BITS Pilani
Pilani | Dubai | Goa | Hyderabad

Probabilistic Graphical Model

Session # 2 : Mathematical Preliminaries

Seetha Parameswaran
seetha.p@pilani.bits-pilani.ac.in

The instructor is gratefully acknowledging the authors who made their course materials freely available online.

Table of Contents

Graph

- Data structure used to represent the probability distribution of data.
- A graph is a data structure $\mathcal{K}$ consisting of a set of nodes and a set of edges.

$$\text{Graph} \qquad \mathcal{K} = (\mathcal{X}, \mathcal{E}) \tag{1}$$

- The set of nodes denote each random variable.

$$\text{Set of Nodes} \qquad \mathcal{X} = \{X_1 \dots X_n\} \tag{2}$$

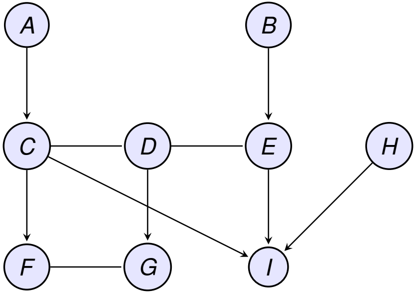
- A pair of nodes X_i, X_j can be connected by a directed edge $X_i \rightarrow X_j$ or an undirected edge $X_i - X_j$.

$$\text{Set of Edges} \qquad \mathcal{E} = X_i \rightarrow X_j \quad \text{or} \quad X_i - X_j \tag{3}$$

Directed Graph

- A graph is directed if all edges are directed. $X_i \rightarrow X_j$.

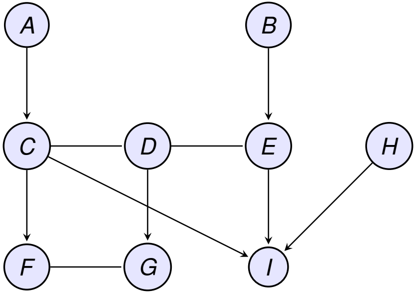
$$\mathcal{G} = (\mathcal{X}, \mathcal{E}) \quad \textit{where} \quad \mathcal{E} = \{X_i \rightarrow X_j\} \tag{4}$$



Undirected Graph

- A graph is undirected if all edges are undirected. $X_i - X_j$.

$$\mathcal{H} = (\mathcal{X}, \mathcal{E}') \quad where \quad \mathcal{E}' = \{X_i - X_j\} \tag{5}$$

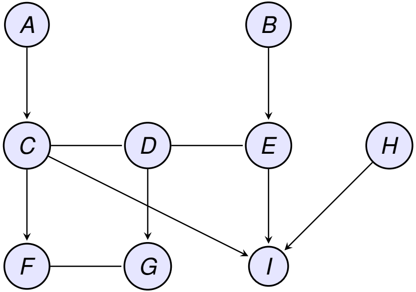


Parent and Child

Graph $\mathcal{K} = (\mathcal{X}, \mathcal{E})$ *where* $\mathcal{E} = \{X \rightleftharpoons Y\}$

- Parent
  - X is called the parent of Y.
  - Pa_X denote parents of X.
- Child
  - Y is called the child of X.
  - Ch_X denote children of X.
- Example: Identify the parents and children of Node E.

  Ans: $Pa_E = B$ $Ch_E = I$

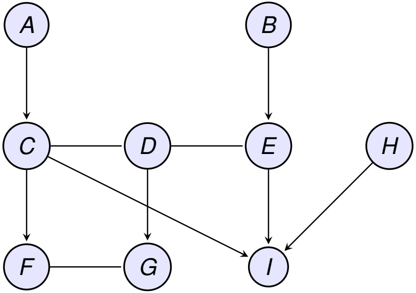


Neighbor and Boundary

- Neighbor
  - Whenever $X \rightleftharpoons Y \in \mathcal{E}$, X and Y are adjacent.
  - Nb_X denote neighbors of X.
- Boundary
  - In Directed graph, $Boundary_X = Pa_X$.
  - In Undirected graph, $Boundary_X = Nb_X$.

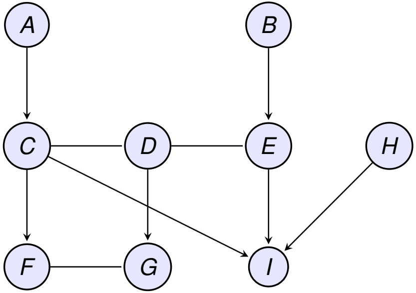
$$Boundary_X = Pa_X \cup Nb_X$$

- Example: Identify the neighbors and boundary of Node C.
  Ans: $Nb_C = F, D, I \qquad B_C = A, D, F, I$



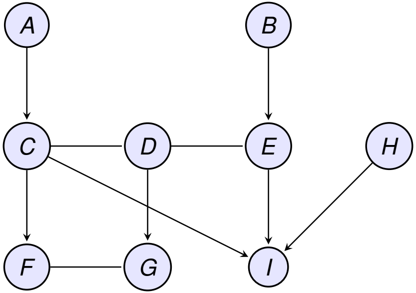
Degree of a Graph

- The degree of a node X is the number of edges in which it participates.
- The in-degree of a node is the number of directed edges $Y \rightarrow X$.
- The degree of a graph is the maximal degree of a node in the graph.
- Example: Identify the degree Node I. Ans=3



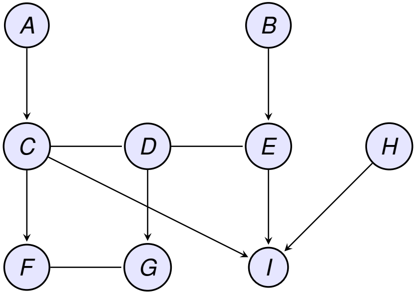
Path

- $X_1, \ldots, X_k$ form a path in the graph $\mathcal{K} = (\mathcal{X}, \mathcal{E})$, if for every $i = 1, \ldots, k-1$ we have either $X_i \rightarrow X_{i+1}$ or $X_i - X_{i+1}$.
- Example: Identify a path. Ans: $A \rightarrow C \rightarrow I$



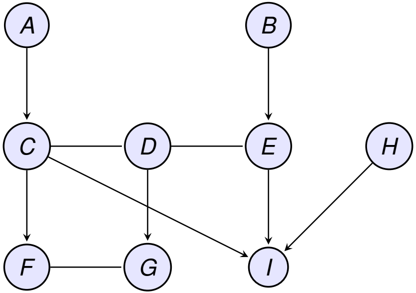
Trail

- $X_1, \ldots, X_k$ form a trail in the graph $\mathcal{K} = (\mathcal{X}, \mathcal{E})$, if for every $i = 1, \ldots, k-1$ we have either $X_i \rightleftharpoons X_{i+1}$.
- A graph is connected, if there is a trail between X_i and X_j.
- Example: Identify a trail. Ans: $A \rightarrow C - D - E \rightarrow I$



Ancestor and Descendant

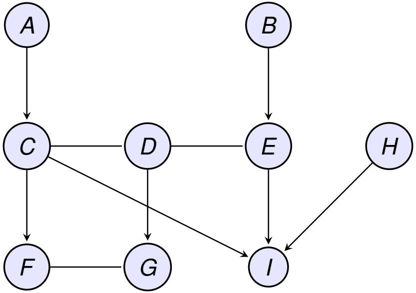
- X is an ancestor of Y in a graph $\mathcal{K}$ if there is a directed path $X_1, \ldots, X_k$ with $X_1 = X$ and $X_k = Y$.
- Y is the descendant of X.
- $Ancestor_X$ and $Descendant_X$



Cycle

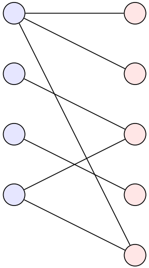
- A cycle in graph $\mathcal{K}$ is a directed path $X_1, \ldots, X_k$ with $X_1 = X_k$.
- Example: Identify a cycle.

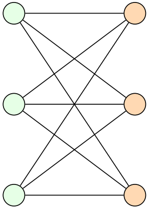
  Ans: No cycle



Bipartite Graphs

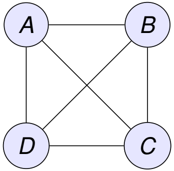
- If the vertex-set of a graph $\mathcal{G}$ can be split into two disjoint sets, V_1 and V_2 , in such a way that each edge in the graph joins a vertex in V_1 to a vertex in V_2 , and there are no edges in $\mathcal{G}$ that connect two vertices in V_1 or two vertices in V_2 , then the graph $\mathcal{G}$ is called a bipartite graph.
- Document-Terms, Student-Class, Movie preference of viewers

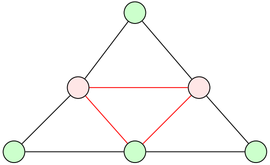




Clique

- Clique is a subset of vertices of an undirected graph such that every two distinct vertices in the clique are adjacent.
- A maximum clique of a graph, is a clique, such that there is no clique with more vertices.





Directed Acyclic Graph (DAG)

- A graph is acyclic if it contains no cycles.
- A directed acyclic graph is a graph that has directed edges but no cycles.
- DAG is the basic graphical representation of Bayesian Networks.

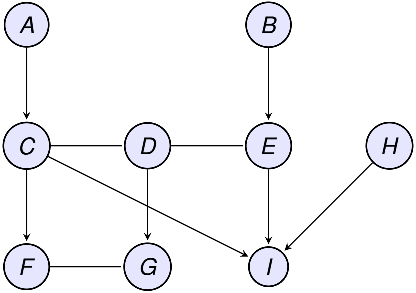


Table of Contents

Uncertainty

- We select a course of actions among many possibilities.
- Decisions may be based on the information obtained from the environment, previous knowledge and the objectives.
- Eg: It looks cloudy. Should I carry umbrella?
- The information and knowledge is incomplete or unreliable. So the decisions made are not certain. We make decisions under uncertainty.
- One of the goals of AI is to develop systems that can reason and make decisions under uncertainty.

Uncertainty

- Due to
  - Partial observability
  - Non-determinism
- Complexity increases
  - Each piece of knowledge may not be independently used to arrive at decisions.
  - Deduced facts are maintained along with new facts. This increases the knowledge base.
- Examples
  - A medical doctor in an emergency.
  - An autonomous vehicle that detects what might be an obstacle in its way.
  - A financial agent needs to select the best investment.

Uncertain Reasoning

Example

- Diagnosing a dental patients' toothache.
- Toothache may be causes of various reasons.
  Equation using propositional logic:

$$Toothache \implies Cavity \lor GumProblem \lor Abscess \lor \ldots$$

- Change to a causal rule.

$$Cavity \implies Toothache$$

  But not all cavity cause toothache.
- So make logically exhaustive.

Uncertain Reasoning

3 reasons for failure when using logic in Judgmental domains [medical diagnosis, law, business, design, automobile repair, gardening, ...]

- Laziness – complete set of antecedents and consequences
- Theoretical ignorance – no complete theory
- Practical ignorance – not all tests can be run

Belief and Degree of Belief

- Belief State is a representation of a set of all possible world states.
- Agent's knowledge can provide only a degree of belief.
- Tool to deal with Degree of Belief is Probability Theory.

Belief is derived from

1. statistical data.
2. some general knowledge.
3. combination of evidence sources.

Probabilistic Statements

- Probability statements instead of propositional logic.
- Probability statements are made with respect to knowledge state.

Example

1. Probability that a patient has a cavity, given that she has toothache is 0.8.
2. Probability that a patient has a cavity, given that she has toothache and a history of gum disease is 0.4.

Table of Contents

Sample Space

- A sample space Ω specifies set of all possible outcomes that we want to consider.

$$\text{Coin toss} \qquad \Omega = \{H, T\}$$

$$\text{Die Roll} \qquad \Omega = \{⚀, ⚁, ⚂, ⚃, ⚄, ⚅\}$$

- Probability of an outcome $P(\omega)$ specifies the chance or probability with each possible outcome.

$$P(H) = 0.5$$

$$P(⚄) = \frac{1}{6}$$

Measurable Event

- An event space $\mathcal{S}$ or Φ is a subset of outcomes, where probabilities can be assigned.
- We are interested in the set of outcomes.

Even die roll $E = \{$⚁, ⚃, ⚅$\}$

Prime die roll $M = \{$⚁, ⚂, ⚄$\}$

- Properties of Event Space
  - Event space contains empty event ϕ and the trivial event Ω.
  - It is closed under union.

$$\text{If} \quad \alpha, \beta \in \mathcal{S}, \quad \text{then} \quad \alpha \cup \beta \in \mathcal{S}$$

  - It is closed under complementary.

$$\text{If} \quad \alpha \in \mathcal{S}, \quad \text{then} \quad \Omega - \alpha \in \mathcal{S}$$

Probability of Event

- Probability of an event is given by the sum of the probabilities of the outcomes it contains.

$$P(\alpha) = \sum_{\omega \in \alpha} P(\omega) \tag{6}$$

$$\text{Even die roll} \qquad P(E) = \frac{3}{6} = 0.5$$

$$\text{Prime die roll} \qquad P(M) = \frac{3}{6} = 0.5$$

Prior Probability

- Prior or Unconditional probabilities refer to degree of belief in the absence of any other information.

$$P(DieTotal = 11) = P((5, 6)) + P((6, 5)) = 1/18$$

Evidence

- The information that has already been revealed is called evidence.
  - She is having toothache. $Toothache = True$ *or* $toothache$
  - We roll a dice and we get 5. $Die_1 = 5$

Posterior Probability

- Conditional or Posterior probability refer to the probability of some event occurring given a particular condition.

$$P(cavity \mid toothache) = 0.6$$

- Condition on all evidences that has been observed.

$$P(\alpha|\beta) = \frac{P(\alpha \wedge \beta)}{P(\beta)} \qquad where\ P(\beta) > 0 \tag{7}$$

Probability Model

- Associate a numerical probability $P(\omega)$ with each event $\mathcal{S}$.
- Axioms of probability theory

$$P(\omega) \geq 0 \tag{8}$$

$$P(\Omega) = \sum_{\omega \in \Omega} P(\omega) = 1 \tag{9}$$

$$P(\phi) = 0 \tag{10}$$

$$P(\alpha \mid \beta) = \frac{P(\alpha \wedge \beta)}{P(\beta)} \qquad \textit{where } P(\beta) > 0 \tag{11}$$

$$P(\alpha \vee \beta) = P(\alpha) + P(\beta) - P(\alpha \wedge \beta) \tag{12}$$

Bayes Rule

- Conditional probabilities can be derived from the prior given the evidence.

$$P(\alpha|\beta) = \frac{P(\beta|\alpha)P(\alpha)}{P(\beta)} \qquad \textit{where } P(\beta) > 0 \tag{13}$$

Example 1

- Consider the student population, and let Smart denote smart students and GradeA denote students who got grade A. Based on estimates from past statistics assume that $P(GradeA|Smart) = 0.6$, the probability for students being smart is 0.3 and the prior probability of students receiving high grades is 0.2. Estimate the probability that the student is smart.

$$\text{Given,} P(Smart) = 0.3$$
$$P(GradeA) = 0.2$$
$$P(GradeA|Smart) = 0.6$$

According to Bayes' rule

$$P(Smart|GradeA) = \frac{0.6 * 0.3}{0.2} = 0.9$$

Example 2

- Suppose that a tuberculosis (TB) skin test is 95 percent accurate. Suppose that 1 in 1000 of the subjects who get tested is infected. What is the probability of getting a positive test result?

$$\begin{aligned}
\text{Given}, P(TB) &= 0.001 \\
P(\text{infected subjects get a positive result}) &= 0.001 * 0.95 \\
P(\text{uninfected subjects get a positive result}) &= 0.999 * 0.05 \\
P(Positive) &= 0.001 * 0.95 + 0.999 * 0.05 = 0.0509 \\
\text{According to Bayes' rule,} & \\
P(TB|Positive) &= \frac{0.001 * 0.95}{0.0509} = 0.0187 \\
&< 2\%
\end{aligned}$$

Table of Contents

Random Variable

- Variables used in probability theory.
- Uppercase letter
- A random variable X is defined by a function that associates with each outcome in Ω a value or a state.

$$\mathbf{P}(attribute = value) = \mathbf{P}(\omega \in \Omega : f_{attribute} = value)$$

$$\mathbf{P}(X = x) = \mathbf{P}(\omega \in \Omega : X(\omega) = x) \tag{14}$$

- Use $P(x)$ as a shorthand for $P(X = x)$.

$$\sum_{x \in Val(X)} P(X = x) = \sum_{x} P(x) = 1 \tag{15}$$

Domain of Random Variable

- Every random variable has a domain, the set of possible values it can take.
- Finite random variable or Infinite random variable.
- Domain can be discrete or continuous (integer or real).
  Weather random variable – Domain $= \{sunny, overcast, rainy, cloudy, snow\}$
- Boolean Random variable
  - $Domain = Val(X) = \{true, false\}$
  - x^1 to denote *true* and x^0 to denote *false*.
  - The distribution of binary random variable is called a Bernoulli distribution.

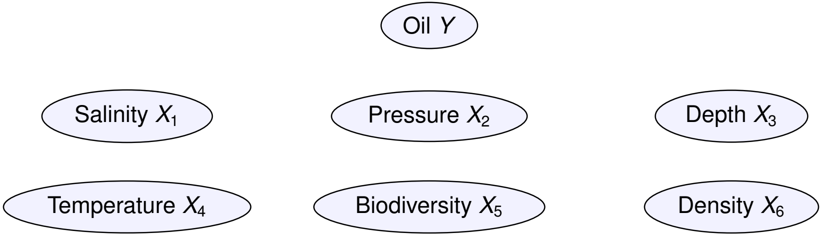
Joint Distribution

- The distribution over several random variables are described using joint distribution.
- Set of random variables $\mathcal{X} = \{X_1, X_2, \ldots, X_n\}$
- Joint Distribution $\mathbf{P}(\mathcal{X}) = P(X_1, X_2, \ldots, X_n)$.
- Eg: Suppose random variable *Grade* reports the final grade of a student and the student's intelligence is given by *Intelligence*. Then the joint distribution $\mathbf{P}(Intelligence, Grade)$

		Intelligence	
		low	high
	A	0.07	0.18
Grade	B	0.28	0.09
	C	0.35	0.03

Why Joint Distribution?

- Once we know the joint distribution of the participating random variables, all inference queries regarding Marginal Distribution and Conditional Probability Distribution can be answered.



$$JD = P(Y, X_1, X_2, X_3, X_4, X_5, X_6)$$

Marginal Distribution

- The distribution over events that can be described using X is often referred to as the marginal distribution over the random variable X.
- Summing out
- Marginal distribution is denoted by $P(X)$.
- Row-wise or Column-wise summations in the JD gives MD.

$$\mathbf{P}(Y) = \sum_{z \in Z} \mathbf{P}(Y, z) \tag{16}$$

- Sum over all the possible combinations of values of the set of varaibles of Z.

Marginal Distribution

- Eg: Suppose random variable *Grade* reports the final grade of a student . Then the marginal distribution of *Grade*

$$P(Grade = A) = 0.25$$
$$P(Grade = B) = 0.37$$
$$P(Grade = C) = 0.38$$
$$\sum P(Grade) = 1$$

Conditional Probability Distribution

- The conditional distribution over a random variable given an observation of the value of another random variable is referred to as Conditional Probability Distribution.
- Compute conditional probability of some variable given evidence about others.

$$\mathbf{P}(Y) = \sum_{z} \mathbf{P}(Y \mid z) P(z) \tag{17}$$

- Eg: What is the probability for the student to have high intelligence given that the grade scored is A.

$$\begin{aligned} P(Intelligence = high | Grade = A) &= \frac{0.18}{0.25} = 0.72 \\ &\neq P(Intelligence) \\ &\neq \text{Marginal} \quad \text{Distribution} \end{aligned}$$

Exercise

Given Joint Distribution $\mathbf{P}(Cavity, Toothache, Catch)$ of 3 binary random variables.

	toothache		*¬toothache*	
	catch	*¬catch*	*catch*	*¬catch*
cavity	0.108	0.012	0.072	0.008
¬cavity	0.016	0.064	0.144	0.576

1. Compute $P(cavity \lor toothache)$.
2. Compute $P(cavity)$.
3. Compute $P(cavity)$ given the evidence of toothache .

Exercise - Answer to Q1

	toothache		$\neg$*toothache*	
	catch	$\neg$*catch*	*catch*	$\neg$*catch*
cavity	0.108	0.012	0.072	0.008
$\neg$*cavity*	0.016	0.064	0.144	0.576

Compute $P(cavity \lor toothache)$.

$$
\begin{aligned}
P(cavity \lor toothache) &= 0.108 + 0.012 + 0.072 + 0.008 + 0.016 + 0.064 \\
&= 0.28
\end{aligned}
$$

Exercise - Answer to Q2

	toothache		*¬toothache*	
	catch	*¬catch*	*catch*	*¬catch*
cavity	0.108	0.012	0.072	0.008
¬cavity	0.016	0.064	0.144	0.576

Compute $P(cavity)$.

$$\begin{aligned} P(cavity) &= 0.108 + 0.012 + 0.072 + 0.008 \\ &= 0.2 \end{aligned}$$

This is the unconditional or marginal probability.

Exercise - Answer to Q3

Compute $P(cavity)$ given the evidence of toothache.

$$
\begin{aligned}
\mathbf{P}(Cavity \mid toothache) &= \frac{\mathbf{P}(Cavity \vee toothache)}{\mathbf{P}(toothache)} \\
&= \alpha \mathbf{P}(Cavity \vee toothache) \qquad \text{Let } \alpha = 1/\mathbf{P}(toothache) \\
&= \alpha \left[P(Cavity, toothache, catch) + P(Cavity, toothache, \neg catch) \right] \\
&= \alpha [< 0.108, 0.016 > + < 0.012, 0.064 >] \\
&= \alpha < 0.12, 0.08 > \qquad \alpha = \frac{1}{0.12 + 0.08} = \frac{1}{0.2} \\
&= < 0.6, 0.4 > \\
P(cavity) &= 0.6
\end{aligned}
$$

Challenges with Joint Distribution

- Computational: Expensive to manipulate and too large to to store.
- Cognitive: Impossible to acquire so many numbers from a human.
- Statistical: Need huge amounts of data to learn the parameters.

Need to compactly represent joint distribution.
Solution: Exploit independence or conditional independence among random variables.

Independence of Events

- An event α is independent of event β in P,

$$P(\alpha \perp \beta) \quad \text{if} \quad \begin{cases} P(\alpha \mid \beta) = P(\alpha) & \text{or} \\ P(\beta \mid \alpha) = P(\beta) & \text{or} \\ P(\alpha \wedge \beta) = P(\alpha)P(\beta) \end{cases} \tag{18}$$

- $\alpha \perp \beta$ implies $\beta \perp \alpha$
- Eg: Tossing two coins, Rolling a die.

Independence of Random Variables

- Two random variables X and Y can be independent of each other.
- A random variable X is independent of another random variable Y,

$$P(X \perp Y) \quad \text{if} \quad \begin{cases} P(X \mid Y) = P(X) & \text{or} \\ P(Y \mid X) = P(Y) & \text{or} \\ P(X \wedge Y) = P(X)P(Y) \end{cases} \tag{19}$$

- $X \perp Y$ implies $Y \perp X$
- Eg: Weather is independent of dental problems.

Conditional Independence of Events

- An event α is conditionally independent of event β given event γ in P

$$P(\alpha \perp \beta \mid \gamma) \quad \text{if} \quad \begin{cases} P(\alpha \mid \beta \wedge \gamma) = P(\alpha \mid \gamma) \\ P(\alpha \mid \beta \wedge \gamma) = P(\beta \mid \gamma) = 0 \\ P(\alpha \wedge \beta \mid \gamma) = P(\alpha \mid \gamma)P(\beta \mid \gamma) \end{cases} \tag{20}$$

- Eg: Getting Admission in MIT is independent of getting admission in Stanford, given the student has scored Grade A.

Conditional Independence of Random Variables

- A random variable X is conditionally independent of random variable Y given random variable Z

$$P(X \perp Y \mid Z) \quad \text{if} \quad \begin{cases} P(X \mid Y, Z) = P(X \mid Z) \\ P(X \mid Y, Z) = P(Y \mid Z) \\ P(X, Y \mid Z) = P(X \mid Z)P(Y \mid Z) \end{cases} \tag{21}$$

- Eg: The random variables Toothache and Catch are independent given the presence or absence of Cavity.

Student Example

- Model the difficulty of a course, intelligence of students, Grade the students score in a particular course.
- Let D represent the difficulty of a course.

$$
\begin{aligned}
\textit{Domain of } \mathbf{D} &= \{easy, hard\} \\
\mathbf{P}(\mathbf{D}) &= \{d^0, d^1\} \\
&= \{0.6, 0.4\}
\end{aligned}
$$

- Let I represent the intelligence of a student.

$$
\begin{aligned}
\textit{Domain of } \mathbf{I} &= \{low, high\} \\
\mathbf{P}(\mathbf{I}) &= \{i^0, i^1\} \\
&= \{0.7, 0.3\}
\end{aligned}
$$

Student Example

- Let G represent the grade a student gets for a course.

$$\textit{Domain of } \mathbf{G} = \{A, B, C\}$$
$$\mathbf{P}(\mathbf{G}) = \{g^1, g^2, g^3\}$$

- $\mathbf{P}(\mathbf{D}, \mathbf{I}, \mathbf{G})$ denotes the probabilities of all combinations of the values of the 3 random variables.
- These $2 * 3 * 3 = 12$ values can be represented using a Joint Distribution Table.

Joint Probability Distribution

- Joint Probability Distribution completely represents the joint distribution for all random variables.
- In the students example, the $\mathbf{P}(\mathbf{D}, \mathbf{I}, \mathbf{G})$, the 12 parameters cannot be determined by the value of the other parameters. Hence called Independent parameters.
- Independent parameters are parameters whose values are not completely determined by the values of the other parameters.

Student Example - Joint Distribution

I	D	G	P
i^0	d^0	g^1	0.126
		g^2	0.168
		g^3	0.126
i^0	d^1	g^1	0.009
		g^2	0.045
		g^3	0.126
i^1	d^0	g^1	0.052
		g^2	0.0224
		g^3	0.0056
i^1	d^1	g^1	0.069
		g^2	0.036
		g^3	0.024

$$\sum \mathbf{P(D, I, G)} = 1$$

REFERENCES

1. Probabilistic Graphical Models: Principles and Techniques by Daphne Koller and Nir Friedman. MIT Press. 2009
2. Artificial Intelligence: A Modern Approach (3rd Edition) by Stuart Russell, Peter Norvig
3. Mastering Probabilistic Graphical Models using Python by Ankur Ankan, Abhinash Panda. Packt Publishing 2015.
4. Learning in Graphical Models by Michael I. Jordan. MIT Press. 1999

Thank You !!!