass.edu/~mccallum/courses/gm2011

• Mailing Lists (coming soon):
  691gm-staff@cs.umass.edu
  691gm-all@cs.umass.edu
Summary So Far

• Graphical model motivation.
• Basic definitions of probabilities, independence, conditional independence, chain rule, bayes rule
• Basic idea: exploiting conditional independence to compress joint prob table.

• Next time: semantics of Bayesian networks