Introduction to Graphical Model

November 12, 2018

1 Lecture 1

Basic Rules of Probability

Frequently we will say $p(x) \propto f(x)$ for some non-negative function $f(x)$

Then we can conclude that:

$$p(x) = \frac{f(x)}{\sum_y f(y)}$$

For joint distribution,

$$\sum_x p(x) = \sum_x \sum_y p(x, y) = 1$$

Independence

If $p(x|y) = p(x)$ for all states of x and y, then the variables x and y are said to be independent as $x \perp\!\!\!\perp y$.

If x and y are independent, then x and y are uncorrelated. However, in general, x and y are uncorrelated, then cannot conclude that x and y are independent.

Conditional Independence

$$\mathcal{X} \perp\!\!\!\perp \mathcal{Y}|\mathcal{Z}$$

$$p(\mathcal{X}, \mathcal{Y}|\mathcal{Z}) = p(\mathcal{X}|\mathcal{Z})p(\mathcal{Y}|\mathcal{Z})$$

and

$$p(\mathcal{X}|\mathcal{Y}, \mathcal{Z}) = p(\mathcal{X}|\mathcal{Z})$$

Conditional independence does not imply marginal independence:

$$p(x,y) = \sum_z p(x|z)p(y|z)p(z) \neq \sum_z p(x|z)p(z) \sum_z p(y|z)p(z)$$

2 Lecture 2

Graphs

Definition:

A graph consists of nodes (vertixes) and undirected or directed links (edges) between nodes.

Path:

A path from X_i to X_j is a sequence of connected nodes starting at X_i and ending at X_j. (no direction)

Directed Acyclic Graph:

Graph in which by following the direction of the arrows a node will **never** be visited **more than once**.

Parents and Children:

Xi is a parent of X_j if there is a link from X_i to X_j. Xi is a child of X_j if there is a link from X_j to X_i.

Ancestors and Descendants:

The ancestors of a node X_i are the nodes with a directed path ending at X_i. The descendants of X_i are the nodes with a directed path beginning at X_i.

Undirected Graph:

Clique:

A clique is a fully connected subset of nodes.

Maximal Clique:

Clique which is not a subset of a larger clique.

Connected graph:

There is a path between every pair of vertices.

Connected components:

In a non-connected graph, the connected components are the connected-subgraphs.

Connectedness: Singly-connected

There is only one path from any node α to another node b

Multiply-connected

A graph is multiply-connected if it is not singly connected.

Belief Networks (Bayesian Networks)

A belief network is a **directed acyclic graph** in which each node is associated with the conditional probability of the node given its parents.

Belief networks are a convenient framework for representing such independence assumptions.

A belief network is a distribution of the form

$$p(x_1, \ldots, x_D) = \prod_{i=1}^{D} p(x_i|pa(x_i))$$

Processing the network

Firstly write the whole joint distribution such as:

$$p(A, R, E, B) = p(A|R, E, B)p(R|E, B)p(E|B)p(B)$$

Then, according to the assumption, remove some independent variable from the joint distribution. **It does matter that the order of joint distribution influence the processing**.

Uncertain Evidence

In soft/uncertain evidence the variable is in more than one state, with the strength of our belief about each state being given by probabilities. For example, if y has the states $dom(y) = \{\text{red, blue, green}\}$ the vector $(0.6, 0.1, 0.3)$ could represent the probabilities of the respective states

In the calculation, we can do this: Given $P(A = tr) = 0.7$

$$p(B = tr|\widetilde{A}) = \sum_{A} p(B = tr|A)p(A|\widetilde{A})$$

Independence

If C has more than one incoming link, then $A \perp\!\!\!\perp B$ and A is not conditional independent with B under C condition. In this case C is called collider. If C has at most one incoming link, then $A \perp\!\!\!\perp B|C$ and A is not independent with B. In this case C is called non-collider.

d-connected/separated

X and Y are d-connected by Z if there is any path from X to Y that is not blocked by Z

If all of the paths are blocked then we say X and Y are d-separated by Z.

Markov Equivalence

skeleton

Formed from a graph by removing the arrows.

immorality

An immorality in a DAG is a configuration of three nodes, A,B,C such that C is a child of both A and B, with A and B not directly connected.

Markov Equivalence

Two graphs represent the same set of independence assumptions if and only if they have the same skeleton and the same set of immoralities.

BN representation

BN cannot represent whatever independence statements are present in p.

Fundamentally, the actual **numerical distribution p contains much more information than a graph can represent**.

Markov Network

A Markov Network is an undirected graph in which there is a potential (non-negative function) ψ defined on each maximal clique.

The joint distribution is proportional to the product of all clique potentials:

$$\begin{aligned} P(Y) &= \frac{1}{Z}\prod_{C}\Psi_C(Y_C) \\ z &= \sum_{Y}\prod_{C}\Psi_C(Y_C) \end{aligned}$$

where the $\Psi_C(Y_C)$ is a potential function (strict positive). The potential function would usually indicated as

$$\Psi_C(Y_C) = \exp\{-E(Y_C)\}$$

For the image recovery, the probability of pixel $Pr(x)$ is only related to the sum of neighboring points and the y.

Iterated conditional Modes

To find the minimum of a function $f(x_1, \ldots, x_D)$, firstly make a guess $x_1^*, \ldots, x_D^*$. Then look at the single variable x_i keeping all others fixed. Then set $x_i^* = \text{argmin}_{x_i} F(x_i)$. Then repeat this, sweeping through all variables, usually in a random order.Then repeat the whole process until convergence.

Boltzmann machine

Ising model

Given $x_i \in +1, -1$, then the joint distribution would be:

$$\begin{aligned} p(x_1, \ldots, x_9) &= \frac{1}{Z}\prod_{i\sim j}\phi_{ij}(x_i, x_j) \\ \phi_{ij}(x_i, x_j) &= e^{-\frac{1}{2T}(x_i - x_j)^2} \end{aligned}$$

Rule for independence in Markov Networks

Remove all links neighboring the variables in the conditioning set Z.

If there is no path from any member of x to any member of y,then x and y are conditionally independent given z.

Chain Graphical Model

Chain Graphs (CGs) contain both directed and undirected links. To develop the intuition. The only terms that we can unambiguously specify from this depiction are $p(a)$ and $p(b)$ since there is no mixed interaction of directed and undirected edges at the a and b vertices.

Factor Graphs

A square node represents a factor (non negative function) of its neighboring variables. FGs are most commonly used for inference.

$$f(x_1, \ldots, x_n) = \prod_i \psi_i(x_i)$$

When used to represent a distribution

$$p(x_1, \ldots, x_n) = \frac{1}{Z} \prod_i \psi_i(x_i)$$

For a factor $\psi i(X_i)$ which is a conditional distribution $p(x_i|pa(x_i))$.

Markov Models

Time-Series

A time-series is an ordered sequence:

$$x_{a:b} = \{x_a, x_{a+1}, \ldots, x_b\}$$

For the time series data, we need a model $p(v_{1:T})$. For the causal consistency, it is meaningful to consider the decomposition.

$$p(v_{1:T}) = \prod_{t=1}^{T} p(v_t|v_{1:t-1})$$

with the convention $p(v_t|v_{1:t-1}) = p(v_1)$ for $t = 1$.

Only the recent past is relevant:

$$p(v_t|v_1, \ldots, v_{t-1}) = p(v_t|v_{t-L}, \ldots, v_{t-1})$$

where $L \geq 1$ is the order of the Markov chain.

$$p(v_{1:T}) = p(v_1)p(v_2|v_1)p(v_3|v_2) \ldots p(v_T|v_{T-1})$$

For a stationary Markov chain the transitions $p(v_t = s'|v_{t-1} = s) = f'(s', s)$ are time-independent. $p(v_t|v_{t-1})$ is the transition matrix.

Equilibrium distribution

$$\mathbf{p}_t = \mathbf{M}^{t-1}\mathbf{p}_1$$

where $t \to \infty$ is independent of initial distribution $\mathbf{p}_1$, then $\mathbf{p}_\infty$ is called the equilibrium distribution of the chain:

$$\mathbf{p}_\infty = \mathbf{M}\mathbf{p}_\infty$$

Due to the Markov convergence (transition matrix not change, finite states, transit to any states, not a simple loop), the matrix eigenvalue has 1 value at most. Therefore, it would converge.

Mixture of Markov models

$$p(v_{1:T}) = \sum_{h=1}^{H} p(h)p(v_{1:T}|h) = \sum_{h=1}^{H} p(h) \prod_{t=1}^{T} p(v_t|v_{t-1}, h)$$

Applying EM later.

Hidden Markov Models

$$p(h_{1:t}, v_{1:T}) = p(v_1|h_1)p(h_1) \prod_{t=2}^{T} p(v_t|h_t)p(h_t|h_{t-1})$$

where $p(h_t|h_{t-1})$ and $p(v_t|h_t)$ are constant through time.

Inference

Inference corresponds to using the distribution to answer question about environment.

Computational Efficiency

Inference can be computationally very expensive and we wish to characterize situation in which inferences can be computed efficiently.

For **singly-connected** graphical models, and certain inference questions, there (usually) exist **efficient algorithms** based on the concept of **message passing**.

In general, the case of **multiply-connected** models is computationally **inefficient**.

Sum-Product algorithm

Both Markov and belief networks can be represented using factor graphs (convenient to derive a marginal inference).
This termed the sum-product algorithm since to compute marginals we need to distribute the **sum over variable states over** the **product of factors**. This also refereed to as belief propagation. Message Passing for multiply-connected graph Using cut-set conditioning.

HMM

Forward algorithm

A matrix ($p(h_t = j|h_{t-1} = i) = A_{ji}$)is the transition matrix, B matrix($p(v_t = j|h_t = i) = B_{ji}$) is the emission matrix, $\pi = (\pi_i)$ is the initial state distribution. The hidden Markov model could be defined as $\lambda = (A, B, \pi)$.

$$\begin{aligned} \alpha_t(i) &= P(0_1, o_2, \ldots, o_t, i_t = q_i|\lambda) \\ &= [\sum_{j=1}^{N} \alpha_{t-1}(j)a_{ji}]b_i(o_{t+1}) \\ P(O|\lambda) &= \sum_{i=1}^{N} \alpha_T(i) \end{aligned}$$

Backward algorithm

$$\begin{aligned} \beta_t(i) &= P(o_{t+1}, o_{t+2}, \ldots, o_T|i_t = q_i, \lambda) \\ &= \sum_{j=1}^{N} a_{ij}b_j(o_{t+1})\beta_{t+1}(j) \\ P(O|\lambda) &= \sum_{i=1}^{N} \pi_i b_i(o_1)\beta_1(i) \end{aligned}$$

Here, define $\gamma_t(i) = P(i_t = q_i|O, \lambda) \propto \alpha_t(i)\beta_t(i)$

Baum-Welch algorithm

Viterbi algorithm

3 Conditional Random Field

4 Markov Random Field

A MRF is de ned by a set of distributions $p(x_i|ne(x_i))$ where $i \in 1, \ldots, n$ indexes the distributions and $ne(x_i)$ are the neighbours of variable x_i, namely that subset of the variables $x1, \ldots, xn$ that the distribution of variable x_i depends on. The term Markov indicates that this is a proper subset of the variables. A distribution is an MRF with respect to an undirected graph G if

$$p(x_i|x_{/i}) = p(x_i|ne(x_i))$$

where $ne(x_i)$ are the neighbouring variables of variable x_i, according to the undirected graph G. The notation $x_{/i}$ is shorthand for the set of all variables X excluding variable x_i, namely X/x_i in set notation.

5 Vision

$$E(\mathbf{x}) = \sum_{i \in V} \psi_i(x_i) + \sum_{i \in V, j \in N_i} \psi_{ij}(x_i, x_j)$$

For the Data term:

$$\psi_i(x_i = 0) = -\log(p(x_i \notin FG))$$
$$\psi_i(x_i = 1) = -\log(p(x_i \in FG))$$

For the Smoothness term:

$$\begin{aligned} \psi_i j(x_i, x_j) &= K_{ij} \delta(x_i \neq x_j) \\ where K_{ij} &= \lambda_1 + \lambda_2 \exp(-\beta(I_i - I_j)^2) \end{aligned}$$

Max-Flow Problem

Ford&Fulkerson algorithm Min-cut:$\min_{S,T} \sum_{i \in S, j \in T} c_{ij}$

Efficient Inference in Trees

Message passing (belief propagation.)

variable elimination

Each time sum over the states of a variable we eliminate it from the distribution.
The message pass could be regarded as message passing.
In general, one may view variable elimination as the passing of messages in the form of potentials from nodes to their neighbors. For belief networks variable elimination passes messages that are distributions when following the direction of the edge, and non normalized potentials when passing messages against the direction of the edge

Junction Tree Algorithm

Clique Graphs

A clique graph consists of a set of potentials, $\phi_1(x^1), \ldots, \phi_n(x^n)$ each defined on a set of variables x^i. For neighboring cliques on the graph, defined on sets of variables x^i and x^j, the intersection $x^s = x^i \cap x^j$ is called the separator and has a corresponding potential $\phi_s(x^s)$. A clique graph represents the function.

$$\frac{\prod_c \phi_c(x^c)}{\prod_s \phi_s(x^s)}$$

absorption

Let v and w be neighbors in a clique graph, let S be their separator, and let $\phi(V), \phi(W)$ and $\phi(S)$ be their potentials. Absorption from V to W through S replaces the tables $\phi(S)$ and $\phi(W)$ with

$$\phi^*(S) = \sum_{v/S} \phi(V)$$

$$\phi^*(W) = \phi(W)\frac{\phi^*(S)}{\phi(S)}$$

Then we say that clique W absorbs information from clique V.

Absorption Schedule

A clique can send a message to a neighbor, provided it has already received messages from all other neighbors.

Junction Trees

A clique tree is a junction tree if, for each pair of nodes, V and W, all nodes on the path between V and W contain the intersection $V \cap W$. This is also called the running intersection property.

Moralisation

For each variable x add an undirected link between all parents of x and replace the directed link to x from its parents by undirected links. This creates a ‘moralised‘Markov network.