Probabilistic Graphical Models

Huynh Xuan Phung - Coursera

Contents

Chapter 1

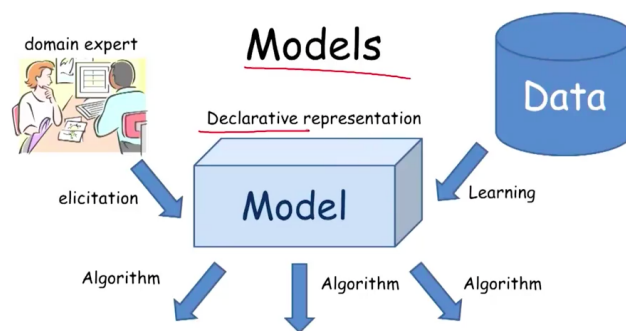# Representation

1.1 Introduction

Model

Figure 1.1: Model is a declarative representation of our understanding of the world

It is important because the same representation, that same model can be used in the the context of one algorithms that might answer different kind of questions. Or the same question in more efficient way.

We can construct methodologies the elicit these models from a human extrict, or learn from data or combination.

Uncertainty
— Partial knowledges of state of the world
— Noisy observations
— Phenomena not covered by our model
— Inherent stochasticity

Probability Theory
— Declarative representation with clear semantics
— Powerful reasoning patterns: conditioning decision making

— Established learning methods

Complex Systems

Graphical Models: Bayesian Networks, Markov Networks (directed or undirected graphs)

Graphical Representation

— Intuitive and compact data structure

— Efficient reasoning using general purpose algorithms

— Sparse parameterization

—— feasible elicitation : by hand

—— learning from data automatically

1.1.1   Distribution:Marginalization,Conditioning

Joint Distribution: P(I,D,G)

Conditioning: observation 1 value of variable —> Reduction —> Renormalization $P(I, D, g^1)$ —> $P(I, D|g^1)$

Marginalization: $\sum_I P(I, D) = P(D)$. Example: you have thrown two 6-sided dice, D_1 and D_2. $P(D_1, D_2)$ is a joint probability distribution. The probability that $D_2 = 1$ is equals to $\sum_{i=1}^{6} P(D_1 = i, D_2 = 1)$.

1.1.2   Factors

A factor is a function or table $\phi(X_1, ..., X_k)$

$\phi : Val(X_1, ..., X_k) \rightarrow R$

Scope $= X_1, ..., X_k$

Joint distribution is a factor

Unnormalized measure is a factor

Conditional Probability Distribution (CPD) is a factor P(G—I,D): G in columns while I,D in rows.

Why factors?

— Fundamental building block for defining distributions in high-dimensional spaces

— Set of basic operations for manipulating these probability distributions

1.2   Bayesian Network (Directed Models)

1.2.1   Semantics and Factorization

What does random variable depend on?

Draw nodes, edges, each node with a factor is CPD (conditional probability distribution)

A bayesian network is:

— A directed acyclic graph (DAG) G whose nodes represent the random variables

— For each node X_i a CPD $P(X_i|Par_G(X_i))$

The BN represent a joint distribution via the chain rule for Bayesian Networks

$P(X_1, ..., X_n) = \prod_i P(X_i | Par_G(X_i))$

BN is a Legal distribution : $\sum P = 1$ and $P > 0$

1.2.2 Reasoning Pattern

Causal Reasoning: reasoning going down

Evidential Reasoning: reasoning going up

Inter-causal Reasoning: The probability of class is hard, if we observe the "C" grade and change the posterior probability of high intelligence is goes up

1.2.3 Flow of Probabilistic Influence: active trail

When can X influence Y?

X → Y

X ← Y

Active Trails:

A trail $X_1 - ... - X_k$ is active if: it has no v-structures $X_{i-1} \to X_i \leftarrow X_{i+1}$

When can X influence Y given evidence about Z?

A trail $X_1 — ... — X_k$ is active given Z if:

— for any v-structure $X_{i-1} \to X_i \leftarrow X_{i+1}$ we have that X_i or one of its descendants $\in$ Z

— no other X_i is in Z

Assignment

1. How many parameter to present a CPD. If X have m possibilities the P(X) needs m-1 independent parameters

IF Y have k possibilities, Z have l possibilities then P(X—Y,Z) has (m-1)*k*l independent parameters.

2. Inter-causal reasoning

To calculate the required values, we can apply Bayes' rule. For instance,

$\frac{P(A=1|T=1,P=1)=P(A=1,T=1,P=1)}{P(T=1,P=1)}$

$= \frac{P(A=1,T=1,P=1)}{(P(A=0,T=1,P=1)+P(A=1,T=1,P=1))}.$

We can then use the chain rule of Bayesian networks to substitute the correct values in, e.g.,

$P(A = 1, T = 1, P = 1) = P(P = 1) * P(A = 1) * P(T = 1 | P = 1, A = 1)$

This example of inter-causal reasoning meshes well with common sense: if we see a traffic jam, the probability that there was a car accident is relatively high. However, if we also see that the president is visiting town, we can reason that the president's visit is the cause of the traffic jam; the probability that there was a car accident therefore drops correspondingly.
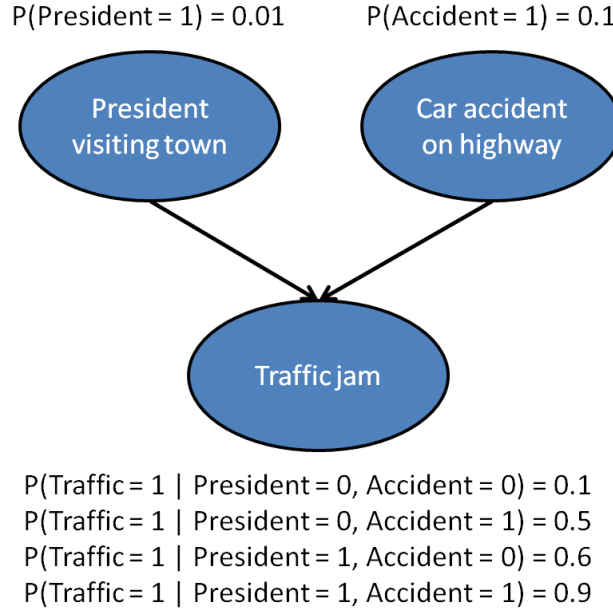
P(President = 1) = 0.01   P(Accident = 1) = 0.1

President
visiting town

Car accident
on highway

Traffic jam

P(Traffic = 1 | President = 0, Accident = 0) = 0.1
P(Traffic = 1 | President = 0, Accident = 1) = 0.5
P(Traffic = 1 | President = 1, Accident = 0) = 0.6
P(Traffic = 1 | President = 1, Accident = 1) = 0.9

Figure 1.2:

1.3   Bayesian Networks: In-dependencies

1.3.1   Conditional Independence

Independence

For events α, β, P satisfy independence if:

— $P(\alpha, \beta) = P(\alpha) * P(\beta)$

— $P(\alpha|\beta) = P(\alpha)$

— $P(\beta|\alpha) = P(\beta)$

Conditional Independence

For random variables X,Y,Z; P satisfy (X independence Y — Z) if

— $P(X, Y|Z) = P(X|Z) * P(Y|Z)$

— $P(X|Y, Z) = P(X|Z)$

— $P(Y|X, Z) = P(Y|Z)$

— $P(X, Y, Z) \propto \phi_1(X, Z) * \phi_2(Y, Z)$

We need to check for all cases

S and G are dependent but conditionally independent given I. I and D are independent but conditionally dependent given G, which active the V-structure here
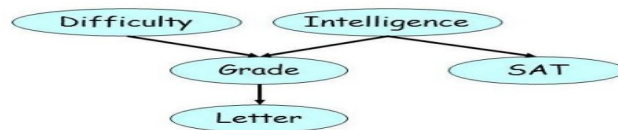
Figure 1.3: Example of student: independence

1.3.2 Independence in Bayesian Networks: d-separate, I-map

Flow of influence and d-separation

X and Y are d-separated in G given Z if there is no active trail in G between X and Y given Z. $d - sep_G(X, Y|Z)$

Factorization to Independences: BNs

Theorem If P factorizes over G, and $d - sep_G(X, Y|Z)$ then P satisfies $(X \perp Y|Z)$.

$P(S) = \sum_I P(I)P(S|I)$ (standard marginalization operation) because P(S,I) = P(I) * P(S—I)

If P factorizes over G, then in P, any variable is independent of its non-descendants given its parents

I-maps

d-separation in G $\rightarrow$ P satisfies corresponding independence statement
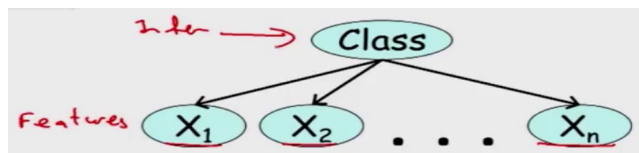
$I(G) = (X \perp Y|Z) : d - sep_G(X, Y|Z)$

Definition: If P satisfies I(G), we say that G is an I-map (in-dependency map) of P

Theorem: If P factorizes over G, then G is an I-map for P

Theorem: If G is an I-map for P, then P factorizes over G

1.3.3 Naive Bayes Model

Assumption: Given the class variable, each observed variable is independent of the other observed variables.

Figure 1.4: Naive Bayes: $(X_i \perp X_j|C)$ for all X_i, X_j

Naive Bayes Classifier

Bernoulli Naive Bayes: each observed variable is binary: 1 if appear and 0 otherwise

Multinomial Naive Bayes for Text: values of each random variable is the actual word in each document

1.3.4 Assignment

If there is no active trails between A and D, then they are independent

I-maps: if G is an I-map of P, all in-dependencies in G also in P. However, this does not mean that all in-dependencies in P also in G

I-maps can also be defined directly on graphs as follow: Let I(G) be the set of in-dependencies encoded by a graph G. Then G_1 is and I-map for G_2 if $I(G_1) \subseteq I(G_2)$

I-map is not a function

1.4 Bayesian Networks: Knowledge Engineering

1.4.1 Medical Diagnosis

1.4.2 Knowledge Engineering: Example SAMIAM

1.4.3 Programming Assignment 1

Compute $P(A, B|C = C_i)$ means that observes the specific value of C and $P(C_i) = 1$ otherwise is 0

— Compute P(A,B,C)

— if $C_j \neq C_i$ then $P(A, B, C_j) = 0$

— $M = \sum_k P(A, B, C_k)$

— Normalize M