Probabilistic Graphical Models

David Sontag

New York University

Lecture 1, January 31, 2013
One of the most exciting advances in machine learning (AI, signal processing, coding, control, ...) in the last decades
How can we gain **global insight** based on **local observations**?
Key idea

1. **Represent** the world as a collection of random variables $X_1, \ldots, X_n$ with joint distribution $p(X_1, \ldots, X_n)$

2. **Learn** the distribution from data

3. **Perform “inference”** (compute conditional distributions $p(X_i \mid X_1 = x_1, \ldots, X_m = x_m)$)
As humans, we are continuously making predictions under uncertainty.

Classical AI and ML research ignored this phenomena.

Many of the most recent advances in technology are possible because of this new, *probabilistic*, approach.
Applications: Deep question answering
The top U.S. general, visiting Israel at a delicate and dangerous moment in the global standoff with Tehran, is expected to press for restraint amid fears that the Jewish state is nearing a decision to attack Iran's nuclear program.

El máximo general de EE.UU., de visita en Israel en un momento delicado y peligroso en el enfrentamiento global con Teherán, se espera que presione a la moderación en medio de temores de que el estado judío se acerca a una decisión de atacar el programa nuclear de Irán.
Applications: Speech recognition

“ I need to hide a body ”

What kind of place are you looking for?

- reservoirs
- metal foundries
- mines
- dumps
- swamps
Applications: Stereo vision

**input:** two images  
**output:** disparity
Key challenges

1. **Represent** the world as a collection of random variables $X_1, \ldots, X_n$ with joint distribution $p(X_1, \ldots, X_n)$
   - How does one *compactly describe* this joint distribution?
   - Directed graphical models (Bayesian networks)
   - Undirected graphical models (Markov random fields, factor graphs)

2. **Learn** the distribution from data
   - Maximum likelihood estimation. Other estimation methods?
   - How much data do we need?
   - How much computation does it take?

3. **Perform “inference”** (compute conditional distributions $p(X_i \mid X_1 = x_1, \ldots, X_m = x_m)$)
We will study Representation, Inference & Learning

First in the simplest case
- Only discrete variables
- Fully observed models
- Exact inference & learning

Then generalize
- Continuous variables
- Partially observed data during learning (hidden variables)
  - Approximate inference & learning

Learn about algorithms, theory & applications
Logistics: class

- **Class webpage:**
  - [http://cs.nyu.edu/~dsontag/courses/pgm13/](http://cs.nyu.edu/~dsontag/courses/pgm13/)
  - Sign up for mailing list!
  - Draft slides posted before each lecture

- **Book:** *Probabilistic Graphical Models: Principles and Techniques* by Daphne Koller and Nir Friedman, MIT Press (2009)
  - Required readings for each lecture posted to course website.
  - Many additional reference materials available!

- **Office hours:** Wednesday 5-6pm and by appointment. 715 Broadway, 12th floor, Room 1204

- **Teaching Assistant:** Li Wan (wanli@cs.nyu.edu)

- **Li’s Office hours:** Monday 5-6pm. 715 Broadway, Room 1231
Prerequisites:
- Previous class on machine learning
- Basic concepts from probability and statistics
- Algorithms (e.g., dynamic programming, graphs, complexity)
- Calculus

Grading: problem sets (65%) + in class final exam (30%) + participation (5%)
- Class attendance is required.
- 7-8 assignments (every 1–2 weeks). Both theory and programming.
- First homework out today, due next Thursday (Feb. 7) at 5pm
- Important: See collaboration policy on class webpage

Important: Solutions to the theoretical questions require formal proofs.

For the programming assignments, I recommend Python, Java, or Matlab. Do not use C++.
An **outcome space** specifies the possible outcomes that we would like to reason about, e.g.

- **Coin toss**
  \[ \Omega = \{ \text{Heads}, \text{Tails} \} \]

- **Die toss**
  \[ \Omega = \{ 1, 2, 3, 4, 5, 6 \} \]

We specify a **probability** \( p(\omega) \) for each outcome \( \omega \) such that

\[
p(\omega) \geq 0, \quad \sum_{\omega \in \Omega} p(\omega) = 1
\]

E.g.,

\[
p(\text{Heads}) = 0.6 \\
p(\text{Tails}) = 0.4
\]
An event is a subset of the outcome space, e.g.

\[ E = \{ \text{dice faces} \} \]

Even die tosses

\[ O = \{ \text{dice faces} \} \]

Odd die tosses

The probability of an event is given by the sum of the probabilities of the outcomes it contains,

\[
p(E) = \sum_{\omega \in E} p(\omega)
\]

E.g., \( p(E) = p(\text{even}) + p(\text{even}) + p(\text{even}) \)

= 1/2, if fair die
Independence of events

Two events $A$ and $B$ are **independent** if

$$p(A \cap B) = p(A)p(B)$$

Are these two events independent?

No!  
$$p(A \cap B) = 0, \quad p(A)p(B) = \left(\frac{1}{6}\right)^2$$

Now suppose our outcome space had two different die:

$$\Omega = \{\text{dice}, \ldots \}$$

2 die tosses

and the probability distribution is such that each die is independent,

$$p(\text{dice}) = p(\text{dice}) p(\text{dice})$$

$$p(\text{dice}) = p(\text{dice}) p(\text{dice})$$
Independence of events

- Two events $A$ and $B$ are **independent** if
  \[
p(A \cap B) = p(A)p(B)
  \]

- Are these two events independent?

![Dice set](image)

\[p(A) = p(\text{dice set})\]

\[p(B) = p(\text{dice set})\]

Yes!

\[p(A \cap B) = p(\text{dice set})\]

\[p(A)p(B) = p(\text{dice set})p(\text{dice set})\]
Conditional probability

- Let $A, B$ be events, $p(B) > 0$.

$$p(A \mid B) = \frac{p(A \cap B)}{p(B)}$$

- Claim 1: $\sum_{\omega \in S} p(\omega \mid S) = 1$
- Claim 2: If $A$ and $B$ are independent, then $p(A \mid B) = p(A)$
Two important rules

1. **Chain rule**  Let $S_1, \ldots S_n$ be events, $p(S_i) > 0$.

$$p(S_1 \cap S_2 \cap \cdots \cap S_n) = p(S_1)p(S_2 \mid S_1) \cdots p(S_n \mid S_1, \ldots, S_{n-1})$$

2. **Bayes’ rule**  Let $S_1, S_2$ be events, $p(S_1) > 0$ and $p(S_2) > 0$.

$$p(S_1 \mid S_2) = \frac{p(S_1 \cap S_2)}{p(S_2)} = \frac{p(S_2 \mid S_1)p(S_1)}{p(S_2)}$$
Discrete random variables

- Often each outcome corresponds to a setting of various attributes (e.g., “age”, “gender”, “hasPneumonia”, “hasDiabetes”)
- A random variable $X$ is a mapping $X : \Omega \rightarrow D$
  - $D$ is some set (e.g., the integers)
  - Induces a partition of all outcomes $\Omega$
- For some $x \in D$, we say

$$p(X = x) = p(\{\omega \in \Omega : X(\omega) = x\})$$

“probability that variable $X$ assumes state $x$”
- Notation: $\text{Val}(X) = \text{set } D \text{ of all values assumed by } X$
  (will interchangeably call these the “values” or “states” of variable $X$)
- $p(X)$ is a distribution: $\sum_{x \in \text{Val}(X)} p(X = x) = 1$
Multivariate distributions

- Instead of one random variable, have random vector
  \[ \mathbf{X}(\omega) = [X_1(\omega), \ldots, X_n(\omega)] \]

- \( X_i = x_i \) is an event. The **joint distribution**
  \[ p(X_1 = x_1, \ldots, X_n = x_n) \]
  is simply defined as \( p(X_1 = x_1 \cap \cdots \cap X_n = x_n) \)

- We will often write \( p(x_1, \ldots, x_n) \) instead of \( p(X_1 = x_1, \ldots, X_n = x_n) \)

- Conditioning, chain rule, Bayes’ rule, etc. **all apply**
Working with random variables

- For example, the **conditional distribution**
  
  \[
p(X_1 \mid X_2 = x_2) = \frac{p(X_1, X_2 = x_2)}{p(X_2 = x_2)}.
  \]

  This notation means
  
  \[
p(X_1 = x_1 \mid X_2 = x_2) = \frac{p(X_1 = x_1, X_2 = x_2)}{p(X_2 = x_2)} \quad \forall x_1 \in \text{Val}(X_1)
  \]

- Two random variables are **independent**, \(X_1 \perp X_2\), if
  
  \[
p(X_1 = x_1, X_2 = x_2) = p(X_1 = x_1)p(X_2 = x_2)
  \]

  for all values \(x_1 \in \text{Val}(X_1)\) and \(x_2 \in \text{Val}(X_2)\).
Consider three binary-valued random variables

\[ X_1, X_2, X_3 \quad \text{Val}(X_i) = \{0, 1\} \]

Let outcome space \( \Omega \) be the cross-product of their states:

\[ \Omega = \text{Val}(X_1) \times \text{Val}(X_2) \times \text{Val}(X_3) \]

\( X_i(\omega) \) is the value for \( X_i \) in the assignment \( \omega \in \Omega \)

Specify \( p(\omega) \) for each outcome \( \omega \in \Omega \) by a big table:

<table>
<thead>
<tr>
<th>( x_1 )</th>
<th>( x_2 )</th>
<th>( x_3 )</th>
<th>( p(x_1, x_2, x_3) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>.11</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>.02</td>
</tr>
<tr>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>.05</td>
</tr>
</tbody>
</table>

How many parameters do we need to specify?

\[ 2^3 - 1 \]
Suppose $X$ and $Y$ are random variables with distribution $p(X, Y)$
\[ X: \text{Intelligence, } \text{Val}(X) = \{ \text{"Very High", "High"} \} \]
\[ Y: \text{Grade, } \text{Val}(Y) = \{ \text{"a", "b"} \} \]

Joint distribution specified by:

\[
\begin{array}{c|cc}
X & \text{vh} & h \\
Y & a & 0.7 & 0.15 \\
& b & 0.1 & 0.05 \\
\end{array}
\]

\[ p(Y = a) = ? = 0.85 \]

More generally, suppose we have a joint distribution $p(X_1, \ldots, X_n)$. Then,
\[
p(X_i = x_i) = \sum_{x_1} \sum_{x_2} \cdots \sum_{x_{i-1}} \sum_{x_{i+1}} \cdots \sum_{x_n} p(x_1, \ldots, x_n)
\]
Suppose $X$ and $Y$ are random variables with distribution $p(X, Y)$

- $X$: Intelligence, $\text{Val}(X) = \{\text{“Very High”}, \text{“High”}\}$
- $Y$: Grade, $\text{Val}(Y) = \{\text{“a”}, \text{“b”}\}$

\[
\begin{array}{c|cc}
X & \text{vh} & h \\
\hline
Y & a & 0.7 & 0.15 \\
   & b & 0.1 & 0.05 \\
\end{array}
\]

Can compute the conditional probability

\[
p(Y = a \mid X = \text{vh}) = \frac{p(Y = a, X = \text{vh})}{p(X = \text{vh})} = \frac{0.7}{0.7 + 0.1} = 0.875.
\]
Example: Medical diagnosis

- Variable for each **symptom** (e.g. “fever”, “cough”, “fast breathing”, “shaking”, “nausea”, “vomiting”)
- Variable for each **disease** (e.g. “pneumonia”, “flu”, “common cold”, “bronchitis”, “tuberculosis”)
- Diagnosis is performed by **inference** in the model:

\[
p(\text{pneumonia} = 1 \mid \text{cough} = 1, \text{fever} = 1, \text{vomiting} = 0)
\]

- One famous model, Quick Medical Reference (QMR-DT), has 600 diseases and 4000 findings
Naively, could represent multivariate distributions with table of probabilities for each outcome (assignment).

How many outcomes are there in QMR-DT? \(2^{4600}\)

Estimation of joint distribution would require a huge amount of data.

Inference of conditional probabilities, e.g.

\[ p(\text{pneumonia} = 1 \mid \text{cough} = 1, \text{fever} = 1, \text{vomiting} = 0) \]

would require summing over exponentially many variables’ values.

Moreover, defeats the purpose of probabilistic modeling, which is to make predictions with previously unseen observations.
Structure through independence

- If $X_1, \ldots, X_n$ are independent, then
  \[
p(x_1, \ldots, x_n) = p(x_1)p(x_2) \cdots p(x_n)
  \]
- $2^n$ entries can be described by just $n$ numbers (if $|\text{Val}(X_i)| = 2$)!
- However, this is not a very *useful* model – observing a variable $X_i$ cannot influence our predictions of $X_j$
- If $X_1, \ldots, X_n$ are *conditionally independent* given $Y$, denoted as $X_i \perp X_{-i} \mid Y$, then
  \[
p(y, x_1, \ldots, x_n) = p(y)p(x_1 \mid y) \prod_{i=2}^{n} p(x_i \mid x_1, \ldots, x_{i-1}, y)
  \]
  \[
  = p(y)p(x_1 \mid y) \prod_{i=2}^{n} p(x_i \mid y).
  \]
- This is a simple, yet *powerful*, model
Example: naive Bayes for classification

- Classify e-mails as spam ($Y = 1$) or not spam ($Y = 0$)
  - Let $1 : n$ index the words in our vocabulary (e.g., English)
  - $X_i = 1$ if word $i$ appears in an e-mail, and 0 otherwise
  - E-mails are drawn according to some distribution $p(Y, X_1, \ldots, X_n)$
- Suppose that the words are conditionally independent given $Y$. Then,
  \[ p(y, x_1, \ldots x_n) = p(y) \prod_{i=1}^{n} p(x_i \mid y) \]

**Estimate** the model with maximum likelihood. **Predict** with:

\[ p(Y = 1 \mid x_1, \ldots x_n) = \frac{p(Y = 1) \prod_{i=1}^{n} p(x_i \mid Y = 1)}{\sum_{y=\{0,1\}} p(Y = y) \prod_{i=1}^{n} p(x_i \mid Y = y)} \]

- Are the independence assumptions made here reasonable?
- Philosophy: Nearly all probabilistic models are “wrong”, but many are nonetheless useful
A **Bayesian network** is specified by a directed *acyclic* graph $G = (V, E)$ with:

1. One node $i \in V$ for each random variable $X_i$
2. One conditional probability distribution (CPD) per node, $p(x_i \mid x_{Pa(i)})$, specifying the variable's probability conditioned on its parents’ values

Corresponds 1-1 with a particular factorization of the joint distribution:

$$p(x_1, \ldots, x_n) = \prod_{i \in V} p(x_i \mid x_{Pa(i)})$$

- Powerful framework for designing *algorithms* to perform probability computations
Consider the following Bayesian network:

![Bayesian network diagram]

- What is its joint distribution?

\[
p(x_1, \ldots, x_n) = \prod_{i \in V} p(x_i \mid x_{\text{Pa}(i)})
\]

\[
p(d, i, g, s, l) = p(d)p(i)p(g \mid i, d)p(s \mid i)p(l \mid g)
\]
More examples

\[ p(x_1, \ldots x_n) = \prod_{i \in V} p(x_i \mid x_{Pa(i)}) \]

Will my car start this morning?

Heckerman et al., Decision-Theoretic Troubleshooting, 1995
More examples

\[ p(x_1, \ldots, x_n) = \prod_{i \in V} p(x_i \mid x_{\text{Pa}(i)}) \]

What is the differential diagnosis?

Fig. 1 The ALARM network representing causal relationships is shown with diagnostic (●), intermediate (○) and measurement (□) nodes. CO: cardiac output, CVP: central venous pressure, LVED volume: left ventricular end-diastolic volume, LV failure: left ventricular failure, MV: minute ventilation, PA Sat: pulmonary artery oxygen saturation, PAP: pulmonary artery pressure, PCWP: pulmonary capillary wedge pressure, Pres: breathing pressure, RR: respiratory rate, TPR: total peripheral resistance, TV: tidal volume

Beinlich et al., The ALARM Monitoring System, 1989
Bayesian networks are *generative models*

Evidence is denoted by shading in a node

Can interpret Bayesian network as a *generative process*. For example, to *generate* an e-mail, we

1. Decide whether it is spam or not spam, by sampling \( y \sim p(Y) \)
2. For each word \( i = 1 \) to \( n \), sample \( x_i \sim p(X_i \mid Y = y) \)
Bayesian network structure implies conditional independencies!

- The joint distribution corresponding to the above BN factors as
  \[ p(d, i, g, s, l) = p(d)p(i)p(g \mid i, d)p(s \mid i)p(l \mid g) \]

- However, by the chain rule, any distribution can be written as
  \[ p(d, i, g, s, l) = p(d)p(i \mid d)p(g \mid i, d)p(s \mid i, d, g)p(l \mid g, d, i, g, s) \]

- Thus, we are assuming the following additional independencies:
  \[ D \perp I, \quad S \perp \{D, G\} \mid I, \quad L \perp \{I, D, S\} \mid G. \quad \text{What else?} \]
Bayesian network structure implies conditional independencies!

- Generalizing the above arguments, we obtain that a variable is independent from its non-descendants given its parents

- **Common parent** – fixing $B$ *decouples* $A$ and $C$

- **Cascade** – knowing $B$ *decouples* $A$ and $C$

- **V-structure** – Knowing $C$ *couples* $A$ and $B$
  - This important phenomenon is called *explaining away* and is what makes Bayesian networks so powerful
A simple justification (for common parent)

We’ll show that \( p(A, C \mid B) = p(A \mid B)p(C \mid B) \) for any distribution \( p(A, B, C) \) that factors according to this graph structure, i.e.

\[
p(A, B, C) = p(B)p(A \mid B)p(C \mid B)
\]

Proof.

\[
p(A, C \mid B) = \frac{p(A, B, C)}{p(B)} = p(A \mid B)p(C \mid B)
\]
D-separation ("directed separated") in Bayesian networks

- Algorithm to calculate whether $X \perp Z \mid Y$ by looking at graph separation
- Look to see if there is **active path** between $X$ and $Z$ when variables $Y$ are observed:

![Diagrams showing active paths between X, Y, and Z.](image)
D-separation ("directed separated") in Bayesian networks

- Algorithm to calculate whether $X \perp Z \mid Y$ by looking at graph separation
- Look to see if there is active path between $X$ and $Z$ when variables $Y$ are observed:
D-separation ("directed separated") in Bayesian networks

- Algorithm to calculate whether $X \perp Z \mid Y$ by looking at graph separation
- Look to see if there is **active path** between $X$ and $Z$ when variables $Y$ are observed:

![Graphical depiction of active paths](image)

- If no such path, then $X$ and $Z$ are **d-separated** with respect to $Y$
- d-separation reduces statistical independencies (hard) to connectivity in graphs (easy)
- Important because it allows us to quickly prune the Bayesian network, finding just the relevant variables for answering a query
D-separation example 1

Graphical Model

David Sontag (NYU)
D-separation example 2
2011 Turing Award was for Bayesian networks

JUDEA PEARL
United States – 2011

CITATION

For fundamental contributions to artificial intelligence through the development of a calculus for probabilistic and causal reasoning.

Judea Pearl created the representational and computational foundation for the processing of information under uncertainty.

He is credited with the invention of Bayesian networks, a mathematical formalism for defining complex probability models, as well as the principal algorithms used for inference in these models. This work not only revolutionized the field of artificial intelligence but also became an important tool for many other branches of engineering and the natural sciences. He later created a mathematical framework for causal inference that has had significant impact in the social sciences.

Judea Pearl was born on September 4, 1936, in Tel Aviv, which was at that time administered under the British Mandate for Palestine. He grew up in Ein Gav, a Biblical town his grandfather went to reestablish in 1924. In 1958, after serving in the Israeli army and joining a Kibbutz, Judea decided to study engineering. He attended the Technion, where he met his wife, Ruth, and received a B.S. degree in Electrical Engineering in 1960. Recalling the Technion faculty members in a 2012 interview in the Technion Magazine, he emphasized the thrill of discovery:
Bayesian networks given by \((G, P)\) where \(P\) is specified as a set of local conditional probability distributions associated with \(G\)’s nodes.

One interpretation of a BN is as a generative model, where variables are sampled in topological order.

Local and global independence properties identifiable via d-separation criteria.

Computing the probability of any assignment is obtained by multiplying CPDs:
- Bayes’ rule is used to compute conditional probabilities.
- Marginalization or inference is often computationally difficult.