COMS 4771
Introduction to Machine Learning

Nakul Verma
Announcements

• Project presentations next class!

• HW4 due next class.

• Final Exam in one week!
• Dimensionality Reduction
  Linear vs non-linear Dimensionality Reduction

• Principal Component Analysis (PCA)

• Introduction to Graphical Models

• Bayesian Networks

• Doing inference and learning on Bayes Nets
A probabilistic model where a graph represents the conditional dependence structure among the variables. Provides a compact representation of the joint distribution!

Example:
Four variables of interest – cloudiness, raining, sprinkler, grass_wet

Possible Causality Structure:

Inference questions:
• What is the probability it rained given the grass is wet?
• What is the chance that the sprinkler was off given grass is wet and it is not cloudy?

Learning questions:
• What is the most likely GM structure and connection weights that models the data?
Graphical Models: Representation

There are two kinds of Graphical Models

Directed models – Bayesian Networks

Undirected models – Markov Random Fields (MRFs)
Bayesian Networks

What is the joint probability for these variables?

\[ P(C, S, R, G) \]

\[ = P(C) P(R|C) P(S|R, C) P(G|S, R, C) \]

\[ = P(C) P(R|C) P(S|C) P(G|S, R) \]

**Chain rule**

due to the (in)dependencies asserted by the parent-child relationships

**In general:**

\[ P(X_1, \ldots, X_d) = \prod_{i=1}^{d} P(X_i \mid \text{parent}(X_i)) \]

That is: a variable is independent of its ancestors given the parents.
Bayesian Networks: Inference

\[ P(C, S, R, G) = P(C)P(R|C)P(S|C)P(G|S, R) \]

These conditional probability tables (CPT) are enough to completely specify the joint distribution!
Bayesian Networks: Inference

\[ P(C, S, R, G) = P(C)P(R|C)P(S|C)P(G|S, R) \]

Q: What is the probability of sprinkler being on given the grass is wet?

\[ P(S = 1|G = 1) = \frac{P(S = 1, G = 1)}{P(G = 1)} = \frac{0.2781}{0.6471} = 0.430 \]

\[ P(G = 1) = \sum_{c,s,r} P(C = c, S = s, R = r, G = 1) \]

\[ = \sum_{c,s,r} P(C = c)P(R = r|C = c)P(S = s|C = c)P(G = 1|S = s, R = r) \]

\[ = 0.6471 \]

\[ P(S = 1, G = 1) = \sum_{c,r} P(C = c, S = 1, R = r, G = 1) = \cdots = 0.2781 \]
Bayesian Networks: Learning Parameters

\[ P(C, S, R, G) = P(C) P(R | C) P(S | C) P(G | S, R) \]

Learning the parameters knowing the structure

\textit{ie, estimate the CPTs from observations}

Simply do the likelihood estimates (ie, counts)

\[ \hat{P}_{ML}(G = g | S = s, R = r) = \frac{\#(G = g, S = s, R = r)}{\#(S = s, R = r)} \]

etc ...

\textit{Issue: assigns zero prob. for unseen combinations in data.}
\textit{How to fix that?}
Bayesian Networks: Learning Structure

\[
P(C, S, R, G) = P(C)P(R|C)P(S|R, C)P(G|S, R, C)
\]

Learning the unknown structure between the variables

General

- Test of conditional independencies in data
- Grow-Shrink Markov Blanket algorithm

Assumed structure:

- Tree structure: Chow-Liu algorithm
- Small cliques: variations on Chow-Liu

NP-hard to find the optimal structure
Markov Random Fields (MRFs)

Graphical models with undirected connections

\[ P(X_1, \ldots, X_d) = \frac{1}{Z} \prod_{C \in \text{max-cliques}(G)} \phi_C(X_C) \]

normalizer (so things integrate to 1), aka the partition function

Clique potentials, typically the relative frequency of variable co-occurrence in a clique

Example: five variable graph

What are the max-cliques?

\[ P(X_1, \ldots, X_5) \propto \phi_1((X_1, X_2, X_3)) \phi_2((X_2, X_3, X_4)) \phi_3((X_3, X_5)) \]
What are the (conditional) independencies asserted by the following graphical models?

*(directed)*

![Directed Graphs](image)

*(undirected)*

![Undirected Graphs](image)
Relation Between Directed & Undirected GM

What are the (conditional) independencies asserted by the following directed model?

What is the equivalent undirected model?
A time series model:

A family of distributions over a sequence of random variables $X_1, X_2, \ldots$ that is indexed by a totally ordered indexing set (often referred to as time).

Many applications:

- Financial/Economic data over time
- Climate data
- Speech and natural language
- ...
Markov Model:

A time series model with the property:

The conditional distribution of the next state $X_{t+1}$ given all the previous states $X_i$ ($i \leq t$) only depends on the current state $X_t$

$$P(X_{t+1} \mid X_t, X_{t-1}, X_{t-2}, \ldots) = P(X_{t+1} \mid X_t)$$

The corresponding graphical model:

... → $X_{t-2}$ → $X_{t-1}$ → $X_t$ → $X_{t+1}$ → $X_{t+2}$ → ...

also known as a Markov chain
To specify a Markov Chain:

Need to specify the distribution of the initial state: \( X_1 \)

Need to specify the conditional distribution: \( X_{t+1} \) given \( X_t \)

\( \text{(We will focus on finite size state space, say, } d \text{ different states) } \)

Initial state distribution:

\[
P(X_1 = i) = \pi_i
\]

Conditional distribution:

\[
P(X_{t+1} = j \mid X_t = i) = A_{i,j}
\]

This is often called the transition matrix

\( A \) is row stochastic
Markov Chain: Example

State space: \{1,2\}

Parameters:

\[ \pi = \begin{pmatrix} 0.1 \\ 0.9 \end{pmatrix}, \quad A = \begin{pmatrix} 0.3 & 0.7 \\ 0.6 & 0.4 \end{pmatrix} \]

What is the probability of seeing the random sequence: 2,2,2,1,1,2,2,1?

\[ \pi_2 \cdot A_{2,2} \cdot A_{2,2} \cdot A_{2,1} \cdot A_{1,1} \cdot A_{1,2} \cdot A_{2,2} \cdot A_{2,1} \approx 0.004355 \]
Markov Chain: Example - PageRank

Web graph: vertices – webpages, edges – links between webpages

Question: how popular is a given webpage $i$?

Possible answer:

proportional to the probability that a random walk ends on page $i$.

$$P(X_t = i) \quad (\text{for some large } t)$$
Let’s calculate the following probabilities:

\[ P(X_1 = i) = \pi_i \]

\[ P(X_2 = i) = \sum_j P(X_1 = j, X_2 = i) \]

\[ = \sum_j P(X_1 = j) \cdot P(X_2 = i \mid X_1 = j) \]

\[ = \sum_j \pi_j A_{j,i} \]

\[ = i^{th} \text{ entry of } \pi^T A = (\pi^T A)_i \]

\[ P(X_3 = i) = \ldots = (\pi^T AA)_i \]

\[ P(X_t = i) = (\pi^T A^{t-1})_i \]

for the PageRank example, does this converge to a stable value for large \( t \)?
Markov Chain: Limiting Behavior

Question does/can $P(X_t) = i$ have a limiting behavior?

Equivalent to asking:

$$\lim_{t \to \infty} A^t \text{ approach a limiting matrix } \begin{pmatrix} \cdots \ q \ \cdots \\ \cdots \ q \ \cdots \\ \cdots \ q \ \cdots \end{pmatrix} \quad \text{(with identical rows)} \ ?$$

For such an $A$, it must satisfy:

$$\lim_{t \to \infty} A^t = \left( \lim_{t \to \infty} A^{t-1} \right) A = \begin{pmatrix} \cdots \ q \ \cdots \\ \cdots \ q \ \cdots \\ \cdots \ q \ \cdots \end{pmatrix} A = \begin{pmatrix} \cdots \ q \ \cdots \\ \cdots \ q \ \cdots \\ \cdots \ q \ \cdots \end{pmatrix}$$

Equivalently:

$$q A = q$$

ie, $q$ is the left eigenvector of $A$ with eigenvalue 1!

$q$ unique whenever there is no multiplicity of eigenvalue 1

such a $q$ is called the stationary distribution of $A$
Web graph doesn’t have a unique stationary distribution, but can add some regularity to the link matrix $A$. That is $\tilde{A} = A + \epsilon 1$

Popularity of a given webpage $i$ is proportional to the $i^{th}$ component of the (regularized) stationary distribution
Hidden Markov Model (HMM): A Markov chain on $\{(X_t, Y_t)\}_t$

Some properties:
- $Y_t$ is unobserved / hidden variable; only $X_t$ is observed.
- Conditioned on $Y_t$, $X_t$ is independent of all other variables!

The corresponding graphical model:
Hidden Markov Models (HMMs) Applications

Natural Language Processing
- Observed: words in a sentence
- Unobserved: words’ part-of-speech or other word semantics

Bioinformatics
- Observed: Amino acids in a protein
- Unobserved: indicators of evolutionary conservation

Speech Recognition
- Observed: Recorded speech
- Unobserved: The phonemes the speaker intended to vocalize
HHMs Parameters

We will focus on discrete state space:

$X_t$ takes values $\{1, \ldots, D\}$ (observed)

$Y_t$ takes values $\{1, \ldots, K\}$ (hidden)

We need the initial state distribution on $Y_1$

$$P(Y_1 = i) = \pi_i$$

Need to specify a $K \times K$ transition matrix $A$ from $Y_t$ to $Y_{t+1}$

$$P(Y_{t+1} = j \mid Y_t = i) = A_{i,j}$$

Need to specify a $K \times D$ emission matrix $B$ from $Y_t$ to $X_t$

$$P(X_t = j \mid Y_t = i) = B_{i,j}$$

Both $A$ and $B$ are row stochastic
HHM: Example – Dishonest Casino

Casino die-rolling game:
Randomly switch between two possible dice: one is fair and one is loaded.

HMM Parameters

\[ A = \begin{pmatrix} 0.95 & 0.05 \\ 0.10 & 0.90 \end{pmatrix}, \quad B = \begin{pmatrix} \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \\ \frac{1}{10} & \frac{1}{10} & \frac{1}{10} & \frac{1}{10} & \frac{1}{10} & \frac{1}{2} \end{pmatrix}, \]

\[ \pi = (1,0) \] [the casino starts off with the fair die]

Problem: based on the sequence of rolls, guess which die was used at each time
Conditional Probabilities (filtering/smoothing)
• Given: parameters $\theta = (\pi, A, B)$, and the observation $X_{1:T}$
• Goal: What is the conditional probability of $Y_{1:T}$?

$$P(Y_{1:T} \mid X_{1:T}, \theta)$$

Most probable sequence (decoding)

$$\arg\max_{Y_{1:T}} P(Y_{1:T} \mid X_{1:T}, \theta)$$

Parameter Estimation
• Given: The observations $X_{1:T}$
• Goal: Find the best parameter estimate of $\theta$
HHM: Example – Dishonest Casino

- Conditional probability
- Decoding
**Filtering Problem**

Can directly compute $P(Y_{1:T} \mid X_{1:T}, \theta)$ using the standard way, but that is slow and doesn’t exploit the conditional independency structure of HMMs.

A popular fast algorithm:

**Forward-Backward algorithm**, can be done in two passes (one forward pass, one backward pass) over the states.

**Decoding Problem**

Most likely posterior setting of the hidden states can be computed efficiently using a dynamic programming algorithm, called **Viterbi decoding algorithm**.

See supplementary material for detail on these algorithms
HHM: Learning the Parameters

We can use the **Expectation Maximization (EM) Algorithm**!

**Input:** $n$ observations sequences $x^{(1)}_{1:T}, x^{(2)}_{1:T}, \ldots, x^{(n)}_{1:T}$

**Initialize:**
Start with an initial setting / guess of parameters $(\hat{\pi}, \hat{A}, \hat{B})$

**E-step:**
Compute conditional expectation $Y$ given $X$ and current parameter guess

*(this can be done using the Forward-Backward algorithm)*

**M-step:**
Given the estimate of $Y$ and the observations $X$, we have the complete likelihood, so simply maximize the likelihood by taking the derivative and examine the stationary points.

*See supplementary material for details*
What We Learned...

• Graphical Models
  Bayesian Networks and Markov Random Fields

• Doing inference and learning on graphical models

• Markov Models

• Hidden Markov Models

• Bayesian Networks
Questions?
Next time...

Project Presentations!
Announcements

• Project presentations next class!

• HW4 due next class.

• Final Exam in one week!