One of the most exciting developments in machine learning (knowledge representation, AI, EE, Stats,...) in the last two (or three, or more) decades...

My expectations are already high... 😊
Speech recognition

Hidden Markov models and their generalizations

Tracking and robot localization

Kalman Filters

[Fox et al.]  [Funiak et al.]
Evolutionary biology

Bayesian networks

[Friedman et al.]

Modeling sensor data

Undirected graphical models

[Guestrin et al.]
Planning under uncertainty

Dynamic Bayesian networks
Factored Markov decision problems

Images and text data

Hierarchical Bayesian models
Structured data (text, webpages, …)

Probabilistic relational models

- Tom Mitchell (Professor)
- Project-of
- WebKB (Project)
- Member
- Advisor-of
- Sean Slattery (Student)

[Koller et al.]

And many

many

many

many

many

more...
Syllabus

- Covers a wide range of Probabilistic Graphical Models topics – from basic to state-of-the-art
- You will learn about the methods you heard about:
  - Bayesian networks, Markov networks, factor graphs, conditional random fields, decomposable models, junction trees, parameter learning, structure learning, semantics, exact inference, variable elimination, context-specific independence, approximate inference, sampling, importance sampling, MCMC, Gibbs, variational inference, loopy belief propagation, generalized belief propagation, Kikuchi, Bayesian learning, missing data, EM, Chow-Liu, structure search, IPF for tabular MRFs, Gaussian and hybrid models, discrete and continuous variables, temporal and template models, hidden Markov Models, Forwards-Backwards, Viterbi, Baum-Welch, Kalman filter, linearization, switching Kalman filter, assumed density filtering, DBNs, BK, Relational probabilistic models, Causality,…
- Covers algorithms, theory and applications
- It’s going to be fun and hard work 😊

Prerequisites

- 10-701 – Machine Learning, especially:
  - Probabilities
    - Distributions, densities, marginalization…
  - Basic statistics
    - Moments, typical distributions, regression…
- Algorithms
  - Dynamic programming, basic data structures, complexity…
- Programming
  - Matlab will be very useful
- We provide some background, but the class will be fast paced
- Ability to deal with “abstract mathematical concepts”
Review Sessions

- Very useful!
  - Review material
  - Present background
  - Answer questions

- Thursdays, 5:00-6:20 in Wean Hall 5409

- First recitation is **this Thursday**
  - Review of probabilities & statistics

- Sometimes this semester: Especial recitations most likely on Mondays 5:30-7pm
  - Cover special topics that we can’t cover in class
  - These are optional, but you are here to learn… 😊

- Do we need a Matlab review session?

Staff

- Two Great TAs: Great resource for learning, interact with them!
  - Amr Ahmed <amahmed@cs.cmu.edu>
  - Dhruv Batra <batradhruv@cmu.edu>

- Administrative Assistant
  - Michelle Martin
    <michelle324@cs.cmu.edu>,
    Wean 4619, x8-5527
First Point of Contact for HWs

- To facilitate interaction, a TA will be assigned to each homework question – This will be your “first point of contact” for this question
  - But, you can always ask any of us

- For e-mailing instructors, always use:
  - 10708-instr@cs.cmu.edu

- For announcements, subscribe to:
  - 10708-announce@cs
  - https://mailman.srv.cs.cmu.edu/mailman/listinfo/10708-announce

- We will also use a discussion group:
  - Post your questions, discuss projects, etc
  - Be nice… 😊
  - Don’t give away any answers… 😊
  - http://groups.google.com/group/10708-f08

Text Books

- **Primary**: Daphne Koller and Nir Friedman, *Structured Probabilistic Models*, in preparation. These chapters are part of the course reader. You can purchase one from Michelle Martin

Grading

- 5 homeworks (50%)
  - First one goes out next Wednesday!
  - Homeworks are long and hard 😊
    - please, please, please, please, please, please start early!!!

- Final project (30%)
  - Done individually or in pairs
  - Details out soon
  - Proposals due October 6th

- Final (20%)
  - Take home, out Dec. 3rd
  - Due Dec. 10th at NOON (hard deadline)

Homeworks

- Homeworks are hard, start early 😊
- Due in the beginning of class
- 3 late days for the semester
- After late days are used up:
  - Half credit within 48 hours
  - Zero credit after 48 hours
- All homeworks must be handed in, even for zero credit
- Late homeworks handed in to Michelle Martin, WEH 4619

- Collaboration
  - You may discuss the questions
  - Each student writes their own answers
  - Write on your homework anyone with whom you collaborate

- IMPORTANT:
  - We may use some material from previous years or from papers for the homeworks.
  - Unless otherwise specified, please only look at the readings when doing your homework.
  - You are taking this advanced graduate class because you want to learn, so this rule is self-enforced 😊
Enjoy!

- NO CLASS THIS WEDNESDAY 9/10

- Probabilistic graphical models are having significant impact in science, engineering and beyond
- This class should give you the basic foundation for applying GMs and developing new methods
- The fun begins…

What are the fundamental questions of graphical models?

- Representation:
  - What are the types of models?
  - What does the model mean/imply/assume? (Semantics)

- Inference:
  - How do I answer questions/queries with my model?

- Learning:
  - What model is the right for my data?
More details??

**Representation:**
- Graphical models represent exponentially large probability distributions compactly
- **Key concept:** *Conditional Independence*

**Inference:**
- What is the probability of $X$ given some observations?
- What is the most likely explanation for what is happening?
- What decisions should I make?

**Learning:**
- What are the right/good parameters for the model?
- How do I obtain the structure of the model?

Where do we start?

- **From Bayesian networks**
- "Complete" BN presentation first
  - Representation
  - Exact inference
  - Learning
  - Only discrete variables for now
- Later in the semester
  - Undirected models
  - Approximate inference
  - Continuous
  - Temporal models
  - And more...
- Class focuses on fundamentals – Understand the foundation and basic concepts
Today

- Probabilities
- Independence
- Two nodes make a BN
- Naïve Bayes

Should be a review for everyone – Setting up notation for the class

Random variable

- Probability distributions usually defined by events
- Events are complicated – we think about attributes
  - Age, Grade, HairColor
- Random variables formalize attributes:
  - Grade=A —— shorthand for event \( \{ \omega \in \Omega : f_{\text{Grade}}(\omega) = A \} \)

Properties of random vars, X:

- Val(X) = possible values of random var X
- For discrete (categorical): \( \sum_{i=1}^{\text{|Val(X)|}} P(X=x_i) = 1 \)
- For continuous: \( \int p(X=x)dx = 1 \)
- \( P(x) \geq 0 \)
Interpretations of probability – A can of worms!

- **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 interpretation**
  - $P(\alpha)$ is my degree of belief that $\alpha$ will happen
  - What the …. does “degree of belief mean”?
  - If I say $P(\alpha)=0.8$, then I am willing to bet!!!

- For this class, we (mostly) don’t care what camp you are in

---

Conditional probabilities

- After learning that $\alpha$ is true, how do we feel about $\beta$?
- $P(\beta|\alpha)$
Two of the most important rules of the semester: 1. The chain rule

- \( P(\alpha \cap \beta) = P(\alpha)P(\beta | \alpha) \)

- More generally:
  - \( P(\alpha_1 \cap \ldots \cap \alpha_k) = P(\alpha_1)P(\alpha_2 | \alpha_1)\ldots P(\alpha_k | \alpha_1 \cap \ldots \cap \alpha_{k-1}) \)

Two of the most important rules of the semester: 2. Bayes rule

- \( P(\alpha | \beta) = \frac{P(\beta | \alpha)P(\alpha)}{P(\beta)} \)

- More generally, external event \( \gamma \):
  - \( P(\alpha | \beta \cap \gamma) = \frac{P(\beta | \alpha \cap \gamma)P(\alpha | \gamma)}{P(\beta | \gamma)} \)
Most important concept:
a) Independence

- $\alpha$ and $\beta$ \textit{independent}, if $P(\beta|\alpha) = P(\beta)$
  - $P \rightarrow (\alpha \perp \beta)$

**Proposition:** $\alpha$ and $\beta$ \textit{independent} if and only if $P(\alpha \cap \beta) = P(\alpha)P(\beta)$

Most important concept:
b) Conditional independence

- Independence is rarely true, but conditionally…

- $\alpha$ and $\beta$ \textit{conditionally independent} given $\gamma$ if $P(\beta|\alpha \cap \gamma) = P(\beta|\gamma)$
  - $P \rightarrow (\alpha \perp \beta | \gamma)$

**Proposition:** $P \rightarrow (\alpha \perp \beta | \gamma)$ if and only if $P(\alpha \cap \beta | \gamma) = P(\alpha | \gamma)P(\beta | \gamma)$
Joint distribution, Marginalization

- Two random variables – Grade & Intelligence

Marginalization – Compute marginal over single var

Marginalization – The general case

- Compute marginal distribution $P(X_i)$:

$$P(X_1, X_2, \ldots, X_i) = \sum_{x_{i+1}, \ldots, x_n} P(X_1, X_2, \ldots, X_i, x_{i+1}, \ldots, x_n)$$

$$P(X_i) = \sum_{x_1, \ldots, x_{i-1}} P(x_1, \ldots, x_{i-1}, X_i)$$
Basic concepts for random variables

- Atomic outcome: assignment $x_1, \ldots, x_n$ to $X_1, \ldots, X_n$

- Conditional probability: $P(X, Y) = P(X)P(Y|X)$

- Bayes rule: $P(X|Y) = \ldots$

- Chain rule:
  - $P(X_1, \ldots, X_n) = P(X_1)P(X_2|X_1)\ldots P(X_k|X_1, \ldots, X_{k-1})$

Conditionally independent random variables

- Sets of variables $X$, $Y$, $Z$

- $X$ is independent of $Y$ given $Z$ if
  - $P \rightarrow (X=x \perp Y=y | Z=z), \forall x \in \text{Val}(X), y \in \text{Val}(Y), z \in \text{Val}(Z)$

- Shorthand:
  - Conditional independence: $P \rightarrow (X \perp Y | Z)$
  - For $P \rightarrow (X \perp Y | \emptyset)$, write $P \rightarrow (X \perp Y)$

- Proposition: $P$ satisfies $(X \perp Y | Z)$ if and only if
  - $P(X, Y|Z) = P(X|Z)P(Y|Z)$
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:
  - $(X \perp Y \mid W,Z) \land (X \perp W \mid Y,Z) \Rightarrow (X \perp Y,W \mid Z)$
  - Only for positive distributions!
  - $P(\alpha) > 0, \forall \alpha, \alpha \neq \emptyset$

- Notation: $I(P)$ – independence properties entailed by $P$

---

Bayesian networks

- One of the most exciting recent advancements in statistical AI
- Compact representation for exponentially-large probability distributions
- Fast marginalization too
- Exploit conditional independencies
Handwriting recognition

Handwriting recognition 2
Webpage classification

Company home page vs Personal home page vs University home page vs ...

Webpage classification 2
Let’s start on BNs…

- Consider $P(X_i)$
  - Assign probability to each $x_i \in \text{Val}(X_i)$
  - Independent parameters

- Consider $P(X_1,\ldots,X_n)$
  - How many independent parameters if $|\text{Val}(X_i)|=k$?

What if variables are independent?

- What if variables are independent?
  - $(X_i \perp X_j), \forall i,j$
  - Not enough!!! (See homework 1 😊)
  - Must assume that $(X \perp Y), \forall X,Y$ subsets of $\{X_1,\ldots,X_n\}$

- Can write
  - $P(X_1,\ldots,X_n) = \prod_{i=1}^{n} P(X_i)$

- How many independent parameters now?
Conditional parameterization – two nodes

- Grade is determined by Intelligence

Conditional parameterization – three nodes

- Grade and SAT score are determined by Intelligence
- \((G \perp S \mid I)\)
The naïve Bayes model – Your first real Bayes Net

- Class variable: C
- Evidence variables: $X_1, \ldots, X_n$
- Assume that $(X \perp Y | C), \forall X, Y$ subsets of \{X_1, \ldots, X_n\}

What you need to know

- Basic definitions of probabilities
- Independence
- Conditional independence
- The chain rule
- Bayes rule
- Naïve Bayes
Next class

- We’ve heard of Bayes nets, we’ve played with Bayes nets, we’ve even used them in your research
- Next class, we’ll learn the semantics of BNs, relate them to independence assumptions encoded by the graph