Probabilistic Graphical Models

Introduction to GM

Eric Xing
Lecture 1, January 13, 2020

Reading: see class homepage
Logistics

- Class webpage: http://www.cs.cmu.edu/~epxing/Class/10708-20/

10-708 – Probabilistic Graphical Models
2020 Spring

Many of the problems in artificial intelligence, statistics, computer systems, computer vision, natural language processing, and computational biology, among many other fields, can be viewed as the search for a coherent global conclusion from local information. The probabilistic graphical models framework provides an unified view for this wide range of problems, enables efficient inference, decision-making and learning in problems with a very large number of attributes and huge datasets. This graduate-level course will provide you with a strong foundation for both applying graphical models to complex problems and for addressing core research topics in graphical models.

- Instructor: Eric P. Xing (epxing@cs)
- Time: MW 12:00-1:30pm
- Location: Wean 7500
- Office Hours: TBA
- Piazza: https://www.piazza.com/cmu/spring2020/10708
- Gradescope: https://www.gradescope.com/courses/80181
- TAs (email, office hours):
  - Xun Zheng (xzheng1@andrew, TBA)
  - Ben Lengerich (blengeri@andrew, TBA)
  - Haohan Wang (haohanw@andrew, TBA)
  - Yiwen Yuan (yiweny@andrew, TBA)
  - Xiang Si (xsi@andrew, TBA)
  - Junxian He (junxian1@andrew, TBA)
Logistics

- **Textbooks:**
  - Daphne Koller and Nir Friedman, Probabilistic Graphical Models
  - M. I. Jordan, An Introduction to Probabilistic Graphical Models (chapters will be made available)

- **Class announcements and discussion:** Piazza

- **Homework submission:** Gradescope

- **TAs:**
  - Xun Zheng
  - Ben Lengerich
  - Haohan Wang
  - Yiwen Yuan
  - Xiang Si
  - Junxian He

- **Lecturer:** Eric Xing

- **Class Assistant:** Amy Protos
Logistics

- 4 homework assignments: 50% of grade
  - Theory exercises, Implementation exercises
- Scribe duties: 10% (~once to twice for the whole semester)
- Final project: 40% of grade
  - Applying PGM to the development of a real, substantial ML system
    - Design and Implement a (record-breaking) distributed Logistic Regression, Gradient Boosted Tree, Deep Network, or Topic model on Petuum and apply to ImageNet, Wikipedia, and/or other data
    - Build a web-scale topic or story line tracking system for news media, or a paper recommendation system for conference review matching
    - An online car or people or event detector for web-images and webcam
    - An automatic “what’s up here?” or “photo album” service on iPhone
  - Theoretical and/or algorithmic work
    - a more efficient approximate inference or optimization algorithm, e.g., based on stochastic approximation, proximal average, or other new techniques
    - a distributed sampling scheme with convergence guarantee
  - 3 or 4-member team to be formed in the first three weeks, proposal, mid-way report, presentation & demo, final report → possibly conference submission!
- Bonus:
  - Contribution to discussion on Piazza
  - Complete mid-semester evaluation
Past projects:

- We will have a prize for the best project(s) ...

- **Award Winning Projects:**
  Others … such as KDD 2014 best paper

- **Other projects:**
Recap of Basic Prob. Concepts

- Representation: what is the joint probability dist. on multiple variables?
  \[ P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8) \]
  - How many state configurations in total? --- \(2^8\)
  - Are they all needed to be represented?
  - Do we get any scientific/medical insight?

- Learning: where do we get all this probabilities?
  - Maximal-likelihood estimation? but how many data do we need?
  - Are there other est. principles?
  - Where do we put domain knowledge in terms of plausible relationships between variables, and plausible values of the probabilities?

- Inference: If not all variables are observable, how to compute the conditional distribution of latent variables given evidence?
  - Computing \(P(H|A)\) would require summing over all \(2^8\) configurations of the unobserved variables
Multivariate Distribution in High-D Space

A possible world for cellular signal transduction:

- Receptor A: $X_1$
- Receptor B: $X_2$
- Kinase C: $X_3$
- Kinase D: $X_4$
- Kinase E: $X_5$
- TF F: $X_6$
- Gene G: $X_7$
- Gene H: $X_8$
A Structured View From Domain Experts

- Dependencies among variables

Receptor A \( \rightarrow \) Kinase C \( \rightarrow \) TF F \( \rightarrow \) Gene G

Receptor B \( \rightarrow \) Kinase D \( \rightarrow \) Kinase E \( \rightarrow \) Gene H

Membrane \( \rightarrow \) Cytosol

\( X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8 \)
What are graphical models?

- Informally, a GM is just a graph representing **relationship** among random variables
- Nodes: random variables (features, not examples)
- Edges (or absence of edges): relationship

- Looks simple!
  - But detail matters, as always.
  - What exactly do we mean by **relationship**?
Relationship between two random variables

- Many types of relationships exist:
  - $X$ and $Y$ are correlated
  - $X$ and $Y$ are dependent
  - $X$ and $Y$ are independent
  - $X$ and $Y$ are partially correlated given $Z$
  - $X$ and $Y$ are conditionally dependent given $Z$
  - $X$ and $Y$ are conditionally independent given $Z$
  - $X$ causes $Y$
  - $Y$ causes $X$
  - ...
- Many of them can be measured by an “one number summary”
Measure of association between two random variables

- Measures of association:
  - Pearson’s correlation
  - Mutual information
  - Hilbert-Schmidt Independence Criterion (HSIC)
  - Partial correlation
  - …

- Why study them? (rather than directly diving into graphical models?)
  - Gives better understanding of what graphical models really mean
  - Useful when estimating graph from data (later in the course)
Probability 101: Pearson’s correlation

- Normalized covariance
  \[
  \rho(X,Y) = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}
  \]

- Captures \textbf{linear} dependency
  - Linear regression from X to Y gives
  \[
  \beta = \frac{\text{Cov}(X,Y)}{\text{Var}(X)}
  \]

- Important properties:
  - X \perp Y implies \( \rho(X,Y) = 0 \) (Why?)
  - \( \rho(X,Y) = 0 \) does not imply X \perp Y (Counterexamples?)

- Q1: Is there any measure that implies independence?
- Q2: What kind of dependency should they consider?
Q1: Is there any measure that implies independence?
   A1: Yes, many! We will mention two of them today.

Q2: What kind of dependency should they consider?
   A2: Nonlinear dependency.

One way to construct such a measure of dependence:
   - If $X \perp Y$ then joint pdf factorizes $P_{XY} = P_X P_Y$
   - Measure "distance" between $P_{XY}$ and $P_X P_Y$
   - distance $== 0$ if and only if $X \perp Y$
Mutual information

- Distance between two distributions?
- Recall our old friend – the Kullback–Leibler divergence

\[
KL(P, Q) = \int_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{Q(x)} dx
\]

- Apply then we get another old friend – mutual information

\[
I(X, Y) = KL(P_{XY}, P_X P_Y)
\]

- Foundation of many topics later in the course
- \( I(X, Y) = 0 \) if and only if \( X \perp Y \)
Hilbert-Schmidt Independence Criterion (HSIC)

- A relatively recent(?) finding by Gretton et al. 2005
- Use **maximum mean discrepancy** (MMD) as the distance metric

\[
\text{HSIC}(X, Y) = \text{MMD}(P_{XY}, P_X P_Y)
\]

\[
\text{MMD}(P, Q) = \|\mu_k(P) - \mu_k(Q)\|_{\mathcal{H}_k}
\]

\[
\mu_k(P) = \mathbb{E}_{Z \sim P}[\phi(Z)] \quad \text{(kernel embedding of } P)\]

\[
\phi(Z) = \text{feature map of kernel } k
\]

- Looks scary! No need to know what it means for now.
- Will cover later if anyone’s interested 😊

- \( \text{HSIC}(X, Y) = 0 \) if and only if \( X \perp Y \)
But what do they have to do with graphical models?

- **Marginal** correlation/dependency graph for $\mathbf{X} = \{X_1, \ldots, X_d\}$
  - Most **primitive** form of graphical models one can think of
  - Connect variables that have nontrivial pairwise correlation/mutual information/HSIC/etc.

- **Not** very informative. Why?
  - $X$ = height of a kid
  - $Y$ = vocabulary of a kid
  - $Z$ = age of a kid
  - Q1: What is the marginal dependency graph?
  - Q2: What is the graph that you think will make more sense?
Partial correlation: accounting for other variables

- Partial correlation between X and Y given a random vector Z
  - Correlation measured after eliminating **linear** effect of Z
  - i.e. correlation between **residuals** from regressing Z to X and Z to Y

\[
\rho(X, Y|Z) = \rho(e_X, e_Y) = \frac{\text{Cov}(e_X, e_Y)}{\sqrt{\text{Var}(e_X) \text{Var}(e_Y)}}
\]

\[
e_X = X - (\beta_X^T Z + \text{intercept}_X)
\]

\[
e_Y = Y - (\beta_Y^T Z + \text{intercept}_Y)
\]

- Similar to Pearson’s correlation:
  - \(X \perp Y \mid Z\) implies \(\rho(X,Y \mid Z) = 0\)
  - \(\rho(X,Y \mid Z) = 0\) does not imply \(X \perp Y \mid Z\)
Partial correlation graphs

- Partial correlation graph for \( X = \{X_1, \ldots, X_d\} \)
  - A more informative graphical model than marginal dependency graph
  - Connect variables with nontrivial partial correlation given the rest
  - Recall the height-vocab-age example (assuming everything is linear)

- A deeper look at the \( d \times d \) partial correlation matrix \( R \) with

\[
R_{ij} = \rho(X_i, X_j | X_{-ij})
\]

- Looks scary! (So many regressions to run?!) 
- But turns out \( R \) is just some version of inverse covariance matrix \( \Theta \) 
- Homework \( \smile \)

\[
R_{ij} = -\frac{\Theta_{ij}}{\sqrt{\Theta_{ii}\Theta_{jj}}}
\]
Conditional independence

- How do we measure conditional (in)dependence?
  - After seeing strong dependency measures and partial correlation, conditional independency appears to be harder than we thought…

- Ancient wisdom: if something is hard, assume Gaussian.

- If \((X, Y, Z)\) are jointly Gaussian, \(\rho(X,Y \mid Z) = 0\) if and only if \(X \perp Y \mid Z\)
  - We will see later that many papers with Gaussian assumption rely on this fact, even though it is rarely explicitly stated
Short Summary

Measure of association between X and Y

Marginal

Non-marginal (partial)

Linear

Nonlinear

Linear

Nonlinear

Pearson's correlation

$\text{dist}(P_{XY}, P_X P_Y)$

Partial correlation

Conditional independence

KL divergence

Max mean discrepancy

Mutual information

Hilbert-Schmidt Independence Criterion

What's next?
Lecture 2: Conditional independence graph

- Go by many different names
  - Conditional independence graphs (CIG)
  - Markov networks (MN)
  - Markov random fields (MRF)
  - Undirected graphical models (UG)

- Many interesting properties, widely used in physics, statistics, computer vision, NLP, deep learning, bioinformatics, coding theory, finance, …
Lecture 3: Directed graphical models

- Another major class of models, also has many names:
  - Directed graphical models
  - Directed acyclic graphs (DAG) (cyclic model exists but hard to work with)
  - Bayesian networks (BN)
  - Structural equation models (SEM)
  - Structural causal models (SCM)
  - …

- Powerful language to express structured knowledge

\[
P(X_{16}) = P(X_1)P(X_2)P(X_3 | X_1, X_2)P(X_4 | X_2)P(X_5 | X_2)
P(X_6 | X_3, X_4)P(X_7 | X_6)P(X_8 | X_5, X_6)
\]
Lecture 4-13 (tentative): Inference and Learning

- Given a graphical model representing our knowledge

- Inference:
  - What is the marginal/conditional density?
  - What is the mean of the marginal/conditional?
  - What is the mode of the marginal/conditional?
  - Can we draw samples from the marginal/conditional?
  - ...

- Learning: Statistical parameter estimation and model selection
Lecture 15-end (tentative): Modern GMs

- Relationship between deep learning and graphical models
- Deep generative models and their unified view
- Reinforcement learning as probabilistic inference
- GMs on functions and sets
- Bayesian nonparametrics
- Large-scale algorithms and systems

2-3 open slots:
- We will list several candidate topics
- Your voice matters!
Why graphical models

- A language for communication
- A language for computation
- A language for development

Origins:
- Wright 1920’s
- Independently developed by Spiegelhalter and Lauritzen in statistics and Pearl in computer science in the late 1980’s
Probabilistic Graphical Models

- If $X_i$'s are conditionally independent (as described by a PGM), the joint can be factored to a product of simpler terms, e.g.,

$$P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8) = P(X_1) P(X_2) P(X_3|X_1) P(X_4|X_2) P(X_5|X_2) P(X_6|X_3, X_4) P(X_7|X_6) P(X_8|X_5, X_6)$$

- Why we may favor a PGM?
  - Incorporation of domain knowledge and causal (logical) structures
  - Modular combination of heterogeneous parts – data fusion
  - Bayesian Philosophy
    - Knowledge meets data

© Eric Xing @ CMU, 2005-2020 26
Why graphical models

- **Probability theory** provides the glue whereby the parts are combined, ensuring that the system as a whole is consistent, and providing ways to interface models to data.

- The **graph theoretic** side of graphical models provides both an intuitively appealing interface by which humans can model highly-interacting sets of variables as well as a data structure that lends itself naturally to the design of efficient general-purpose algorithms.

- Many of the classical multivariate probabilistic systems studied in fields such as statistics, systems engineering, information theory, pattern recognition and statistical mechanics are special cases of the general graphical model formalism.

- The graphical model framework provides a way to view all of these systems as instances of a common underlying formalism.

--- M. Jordan
Questions?
What Are Graphical Models?

**Graph**

**Model**

\[ M_G \]

**Data**

\[ D = \{ X_1^{(i)}, X_2^{(i)}, \ldots, X_m^{(i)} \}_{i=1}^N \]
Reasoning under **uncertainty**!
The Fundamental Questions

- **Representation**
  - How to capture/model uncertainties in possible worlds?
  - How to encode our domain knowledge/assumptions/constraints?

- **Inference**
  - How do I answer questions/queries according to my model and/or based given data?
    
    e.g.: \( P(X_i | \mathcal{D}) \)

- **Learning**
  - What model is "right" for my data?
    
    e.g.: \( M = \arg \max_{M \in \mathcal{M}} F(\mathcal{D}; M) \)
Recap of Basic Prob. Concepts

- Representation: what is the joint probability dist. on multiple variables?

\[ P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8) \]

- How many state configurations in total? --- 2^8
- Are they all needed to be represented?
- Do we get any scientific/medical insight?

- Learning: where do we get all this probabilities?
  - Maximal-likelihood estimation? but how many data do we need?
  - Are there other est. principles?
  - Where do we put domain knowledge in terms of plausible relationships between variables, and plausible values of the probabilities?

- Inference: If not all variables are observable, how to compute the conditional distribution of latent variables given evidence?
  - Computing \( p(H|A) \) would require summing over all 2^6 configurations of the unobserved variables.
What is a Graphical Model?
--- Multivariate Distribution in High-D Space

- A possible world for cellular signal transduction:

  - Receptor A $X_1$
  - Receptor B $X_2$
  - Kinase C $X_3$
  - Kinase D $X_4$
  - Kinase E $X_5$
  - TF F $X_6$
  - Gene G $X_7$
  - Gene H $X_8$

What is a Graphical Model?
--- Multivariate Distribution in High-D Space

- A possible world for cellular signal transduction:
GM: Structure Simplifies Representation

- Dependencies among variables

- Receptor A \( \mathbf{x}_1 \) → Kinase C \( \mathbf{x}_3 \) → TF F \( \mathbf{x}_6 \) → Gene G \( \mathbf{x}_7 \)
- Kinase E \( \mathbf{x}_5 \) → Kinase D \( \mathbf{x}_4 \) → TF F \( \mathbf{x}_6 \) → Gene H \( \mathbf{x}_8 \)
- Receptor B \( \mathbf{x}_2 \) → Kinase E \( \mathbf{x}_5 \)
Probabilistic Graphical Models

- If $X_i$'s are conditionally independent (as described by a PGM), the joint can be factored to a product of simpler terms, e.g.,

\[
P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8) = P(X_1) P(X_2) P(X_3|X_1) P(X_4|X_2) P(X_5|X_2)
\]
\[
P(X_6|X_3, X_4) P(X_7|X_6) P(X_8|X_5, X_6)
\]

Stay tune for what are these independencies!

- Why we may favor a PGM?
  - Incorporation of domain knowledge and causal (logical) structures

1+1+2+2+2+4+2+4=18, a 16-fold reduction from $2^8$ in representation cost!
GM: Data Integration

- Receptor A
- Kinase C
- TF F
- Gene G

- Kinase D
- Gene H

- Receptor B
- Kinase E

Factors: $X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9$
More Data Integration

- Text + Image + Network ➔ Holistic Social Media

- Genome + Proteome + Transcriptome + Phenome + … ➔ PanOmic Biology
If $X_i$'s are conditionally independent (as described by a PGM), the joint can be factored to a product of simpler terms, e.g.,

$$P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8) = P(X_2) P(X_4 | X_2) P(X_5 | X_2) P(X_1) P(X_3 | X_1) P(X_6 | X_3, X_4) P(X_7 | X_6) P(X_8 | X_5, X_6)$$

Why we may favor a PGM?

- Incorporation of domain knowledge and causal (logical) structures
  
  $2+2+4+4+4+8+4+8=36$, an 8-fold reduction from $2^8$ in representation cost!

- Modular combination of heterogeneous parts – data fusion
This allows us to capture uncertainty about the model in a principled way.

But how can we specify and represent a complicated model?

Typically, the number of genes need to be modeled are in the order of thousands!
GM: MLE and Bayesian Learning

- Probabilistic statements of $\Theta$ is conditioned on the values of the observed variables $A_{\text{obs}}$ and prior $p(\mid \chi)$

$$\Theta_{\text{Bayes}} = \int \Theta \ p(\Theta \mid A, \chi) \ d\Theta$$

$$p(\Theta; \chi)$$

$$p(\Theta \mid A; \chi) \propto p(A \mid \Theta) p(\Theta; \chi)$$

posterior  likelihood  prior
If \( X_i \)'s are conditionally independent (as described by a PGM), the joint can be factored to a product of simpler terms, e.g.,

\[
P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8) = P(X_1) P(X_2) P(X_3| X_1) P(X_4| X_2) P(X_5| X_2) P(X_6| X_3, X_4) P(X_7| X_6) P(X_8| X_5, X_6)
\]

Why we may favor a PGM?
- Incorporation of domain knowledge and causal (logical) structures
- Modular combination of heterogeneous parts – data fusion 
- Bayesian Philosophy
  - Knowledge meets data
So What Is a PGM After All?

In a nutshell:

PGM = Multivariate Statistics + Structure

GM = Multivariate Obj. Func. + Structure
So What Is a PGM After All?

- The informal blurb:
  - It is a smart way to write/specify/compose/design exponentially-large probability distributions without paying an exponential cost, and at the same time endow the distributions with **structured semantics**.

- A more formal description:
  - It refers to a family of distributions on a set of random variables that are compatible with all the probabilistic independence propositions encoded by a graph that connects these variables.

\[
P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8) = P(X_1)P(X_2|X_1)P(X_3|X_1X_2)P(X_4|X_2)P(X_5|X_2)P(X_6|X_3X_4)P(X_7|X_6)P(X_8|X_5X_6)
\]
Two types of GMs

- **Directed edges** give causality relationships (Bayesian Network or Directed Graphical Model):

  \[ P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8) = P(X_1) P(X_2) P(X_3 | X_1) P(X_4 | X_2) P(X_5 | X_2) P(X_6 | X_3, X_4) P(X_7 | X_6) P(X_8 | X_5, X_6) \]

- **Undirected edges** simply give correlations between variables (Markov Random Field or Undirected Graphical model):

  \[ P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8) = \frac{1}{Z} \exp\{E(X_1)+E(X_2)+E(X_3, X_1)+E(X_4, X_2)+E(X_5, X_2) + E(X_6, X_3, X_4)+E(X_7, X_6)+E(X_8, X_5, X_6)\} \]
Bayesian Networks

- **Structure:** *DAG*

  - **Meaning:** A node is conditionally independent of every other node in the network outside its Markov blanket.

  - **Local conditional distributions (CPD)** and the **DAG** completely determine the joint dist.

  - **Give causality relationships, and facilitate a generative process**
Markov Random Fields

- **Structure**: *undirected graph*
  - Meaning: a node is conditionally independent of every other node in the network given its Directed neighbors.
  - Local contingency functions (potentials) and the cliques in the graph completely determine the joint dist.
  - Give correlations between variables, but no explicit way to generate samples.
Towards structural specification of probability distribution

- Separation properties in the graph imply independence properties about the associated variables.
- For the graph to be useful, any conditional independence properties we can derive from the graph should hold for the probability distribution that the graph represents.

- **The Equivalence Theorem**
  
  For a graph $G$,
  
  Let $D_1$ denote the family of all distributions that satisfy $I(G)$,
  
  Let $D_2$ denote the family of all distributions that factor according to $G$,
  
  Then $D_1 \equiv D_2$. 
GMs are your old friends

Density estimation
  Parametric and nonparametric methods

Regression
  Linear, conditional mixture, nonparametric

Classification
  Generative and discriminative approach

Clustering
An (incomplete) genealogy of graphical models

(Picture by Zoubin Ghahramani and Sam Roweis)
Fancier GMs: reinforcement learning

- Partially observed Markov decision processes (POMDP)
Fancier GMs: machine translation

The HM-BiTAM model
(B. Zhao and E.P Xing, ACL 2006)
Fancier GMs: genetic pedigree

An allele network
Fancier GMs: solid state physics

Ising/Potts model
Application of GMs

- Machine Learning
- Computational statistics
- Computer vision and graphics
- Natural language processing
- Informational retrieval
- Robotic control
- Decision making under uncertainty
- Error-control codes
- Computational biology
- Genetics and medical diagnosis/prognosis
- Finance and economics
- Etc.