CSE 190 – Lecture 7
Data Mining and Predictive Analytics

Graphical Models
Gradesource

- You should have received an e-mail with your gradesource secret number
- Grades will be posted there once your homework has been marked
- Homework will be returned on Thursday (but I’ll cover it in class next week)
So far we’ve looked at prediction problems of the form

\[ p(\text{label}|\text{data}) \]
Today

E.g.
Estimate a user’s political affiliation from the content of their tweets

Train a model to fit:
\[ p(u \text{ leans right of centre} \mid \text{content of } u\text{'s tweets}) \]
But!
Can we do better by using information from the network?

u’s friends/followers

e.g. train a model to fit:

\[ p(u \text{ leans right of centre} \mid \text{content of } u\text{’s tweets, political affiliation of } u\text{’s friends}) \]
Today

But (part 2)!
friends’ affiliations are also unknowns

u’s friends/followers

e.g. train a model to fit:
\[ p(u \text{ leans right of centre} | \text{content of } u\text{'s tweets, political affiliation of } u\text{'s friends}) \]
Today

Interdependent variables

How can we solve predictive tasks when

- There are multiple unknowns to infer simultaneously
- There are dependencies between the unknowns
- In other words, what can we do when...

\[ p(label_1, label_2 | \text{data}) \neq p(label_1 | \text{data})p(label_2 | \text{data}) \]
Examples

Infer the political affiliation of **every** user on twitter

(kind of did this last week, but we didn’t make any use of evidence at each node)

Graph data from Adamic (2004). Visualization from allthingsgraphed.com
What was said in the missing part of the signal? (or, what was the whole signal)

Sollen wir ? ?(garbled)? ? Berlin fahren

Restore the image
The restored value of each pixel is related to (the restored value of) the pixels surrounding it.
Examples

In all of these examples we can’t infer the values of the unknown variables in isolation (or at least not very well)

**Q:** Can we infer all of the variables *simultaneously* and account for their interdependencies?
Infer the political affiliation of every user on twitter

1 billion variables, 2 states per variable = $2^{(10^9)}$ possible outcomes
What was said in the missing part of the signal? (or, what was the **whole** signal)

5 (or so) variables (words), ~10,000 possible values (dictionary size) = $(10^4)^5$ outcomes

**Sollen wir** ? ?(garbled)? ? Berlin fahren
Examples

Restore the image

The restored value of each pixel is related to (the restored value of) the surrounding pixels.

1 million variables (pixels), $256^3$ states per pixel = $(256^3)^{(10^6)}$ possible outcomes
A: State spaces are *way too big* to enumerate

But the problems are incredible *structured*, meaning that full enumeration may be avoidable
Examples

Infer the political affiliation of every user on twitter

My affiliation is only directly related to that of my friends

Graph data from Adamic (2004). Visualization from allthingsgraphed.com
What was said in the missing part of the signal? (or, what was the whole signal)

Sollen wir ? ?(garbled)? ? Berlin fahren

Each word in a sentence is only directly related to a few neighboring words.
Restore the image
The restored value of each pixel is related to (the restored value of) the surrounding pixels.

Each pixel is only directly related to the few pixels nearby.
Graphical models

• Are a language to describe the interdependencies between variables in multi-variable inference problems

• Give rise to a set of algorithms that exploit the structure of these interdependencies to make inference tractable
Today

- Some definitions
- Inference in chain-structured models (e.g. inference for sequence data)
- Inference in trees and networks that are “tree-like”
- Inference in some other useful and non-useful specific cases (maybe)
- Case study (maybe)
Consider a high dimensional probability distribution such as:

\[ p(a, b, c, d, e, f, g) \]

Such an expression can be rewritten as:

\[ p(a)p(b|a)p(c|a, b)p(d|a, b, c)p(e|a, b, c, d)p(f|a, b, c, d, e)p(g|a, b, c, d, e, f) \]

Which is not so useful as it’s still a function of seven variables, for example:

\[ p(a) = \sum_{b, c, d, e, f, g} p(a, b, c, d, e, f, g) \]

is expensive to compute
Probability distributions

But what if a more useful factorization is possible?

\[ p(a, b, c, d, e, f, g) \]

Imagine this can be rewritten as:

\[ p(a)p(b|a)p(c|b)p(d|c)p(e|d)p(f|e)p(g|f) \]

“a causes b, b causes c, c causes d, d causes e...”
e.g. what is the probability that the following forecast is accurate?

\[ p(\text{Sun}=-6 \mid \text{Sat}=-7)p(\text{Mon}=-8 \mid \text{Sun}=-6)p(\text{Tue}=-6 \mid \text{Mon}=-8) \ldots \]
Probability distributions

What is useful about a distribution that factorizes like

\[ p(a)p(b|a)p(c|b)p(d|c)p(e|d)p(f|e)p(g|f) \]

is that we can compute marginals efficiently:

\[
\begin{align*}
p(g) &= \sum_{a,b,c,d,e,f} p(a, b, c, d, e, f, g) \\
&= \sum_{a,b,c,d,e,f} p(a)p(b|a)p(c|b)p(d|c)p(e|d)p(f|e)p(g|f) \\
&= \sum_{f} p(g|f) \sum_{e} p(f|e) \sum_{d} p(e|d) \sum_{c} p(d|c) \sum_{b} p(c|b) \sum_{a} p(a)p(b|a)
\end{align*}
\]
\[ p(a)p(b|a)p(c|b)p(d|c)p(e|d)p(f|e)p(g|f) \]

\[ = \sum_f p(g|f) \sum_e p(f|e) \sum_d p(e|d) \sum_c p(d|c) \sum_b p(c|b) \sum_a p(a)p(b|a) \]

\[ O(N^2) \]

\[ O(N^2) \]

\[ O(N^2) \]

\[ O(N^2) \]

\[ O(N^2) \]

\[ O(N^2) \]

\[ O(N^2) \]

\[ O(N^2) \]

\[ O(N^2) \]

\[ O(N^2) \]

\[ (N = \text{number of possible states per variable}) \]
We had a problem that was **expensive**:  

\[ p(g) = \sum_{a,b,c,d,e,f} p(a, b, c, d, e, f, g) \]

\((O(N^K), \text{ } N = \text{ number of states, } K = \text{ number of variables})\)

but were able to solve it efficiently \((O(KN^2))\) due to factorization:

\[
= \sum_f p(g|f) \sum_e p(f|e) \sum_d p(e|d) \sum_c p(d|c) \sum_b p(c|b) \sum_a p(a)p(b|a) 
\]

(bonus: we computed the marginal of every variable while we were at it!)
Directed graphical models (Bayes Nets)

Graphical models give us a language to describe such factorization assumptions e.g.

\[ p(a, b, c, d, e, f, g) = p(a)p(b|a)p(c|b)p(d|c)p(e|d)p(f|e)p(g|f) \]

Can be described by the graph
Directed graphical models

A few examples....

\[ p(c)p(a|c)p(b|c) \]
\[ p(a)p(b)p(c|a,b) \]
\[ p(a)p(b|a)p(c|b) \]

**Rule:** terms factorize according to \( p(\text{node}|\text{parents}) \)
Directed graphical models

A few examples....

What about:

\[ p(c)p(a|c)p(b|c) \]

What is: \( p(b) \)?

\[
= \sum_{c,a} p(c)p(a|c)p(b|c) \\
= \sum_c p(c)p(b|c) \sum_a p(a|c) \\
= 1
\]

But what if we knew \( a \)?

\[ p(c)p(a = 1|c)p(b|c) \]

\( p(b) \)?

\[
= \sum_c p(c)p(a = 1|c)p(b|c) \\
= f(b, c)
\]
What are the conditional independence statements implied by this graph?

"c is a common cause for a and b"
"if we know c, then knowing a tells us nothing about b"
\[(a \perp \perp b|c)\]
Recall: Naïve Bayes (week 2)

"features are independent given the label"

\[ (\text{feature}_i \perp \text{feature}_j | \text{label}) \]
Conditional independence

What are the conditional independence statements implied by this graph?

(a → b → c)

\[ p(a)p(b|a)p(c|b) \]

(a ⊥⊥ c|b)

e.g. “Monday’s weather is conditionally independent of Wednesday’s weather, given Tuesday’s weather”
Conditional independence

What are the conditional independence statements implied by this graph?

\[ p(a)p(b)p(c|a, b) \]

\[ (a \perp\!\!\!\!\!\!\!\!\perp b|c) ? \]

No: e.g. think of a system with two points of failure. If I know \( c \), then knowing \( \sim a \) tells me that \( b \) is likely.
Conditional independence

What are the conditional independence statements implied by this graph?

\[ p(a)p(b)p(c|a, b) \]

But... \[ p(a, b) = \sum_c p(a)p(b)p(c|a, b) \]

\[ (a \perp\!\!\!\perp b|\emptyset) \]

“\(a\) and \(b\) are conditionally independent if we know nothing”
So... what parts of the graph can we ignore when doing inference?

e.g. if we know $a$, then we can ignore $d,e,f$ when performing inference about $b/c$

**Case 1:**

$(A \perp B|C)$ if

any path from $a \in A$ to $b \in B$

meets at $c$ or

with $c \in C$
D-separation

So... what parts of the graph can we ignore when doing inference?

e.g. if we know a, then we can ignore d,e,f when performing inference about b/c

Case 2: 
\[ (A \perp B \mid C) \] if

any path from \( a \in A \) to \( b \in B \) meets at \( c \) and neither \( c \) nor any of its descendants are in \( C \)
So... what parts of the graph can we ignore when doing inference?

In these two cases we say that \( C \) \textbf{d-separates} (directionally separates) \( A \) from \( B \), and that \( (A \perp\!\!\!\perp B|C) \)

This means that if we know \( C \), then we can ignore \( B \) when making inferences about \( A \)

These cases fully characterize the independence structure of the distribution (Pearl, 1988)
Further reading:

- Bishop Chapter 8
- Coursera course on PGMs: [https://www.coursera.org/course/pgm](https://www.coursera.org/course/pgm)
CSE 190 – Lecture 7
Data Mining and Predictive Analytics

Undirected Graphical Models
Consider the following social network:
(in which friends influence each other’s decisions)

Who will vote the same way?

(see similar examples in slides from Stanford (Koller), Buffalo (Srihari) etc.)
What graphical model represents this?

Want:

\((\text{julian} \perp \text{jake}|\text{ashton}, \text{bob})\)

\((a \perp d|b, c)\)

\((\text{ashton} \perp \text{bob}|\text{julian}, \text{jake})\)

\((b \perp c|a, d)\)

Who will vote the same way?
Undirected graphical models

Attempt 1:

Want:

\[(\text{julian} \perp \text{jake}|\text{ashton}, \text{bob})\]
\[(\text{a} \perp \text{d}|\text{b}, \text{c}) \text{ yes}\]
\[(\text{ashton} \perp \text{bob}|\text{julian}, \text{jake})\]
\[(\text{b} \perp \text{c}|\text{a}, \text{d}) \text{ no (why?)}\]

Who will vote the same way?
Undirected graphical models

Attempt 2:

Want:

\[(julian \perp jake | ashton, bob)\]  
\[(a \perp d | b, c) \quad \text{yes}\]

\[(ashton \perp bob | julian, jake)\]  
\[(b \perp c | a, d) \quad \text{no (why?)}\]

Who will vote the same way?
There **is no** directed network that will capture exactly these conditional independence assumptions

\[
(a \independent d|b, c) \\
(b \independent c|a, d)
\]

So let’s use an undirected network to represent them!
The edges of the network determine how the distribution \textbf{factorizes}

\[
p(a, b, c, d) = \frac{1}{Z} \psi(a, b) \psi(a, c) \psi(b, d) \psi(c, d)
\]

\[
Z = \sum_{a,b,c,d} \psi(a, b) \psi(a, c) \psi(b, d) \psi(c, d)
\]
Undirected graphical models

Examples:

\[
\psi(a, b) \quad \psi(b, c) \quad \psi(c, d)
\]

\[
p(a, b, c, d) = \frac{1}{Z} \psi(a, b) \psi(b, c) \psi(c, d)
\]

Factors are defined over the (maximal) **cliques** of the graph
Undirected graphical models

How to convert from a directed to an undirected network?

e.g.

$$p(a, b, c, d) = p(a)p(b)p(c|a, b)p(d|c)$$
Undirected graphical models

How to convert from a directed to an undirected network?

e.g.

\[ p(a, b, c, d) = p(a)p(b)p(c|a, b)p(d|c) \]
How to convert from a directed to an undirected network?

\begin{align*}
    p(a, b, c, d) &= p(a)p(b)p(c|a, b)p(d|c) \\
    p(a, b, c, d) &= \frac{1}{\mathcal{Z}} \psi(a, b, c)\psi(c, d)
\end{align*}
How to convert from a directed to an undirected network?

every term from the original graph now appears in a clique

\[ p(a, b, c, d) = p(a)p(b)p(c|a, b)p(d|c) \]

\[ p(a, b, c, d) = \frac{1}{Z} \psi(a, b, c)\psi(c, d) \]

**But:** the construction has “forgotten” some information
Undirected graphical models

How to convert from a directed to an undirected network?

Both directed networks transform to the same undirected network

But we lost the fact that \( a \perp b | \emptyset \) in the undirected version
Undirected graphical models

Inference is similar to the directed case:

\[
p(a, b, c, d) = \frac{1}{Z} \psi(a, b) \psi(b, c) \psi(c, d)
\]

\[
p(a) = \sum_{b, c, d} \frac{1}{Z} \psi(a, b) \psi(b, c) \psi(c, d)
\]

\[
= \frac{1}{Z} \sum_b \psi(a, b) \sum_c \psi(b, c) \sum_d \psi(c, d)
\]

Just normalize the result so that it’s a probability distribution.
Another example...

\[
p(a) = \sum_{b,c,d} \frac{1}{Z} \psi(a, b) \psi(b, c) \psi(b, d)
\]

\[
= \frac{1}{Z} \sum_b \psi(a, b) \sum_c \psi(b, c) \sum_d \psi(b, d)
\]

multiple elimination orderings are possible:

\[
= \frac{1}{Z} \sum_b \psi(a, b) \sum_d \psi(b, d) \sum_c \psi(b, c)
\]
Where are we so far?

We have a “language” for describing the dependencies between variables in multi-variable inference problems

- **Directed** graphical models are a natural way to represent causal relationships, e.g. whether I drive to work depends on whether it rains

  ![Diagram](diagram.png)

  - did my front brakes fail?  
  - did I crash my bike?

- **Undirected** graphical models are a natural way to represent mutual dependencies, e.g. friends are likely to share similar opinions
What do we still need to do?

We still need to use these models to perform inference tasks (i.e., to infer marginals and maximum likelihood states)

- So far we’ve done this by moving around summation signs on the blackboard until things worked
- But how can we define algorithms to do this automatically?
- See lecture supplement on course webpage
Questions?

Next lecture:
• Homework 1
• Some recap of midterm-related material, and any pressing questions
• (maybe) get started on recommender systems