

CS578 – INTERACTIVE AND TRANSPARENT MACHINE LEARNING

TOPIC: 1. PROBABILITIES; 2. GRAPHS



Mustafa Bilgic



<http://www.cs.iit.edu/~mbilgic>



<https://twitter.com/bilgicm>

OUTLINE

○ Motivation

- Learn (or refresh our memory about) basic concepts that will be used throughout this class

○ Topics

- Probability Theory
- Graphs

PROBABILITY DISTRIBUTION

- **Ω : Space** of possible outcomes
 - E.g., Rolling a die $\Omega = \{1, 2, 3, 4, 5, 6\}$
- **S: Measurable events**
 - E.g., An odd roll of die $S = \{1, 3, 5\}$
- A **probability distribution P** over (Ω, S) is a mapping from events in S to real values that satisfies
 - $P(\alpha) \geq 0$ for all $\alpha \in S$
 - $P(\Omega) = 1$
 - If $\alpha, \beta \in S$ and $\alpha \cap \beta = \emptyset$, then $P(\alpha \cup \beta) = P(\alpha) + P(\beta)$

RANDOM VARIABLES

- A problem is represented through variables
 - Age, gender, lab tests, ...
 - Intelligence (of a student), Difficulty (of a class), Grade (of a student in that class), ...
- A variable takes on values from its domain
 - Gender takes on Male, Female
 - Grade takes on A, B, C
- Can be either discrete or continuous
 - Gender is discrete, age is continuous
- In an uncertain world, a variable takes on values from its domain probabilistically
 - For example, Grade can be A, B, or C probabilistically
 - $P(\text{Grade} = A)$, $P(\text{Grade} = B)$, $P(\text{Grade} = C)$

RANDOM VARIABLES – NOTATION

- Capital: X : variable
- Lowercase: x : a particular value of X
- $\text{Val}(X)$: the set of values X can take
- Bold Capital: \mathbf{X} : a set of variables
- Bold lowercase: \mathbf{x} : an assignment to all variables in \mathbf{X}
- $P(X=x)$ will be shortened as $P(x)$
- $P(X=x \cap Y=y)$ will be shortened as $P(x,y)$

TABLE – THE MOST BASIC REPRESENTATION

Intelligence	P(Intelligence)
low	0.7
high	0.3

Grade	P(Grade)
a	0.25
b	0.37
c	0.38

JOINT DISTRIBUTION

- Several random variables
 - $\mathbf{X} = \{X_1, X_2, \dots, X_n\}$
- Joint Distribution
 - $P(\mathbf{X}) = P(X_1, X_2, \dots, X_n)$
 - Specifies a probability value to all possible assignments

JOINT DISTRIBUTION

Intelligence	Grade	P(Intelligence, Grade)
low	a	0.07
low	b	0.28
low	c	0.35
high	a	0.18
high	b	0.09
high	c	0.03

CONDITIONAL PROBABILITY

- $P(X | Y) = P(X, Y) / P(Y)$
- What do the following mean?
 - $P(\text{Grade})$
 - $P(\text{Grade} | \text{Intelligence})$
 - $P(\text{Grade} | \text{Intelligence} = \text{high})$
 - $P(\text{Grade} = a | \text{Intelligence} = \text{high})$

SUMMATION RULE

- Given $P(X, Y)$, $P(X)$ can be computed using
 - $P(X) = \sum_y P(X, y)$ where y ranges over $\text{Val}(Y)$
- Answer the following
 - $\sum_x P(X) = ?$
 - $\sum_x P(X | y) = ?$
 - $\sum_y P(X | Y) = ?$

CHAIN RULE

- $P(X_1, X_2, X_3, \dots, X_k) =$
 - $P(X_1) P(X_2 | X_1) P(X_3 | X_1, X_2) \dots P(X_k | X_1, X_2, X_3, \dots, X_{k-1})$
 - or
 - $P(X_2) P(X_1 | X_2) P(X_3 | X_1, X_2) \dots P(X_k | X_1, X_2, X_3, \dots, X_{k-1})$
 - or
 - $P(X_2) P(X_3 | X_2) P(X_1 | X_3, X_2) \dots P(X_k | X_1, X_2, X_3, \dots, X_{k-1})$
 - or
 - Pick an order, then
 - $P(\text{first})P(\text{second} | \text{first})P(\text{third} | \text{first}, \text{second}) \dots P(\text{last} | \text{all_previous})$

BAYES RULE

- Bayes Rule
 - $P(X | Y) = P(Y | X)P(X) / P(Y)$
- Conditional Bayes Rule
 - $P(X | Y, Z) = P(Y | X, Z)P(X | Z) / P(Y | Z)$

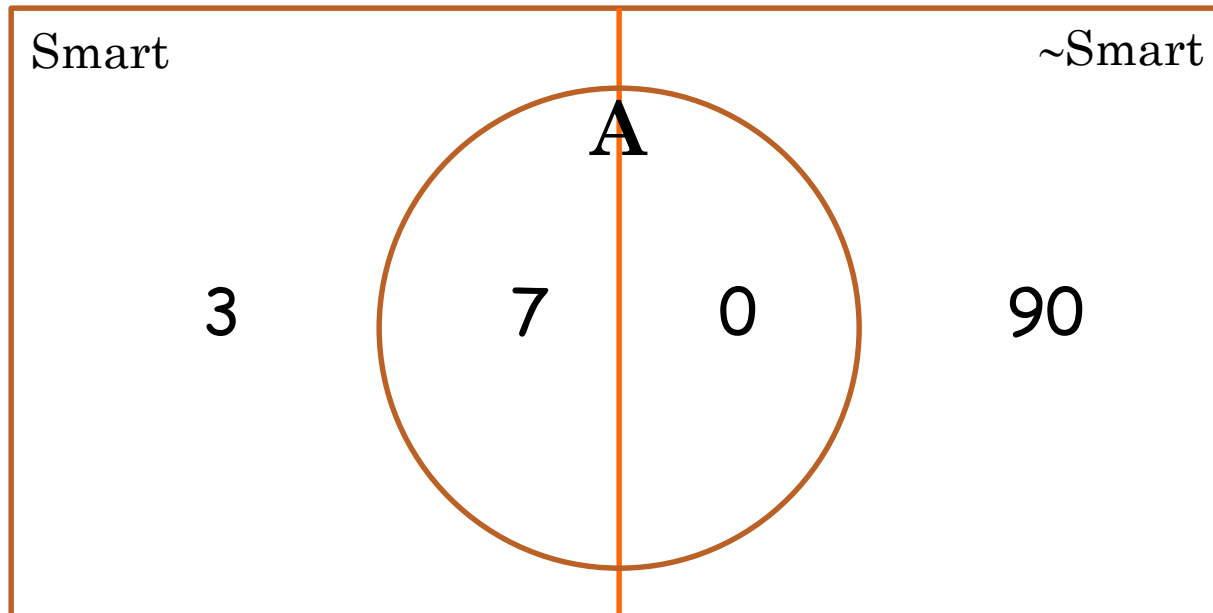
BAYES RULE

- Can we compute $P(\alpha|\beta)$ from $P(\beta|\alpha)$?
- E.g., *Disclaimer: these numbers have nothing to do with the previous tables and numbers*
 - In a class, 70% of the smart students got an A.
 - $P(a \mid \text{smart}) = 0.7$
 - John got an A. What is the probability of John being smart given he got an A?
 - $P(\text{smart} \mid a) = ?$

CLASS EXAMPLE

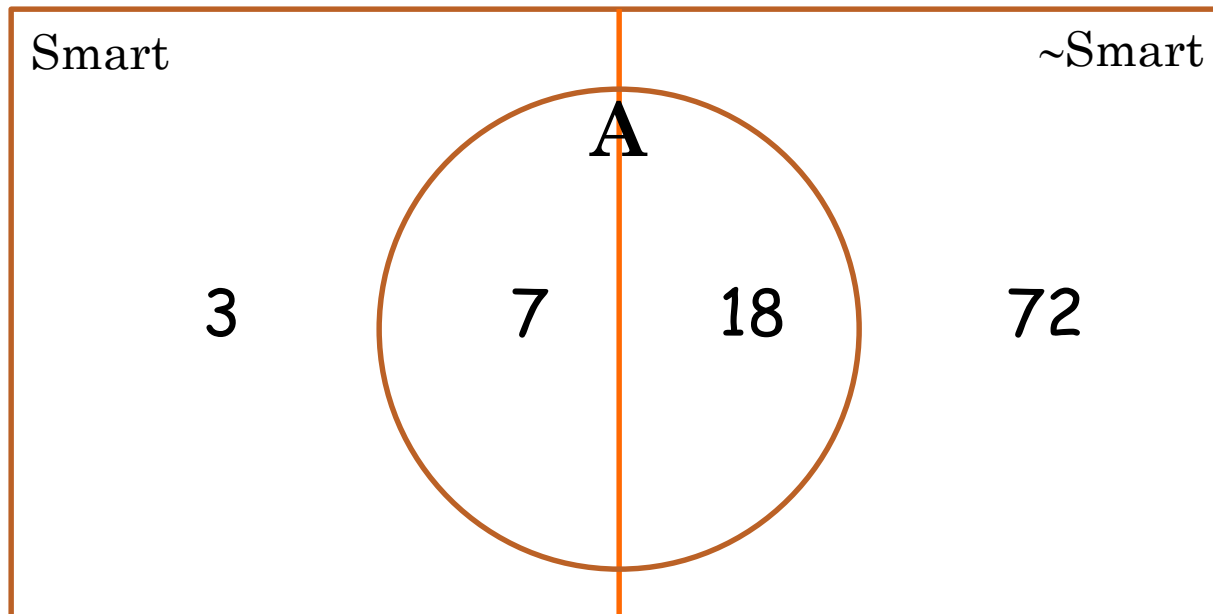
- Let's say there are 100 students in the class
- Let's say 10 of them are Smart, 90 are \sim Smart
- Probability of a randomly picked student being smart
 - $P(s) = 0.1$
- We know that 70% of the smart students got an A.
 - $P(a | s) = 0.7$
 - 7 smart students got an A; 3 did not get an A.
- What is $P(s|a) = ?$
 - Depends on $P(a)$

VERY HARD CLASS



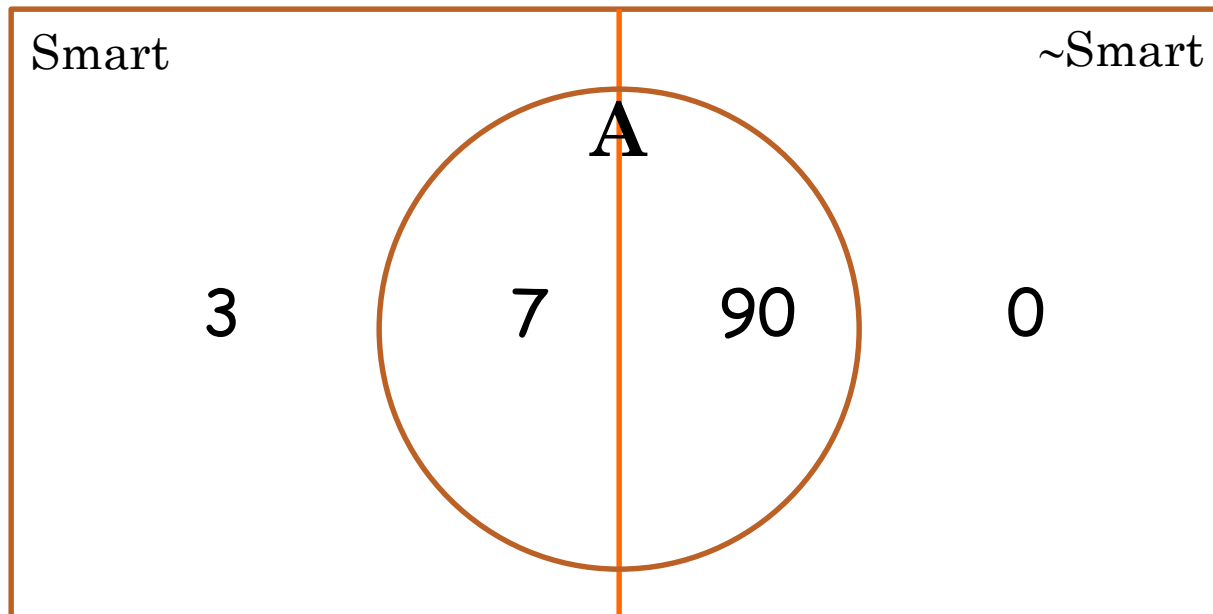
$$P(s | a) = ?$$

MEDIUM HARD CLASS



$$P(s | a) = ?$$

WEIRD CLASS



$$P(s | a) = ?$$

MARGINAL INDEPENDENCE

- An event α is **independent** of event β in P , denoted as $P \models \alpha \perp \beta$, if
 - $P(\alpha \mid \beta) = P(\alpha)$, or
 - $P(\beta) = 0$
- Proposition: A distribution P satisfies $\alpha \perp \beta$ if and only if
 - $P(\alpha, \beta) = P(\alpha) P(\beta)$
 - *Can you prove it?*
- Corollary: $\alpha \perp \beta$ implies $\beta \perp \alpha$

MARGINAL INDEPENDENCE

X	Y	P(X, Y)
t	t	0.18
t	f	0.42
f	t	0.12
f	f	0.28

Is $X \perp Y$?

CONDITIONAL INDEPENDENCE

- Two events are independent given another event
- An event α is **independent** of event β given event γ in P , denoted as $P \models (\alpha \perp \beta \mid \gamma)$, if
 - $P(\alpha \mid \beta, \gamma) = P(\alpha \mid \gamma)$, or
 - $P(\beta, \gamma) = 0$
- Proposition: A distribution P satisfies $\alpha \perp \beta \mid \gamma$ if and only if
 - $P(\alpha, \beta \mid \gamma) = P(\alpha \mid \gamma) P(\beta \mid \gamma)$

CONTINUOUS SPACES

- Assume X is continuous and $\text{Val}(X) = [0,1]$
- If you would like to assign the same probability to all real numbers in $[0, 1]$, what is, for e.g., $P(X=0.5) = ?$
- Answer: $P(X=0.5) = 0$.

PROBABILITY DENSITY FUNCTION

- We define **probability density function**, $p(x)$, a non-negative integrable function, such that $\int_{Val(X)} p(x)dx = 1$

$$P(X \leq a) = \int_{-\infty}^a p(x)dx$$

$$P(a \leq X \leq b) = \int_a^b p(x)dx$$

UNIFORM DISTRIBUTION

- A variable X has a uniform distribution over $[a,b]$ if it has the PDF

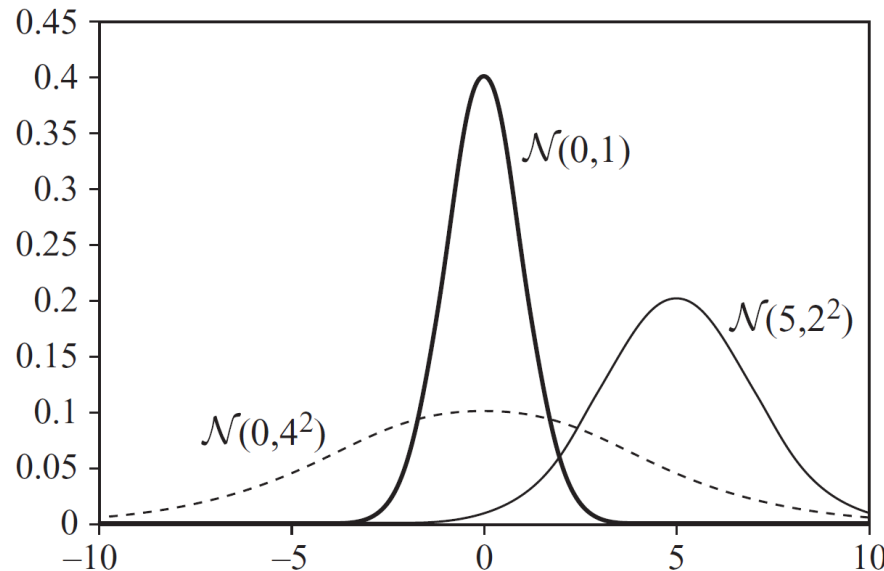
$$p(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

Check and make sure that $p(x)$ integrates to 1.

GAUSSIAN DISTRIBUTION

- A variable X has a Gaussian distribution with mean μ and variance σ^2 , if it has the PDF

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



Can $p(x)$ be ever greater than 1?

CONDITIONAL PROBABILITY

- We want $P(Y | X=x)$ where X is continuous, Y is discrete
- $P(Y | X=x) = P(Y, X=x) / P(X=x)$
 - What's wrong with this expression?
- Instead, we use the following expression

$$P(Y | X = x) = \lim_{\varepsilon \rightarrow 0} P(Y | x - \varepsilon \leq X \leq x + \varepsilon)$$

CONDITIONAL PROBABILITY

- We want $p(Y | X)$ where X is discrete, Y is continuous
- How would you represent it?

EXPECTATION

$$E_P [X] = \sum_x xP(x)$$

$$E_P [X] = \int_x xp(x)dx$$

$$E_P [aX + b] = aE_P [X] + b$$

$$E_P [X + Y] = E_P [X] + E_P [Y]$$

$$E_P [X | y] = \sum_x xP(x | y)$$

What about $E[X*Y]$?

VARIANCE

$$\text{Var}_P[X] = E_P \left[\left(X - E_P[X] \right)^2 \right]$$

$$\text{Var}_P[X] = E_P[X^2] - \left(E_P[X] \right)^2$$

Can you derive the second expression using the first expression?

$$\text{Var}_P[aX + b] = a^2 \text{Var}_P[X]$$

What is $\text{Var}[X+Y]$?

UNIFORM AND GAUSSIAN DISTRIBUTION

- If $X \sim N(\mu, \sigma^2)$, then $E[X] = \mu$, $\text{Var}[X] = \sigma^2$
- What about the expectation and variance of a uniform distribution?

GRAPHS

- A **graph** consists of **nodes** and **edges**
- **Nodes:** $\mathcal{X} = \{X_1, X_2, \dots, X_n\}$
- **Undirected Edge:** $X_i - X_j$
- **Directed Edge:** $X_i \rightarrow X_j$
- Between a pair of nodes, at most one type of edge exists
 - We cannot have $X_i \rightarrow X_j$ and $X_j \rightarrow X_i$ at the same time, and
 - We cannot have $X_i \rightarrow X_j$ and $X_i - X_j$ at the same time
- Some edge: $X_i \rightleftharpoons X_j$

DIRECTED AND UNDIRECTED

- A graph is **directed** if its *all* edges are directed
- A graph is **undirected** if its *all* edges are undirected

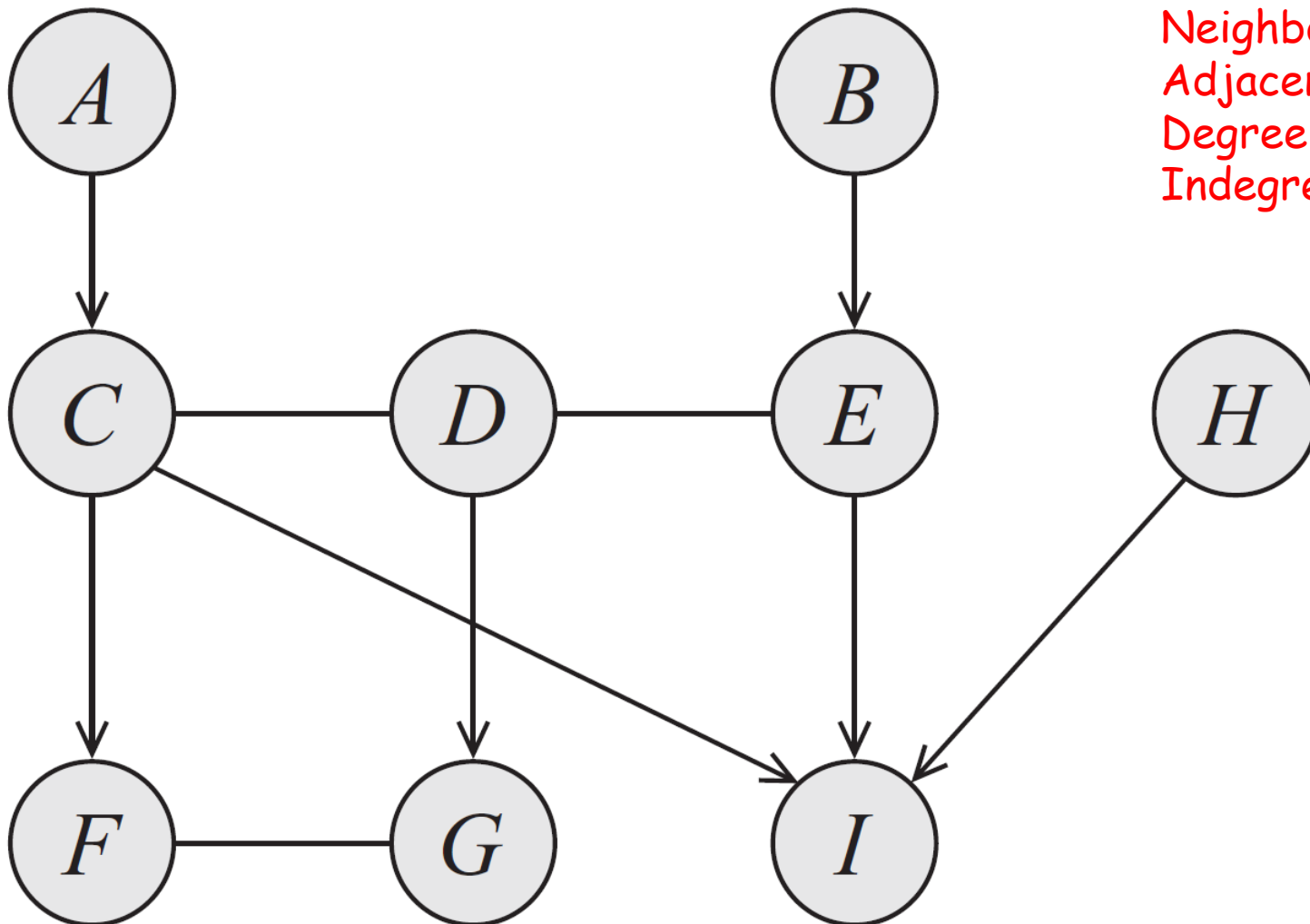
RELATIONSHIPS

- $X_i \rightarrow X_j$
 - X_i is the **parent**
 - X_j is the **child**
- $X_i - X_j$
 - X_i and X_j are **neighbors**
- $X_i \rightleftharpoons X_j$
 - X_i and X_j are **adjacent**
- **Degree** of X_i : The number of edges X_i is part of
- **Indegree** of X_i : The number of directed edges pointing to X_i
- **Degree** of a graph: The maximal degree of a node in the graph

EXAMPLE

Examples of:

Parents
Children
Neighbors
Adjacent nodes
Degree
Indegree



COMPLETE GRAPHS AND CLIQUES

- A subgraph over $\mathbf{X} \subseteq \mathcal{X}$ is **complete** if *every* two nodes in \mathbf{X} are connected by some edge
- Such a set \mathbf{X} is also called a **clique**
- A clique is maximal if for any superset of nodes $\mathbf{Y} \supset \mathbf{X}$, \mathbf{Y} is not a clique

PATHS AND TRAILS

- X_1, X_2, \dots, X_k forms a **path** if, for every $i = 1, 2, \dots, k-1$, we have that either $X_i - X_{i+1}$ or $X_i \rightarrow X_{i+1}$.
- A path is directed if, for *at least one* i , $X_i \rightarrow X_{i+1}$.
- X_1, X_2, \dots, X_k forms a **trail** if, for every $i = 1, 2, \dots, k-1$, we have $X_i \rightleftharpoons X_{i+1}$.
- What is the difference between a path and a trail? Is every path also a trail? Is every trail also a path?

ANCESTORS AND DESCENDANTS

- X_i is an **ancestor** of X_j if there is a directed path from X_i to X_j
- X_i is a **descendant** of X_j if there is a directed path from X_j to X_i
- **Nondescendants**(X_i) $\equiv \mathcal{X} \setminus \text{Descendants}(X_i)$

CYCLES AND LOOPS

- A **cycle** is a directed path from a node to itself
- A graph is **acyclic** if it contains no cycles
- A directed acyclic graph is the one where all edges are directed and there are no cycles
- A **loop** is a trail from a node to itself
- A graph is **singly-connected** if it contains no loops