Probability Theory

A Mathematical Framework for Representing Uncertain Statements

Sanjay Singh*†

*Department of Information and Communication Technology
Manipal Institute of Technology, Manipal University
Karnataka-576104, INDIA
sanjay.singh@manipal.edu

†Centre for Artificial and Machine Intelligence (CAMI)
Manipal University, Karnataka-576104, INDIA

February 15, 2017

Why Probability?

Possible Sources of Uncertainty

- Inherent stochasticity in the system being modeled
- Incomplete observability-even deterministic systems can appear stochastic when one cannot observe all the variables that drive the behavior of the system, e.g Monti Hall problem
- Incomplete modeling-using a model that discard some of the observed information, the discarded information results in uncertainty in the model's predictions

- Frequentist probability (analyzes the frequency of events)
- Bayesian probability (degree of belief, related to qualitative levels of certainty)
- Probability can be seen as the extension of logic to deal with uncertainty
- Probability theory provides a set of formal rules for determining the likelihood of a proposition being true given the likelihood of other propositions

Elements of Probability

- **Sample space** Ω:The set of all the outcomes of a random experiments (each outcome $\omega \in \Omega$)
- **Set of events (event space)** $\mathcal{F}$: A set whose elements $A \in \mathcal{F}$ (called events) are subsets of Ω (i.e., $A \in \subseteq \Omega$ is a collection of possible outcomes of an experiment)
- **Probability measure**: A function $P : \mathcal{F} \mapsto \mathbb{R}$ that satisfies following properties
  - $P(A) \geq 0, \quad \forall A \in \mathcal{F}$
  - $P(\Omega) = 1$
  - If $A_1, A_2, \ldots$ are disjoint events (i.e., $A_i \cap A_j = \emptyset$ whenever $i \neq j$), then

$$P(\cup_i A_i) = \sum_i P(A_i)$$

These properties are called **Axioms of Probability**

Definition

Algebraic Operations Let A and B be two events of the sample space Ω. We will denote

- "A does not occur" by $\overline{A}$
- "Either A or B occur" by $A \cup B$
- "Both A and B occur" by A, B or $A \cap B$
- "A occur and B does not" by $A \backslash B \equiv A \cap \overline{B}$

Conditional Probability

- Let $P(B) \neq 0$, then the conditional probability of any event A given B is defined as $P(A|B) = \dfrac{P(A \cap B)}{P(B)}$
- $P(A|B)$ is the probability of event A after observing the occurrence of event B
- Two events are independent iff $P(A \cap B) = P(A)P(B)$ (or $P(A|B) = P(A)$)
- Total probability theorem

$$P(B) = P(A, B) + P(\overline{A}, B) = P(B|A)P(A) + P(B|\overline{A})P(\overline{A})$$

Random Variables

- Consider an experiment in which we flip 10 coins
- Want to know number of coins that comes up heads
- Sample space, $\omega_0 =< H, H, T, T, H, T, H, H, T, T >\in \Omega$
- We don't care about probability of obtaining any particular sequence of heads and tails, instead we care about real-valued functions of outcome, such as the number of heads that appear among 10 coins
- These function are called as **random variables**
- Formally, a random variable X is a function $X : \Omega \mapsto \mathbb{R}$, denoted as $X(\omega)$
- Probability of a set associated with a random variable X taking on some specific values k is written as

$$P(X = k) = P(\{\omega : X(\omega) = k\})$$

- $P(a \leq X(\omega) \leq b) = P(\{\omega : a \leq X(\omega) \leq b\})$, when $X(\omega)$ is a continuous random variable

Probability Distributions

A probability distribution is a description of how likely a random variable (or set of rv) is to take on each of its possible states. Probability distribution description depends on whether the variables are discrete or continuous.

Notation $X \sim P$: Random variable X has/follows distribution P

- To sepcify probability measures when dealing with r.v, it is convenient to specify alternative functions: CDFs, PMFs, and PDFs
- The probability measures for an experiments can easily be obtained from CDFs, PDFs or PMFs
- **Cumulative distribution function (CDF)** is a function $F_X : \mathbb{R} \mapsto [0, 1]$, which specifies probability measures as

$$F_X(x) \triangleq P(X \leq x)$$

- Properties of CDF
  - $0 \leq F_X(x) \leq 1$
  - $\lim_{x \to -\infty} F_X(x) = 0$
  - $\lim_{x \to \infty} F_X(x) = 1$
  - $x \leq y \Longrightarrow F_X(x) \leq F_X(y)$

Probability Mass Function

- Probability distribution over discrete variables is described using a **probability mass function**(PMF)
- Probability measures associated with a rv is to directly specify the probability of each value that rv can assume
- Probability mass function (PMF) is a function $p_X : \Omega \mapsto \mathbb{R}$ such that

$$p_X(x) \triangleq P(X = x)$$

- Notation $Val(X)$ represents the set of possible values that rv X may assume
- Properties of PMF
  - $0 \leq p_X(x) \leq 1$
  - $\sum_{x \in Val(X)} p_X(x) = 1$
  - $\sum_{x \in A} p_X(x) = P(X \in A)$

Probability Density Function

- For continuous random variables, CDF $F_X(x)$ is differentiable, here we define **probability density function** (PDF) as

$$f_X(x) \triangleq \frac{dF_X(x)}{dx}$$

- PDF at any given point x is not the probability of that event, i.e., $f_X(x) \neq P(X = x)$
- CDF can take value larger than one
- Properties of PDF
  - $f_X(x) \geq 0$
  - $\int_{-\infty}^{\infty} f_X(x)dx = 1$
  - $\int_{x \in A} f_X(x)dx = P(X \in A)$

Marginal Probability

- When we know the probability distribution (PD) over a set of variables and want PD over just a subset of them
- Probability distribution over the subset is known as **marginal probability** distribution
- For example, $P(X, Y)$, we can find $P(X)$ with the sum rule

$$\forall x \in X, P(X = x) = \sum_{y} P(X, Y)$$

- Term "marginal probability" comes from the process of computing probabilities on paper
- When values of $P(X, Y)$ are written in a grid with different values of x in rows and different values of y in column
- Sum across a row of the grid, then write $P(X)$ in the margin of the paper just to the right of the paper

Multiplication Rule

- The joint probability of a set of random variables $X_1, X_2, \ldots, X_n$ can be expressed as

$$P(X_1, X_2, \ldots, X_n) = P(X_1) \prod_{i=2}^{n} P(X_i | X_1, \ldots, X_{i-1})$$

- For example, $P(X_1, X_2, X_3) = P(X_1)P(X_2|X_1)P(X_3|X_1, X_2)$

Independence and Conditional Independence

- Two random variables X and Y are **independent** iff:

$$\forall x \in X, \forall y \in Y, P(X = x, Y = y) = P(X = x)P(Y = y)$$

- Two random variables X and Y are **conditionally independent** given a random variable Z iff

$$\forall x \in X, \forall y \in Y, \forall z \in Z, P(X = x, Y = y|Z = z) = P(X = x|Z = z)P(Y = y|Z$$

- Notation $X \perp Y$ means that X and Y are independent
- $X \perp Y|Z$ means that X and Y are conditionally independent given Z

Expectation, Variance and Covariance

- Expectation or expected value of some function $f(x)$ wrt a probability distribution P is the average or mean value that f takes on when x is drawn from P
- For discrete variables, $E_{X\sim P}[f(x)] = \sum_x P(x)f(x)$
- For continuous variables, it is computed as

$$E_{X\sim p}[f(x)] = \int p(x)f(x)dx$$

- Expectations are liner, for example

$$E_X[\alpha f(x) + \beta g(x)] = \alpha E_X[f(x)] + \beta E_X[g(x)] \quad \alpha, \beta \in \mathbb{R}$$

- **Variance** gives a measure of how much the values of a function of a rv X vary as we sample different values of x from its probability distribution
- $Var(f(x)) = E\left[(f(x) - E[f(x)])^2\right]$
- When variance is low, the values of $f(x)$ cluster near its mean value
- Standard Deviation $= \sqrt{Variance}$
- Standard deviation is the Euclidean distance between the values of $f(x)$ and its mean

- **Covariance** gives some sense of how much two values are linearly related to each other, as well as the scale of these variables

$$Cov(f(x), g(y)) = E\left[(f(x) - E[f(x)])(g(y) - E[g(y)])\right]$$

- High covariance mean that the value change very much and are both far away from their respective mean at the same time
- Positive sign of covariance means both variables tend to take on high values simultaneously
- Negative sign of covariance means, one variable tends to take high value at the times that other takes on low values and vice-versa
- Measures such as *correlation* normalize the contribution of each variable in order to measure only how much the variables are related, rather than also being affected by the scale of the separate variables

- Covariance and dependence are related
- Two variables that are independent have zero covariance
- Two variables that are dependent have non-zero covariance
- For two variables to have zero covariance, there must be no linear dependence between them
- Independence is a stronger requirement than zero covariance, because independence also excludes nonlinear relationship
- Covariance matrix of a random variable $X \in \mathbb{R}^n$ is an $n \times n$ matrix such that

$$Cov(X)_{ij} = Cov(X_i, X_j)$$

- The diagonal elements of the covariance give the variance:

$$Cov(X_i, X_i) = Var(X_i)$$

Common Probability Distribution

Common probabilty distribution in context of machine learning are

- Bernoulli Distribution
- Multinoulli Distribution
- Gaussian Distribution
- Exponential and Laplace Distribution
- Dirac and Emprirical Distribution
- Mixtures of Distribution

Bernoulli Distribution

- It is distribution over a single binary random variable
- It is controlled by a single parameter $\phi \in [0, 1]$, which gives the probability of the random variable being equal to 1
- Properties of Bernoulli distribution
  - $P(X = 1) = \phi$
  - $P(X = 0) = 1 - \phi$
  - $P(X = x) = \phi^x(1 - \phi)^{1-x}$
  - $E_X[X] = \phi$
  - $Var_X(X) = \phi(1 - \phi)$

Multinoulli Distribution

- Multinoulli or categorical distribution is a distribution over a single discrete variable with k different states
- It is parameterized by a vector $\boldsymbol{p} \in [0, 1]^{k-1}$, where p_i gives the probability of the ith state
- Final kth state's probability is given by $1 - \mathbf{1}^T\boldsymbol{p}$
- We must contain $\mathbf{1}^T\boldsymbol{p} \leq 1$
- Multinoulli distributions are used to refer to distributions over categories of objects
- Bernoulli and multinoulli distributions are sufficient to describe any distribution over their domain
- They model discrete variables for which it is feasible to simply enumerate all of the states

Gaussian Distribution

- It is the most commonly used distribution over real numbers (also called normal distribution)

$$\mathcal{N}(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

- Two parameters $\mu \in \mathbb{R}$ and $\sigma \in (0, \infty)$ control the normal distribution
- For frequent PDF evaluation, we use $\beta \in (0, \infty)$ to control precision (or inverse variance of distribution)

$$\mathcal{N}(x; \mu, \beta^{-1}) = \sqrt{\frac{\beta}{2\pi}} \exp\left(-\frac{1}{2}\beta(x-\mu)^2\right)$$

- Without any prior information about the distribution, normal distribution is a good default choice for two reasons
  1. Many distributions are truly close to being normal distributions (courtesy CLT)
  2. Normal distribution encodes the maximum amount of uncertainty over the real numbers (as being the one that inserts least amount of prior knowledge in the model)

- Normal distribution generalizes to $\mathbb{R}^n$, and known as **multivariate normal distribution**

$$\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^n det(\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

- For frequent PDF computation, we can use precision matrix $\boldsymbol{\beta}$

$$\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\beta}^{-1}) = \sqrt{\frac{det(\boldsymbol{\beta})}{(2\pi)^n}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\beta}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

Exponential and Laplace Distributions

- In context of deep learning, we often want to have a probability distribution with a sharp point at $x = 0$
- To accomplish this, we can use exponential distribution

$$p(x; \lambda) = \lambda \mathbf{1}_{x \geq 0} \exp(-\lambda x)$$

- Exponential distribution uses the indicator function $\mathbf{1}_{x \geq 0}$ to assign probability zero to all negative values of x
- Laplace distribution allows to place a sharp peak at an arbitrary point μ

$$Laplace(x; \mu, \gamma) = \frac{1}{2\gamma} \exp\left(-\frac{|x - \mu|}{\gamma}\right)$$

Dirac and Emprirical Distribution

- When we wish to specify that all of the mass in the probability distribution cluster around a single point, we can use Dirac delta function $\delta(x)$,

$$p(x) = \delta(x - \mu)$$

- Dirac delta function is defined to be zero-valued everywhere except 0, yet integrates to 1
- Dirac delta function is a kind of generalized function that is defined in terms of its properties when integrated
- By defining $p(x)$ to be δ shifted by $-\mu$, we obtain an infinitely narrow and infinitely high peak of probability mass at $x = \mu$

- A common use of Dirac delta distribution is as a empirical distribution

$$\hat{p}(x) = \frac{1}{m}\sum_{i=1}^{m}\delta(x - x_i)$$

  which puts probability mass $1/m$ on each of the m points $x_1, \ldots, x_m$
- Dirac delta distribution is only necessary to define the empirical distribution over continuous variables
- For discrete variables, an empirical distribution can be conceptualized as a multinoulli distribution, with a probability associated to each possible input value that is equal to the empirical frequency of that value in the training set

Mixtures of Distributions

- It is common to define probability distribution by combining other simpler probability distributions
- One common way of combining distributions is to construct a mixture distribution
- A mixture distribution is made up of several component distributions
- On each trial, choice of which component distribution generates the sample is determined by sampling a component identity from a multinoulli distribution

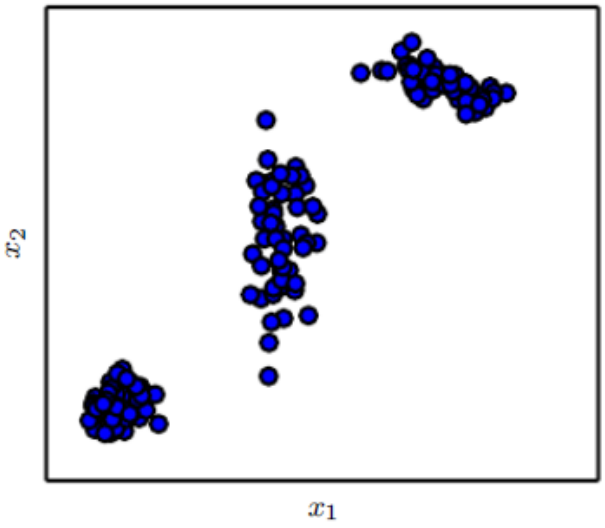
$$P(X) = \sum_i P(c = i)P(X|c = i)$$

  , where $P(c)$ is multinoulli distribution over component identities
- Example: empirical distribution over real-valued variable is a mixture distribution with one Dirac component for each training example

Latent Variable

- A latent variable is a random variable that we cannot observe directly
- Component identity variable c of the mixture model provides an example
- Latent variables may be related to X, $P(X, c) = P(X|c)P(c)$
- Distribution $P(c)$ over latent variable and distribution $P(X|c)$ relating latent to the visible variable determines the shape of distribution $P(X)$

- Gaussian mixture model is a common type of mixture model, in which the components $p(X|c = i)$ are Gaussians
- Each component has a separate parameterized mean $\mu^{(i)}$ and covariance $\Sigma(i)$
- Gaussian mixture model is a universal approximator of densities, in the sense that any smooth density can be approximated with any specific, non-zero amount of error by a Gaussian mixture model with enough component

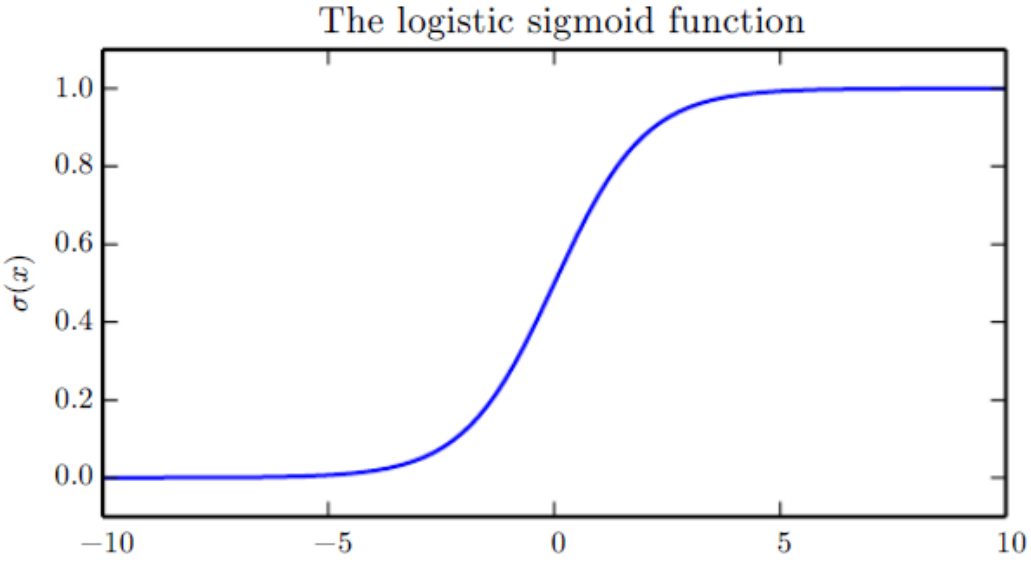


Properties of Common Functions

- Logistic sigmoid

$$\sigma(x) = \frac{1}{1 + \exp(-x)}$$

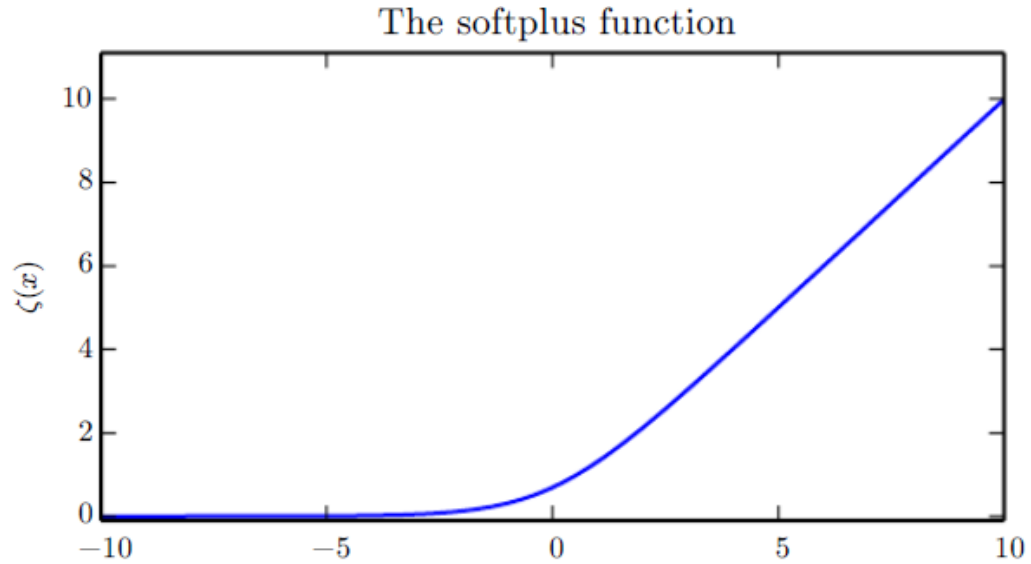
- Commonly used to parametrize Bernoulli distribution



- Softplus function, which is defined as

$$\zeta(x) = \log\left(1 + \exp(x)\right)$$

- It is a smoothed or "softened" version of $x^+ = max(0, x)$
- Softplus can useful for producing σ or β parameter of a normal distribution



Some Useful Properties

- $\sigma(x) = \dfrac{\exp(x)}{\exp(x) + \exp(0)}$
- $\dfrac{d\sigma(x)}{dx} = \sigma(x)(1 - \sigma(x))$
- $1 - \sigma(x) = \sigma(-x)$
- $\log \sigma(x) = -\zeta(-x)$
- $\dfrac{d\zeta(x)}{dx} = \sigma(x)$
- $\forall x \in (0, 1), \sigma^{-1}(x) = \log\left(\dfrac{x}{1 - x}\right)$
- $\forall x > 0, \zeta^{-1}(x) = \log\left(\exp(x) - 1\right)$
- $\zeta(x) = \displaystyle\int_{\infty}^{x} \sigma(y)dy$
- $\zeta(x) - \zeta(-x) = x$

Change of Variables

- Suppose we have two random variables, X and Y such that $Y = g(X)$, where $g(.)$ is an invertible, continuous and differentiable transformation
- One might expect that $p_y(Y) = p_x(g^{-1}(Y))$, which is not true
- For example, suppose we have scalar random variables X and Y, such that $Y = \dfrac{X}{2}$ and $X \sim U(0, 1)$
- If we use the rule $p_y(Y) = p_x(2Y)$ then p_y will be 0 everywhere except the interval [0,1/2] and it will be 1
- This means $\displaystyle\int p_y(Y)dy = \frac{1}{2}$, which violates the definition of probablity distribution

- This is because it fails to account for distortion of space introduced by the function g
- We know that the probability of x lying in an infinitesimal small region with volume δx is given by $p(x)\delta x$
- Since g can expand and contract space, the infinitesimal volume surrounding x in x space may have different volume in y space
- To fix this issue, we need to preserve the property

$$|p_y(g(x))dy| = |p_x(x)dx|$$

- On solving, $p_y(y) = p_x(g^{-1}(y))|\frac{\partial x}{\partial y}|$ or equivalently

$$p_x(x) = p_y(g(x))|\frac{\partial g(x)}{\partial x}|$$

- In higher dimension, the derivative generalizes to the determinant of Jacobian matrix
- For real-valued vectors $\boldsymbol{x}, \boldsymbol{y}$,

$$p_x(\boldsymbol{x}) = p_y(g(\boldsymbol{x}))\left|det\left(\frac{\partial g(\boldsymbol{x})}{\partial \boldsymbol{x}}\right)\right|$$

Information Theory

- Information theory quantifies how much information is present in a signal
- Basic intution behind information theory is that learning that an unlikely event has occured is more informative than learning that a likely event has occured
- A message "the sun rose this morning" is so uninformative
- A message "there was a solar eclipse this morning" is very informative
- Self-information of an event $X = x$ is defined as $I(x) = -\log P(x)$
- Unless specified, we use log to mean natural log, with base e
- $I(x)$ has units in **nats**
- One nat is the amount of information gained by observing an event of probability $\frac{1}{e}$

- Self-information deals only with a single outcome
- We can quantify uncertainty in an entire probability distribution using Shannon entropy

$$H(x) = E_{X\sim P}[I(x)] = -E_{X\sim P}[\log P(x)]$$

, also denoted as $H(P)$
- Shannon entropy of a distribution is the expected amount of information in an event drawn from that distribution
- Entropy gives lower bound on the number of bits (for base=2) needed on average to encode symbols drawn from a distribution P
- Distributions that are nearly deterministic have low entropy, and vice versa
- When X is continuous, the Shannon entropy is known as differential entropy

- Consider two separate probability distributions $P(X)$ and $Q(X)$ over the same random variable X
- We want to measure how different these two distributions are using *Kullback-Leibler (KL) divergence*

$$D_{KL}(P \parallel Q) = E_{X\sim P}\left[\log \frac{P(x)}{Q(x)}\right] = E_{X\sim P}[\log P(x) - \log Q(x)]$$

- KL divergence is non-negative, and it is 0 iff P and Q are the same distribution in case of discrete variables, or "almost everywhere" in case of continuous variables
- KL divergence measures some sort of distance between two distributions
- It is not true distance because it is not symmetric i.e., $D_{KL}(P \parallel Q) \neq D_{KL}(Q \parallel P)$
- Choice of KL direction is problem-dependent
- Cross-entropy

$$\begin{aligned} H(P,Q) &= H(P) + D_{KL}(P \parallel Q) \\ &= -E_{X\sim P} \log Q(x) \end{aligned}$$

Structured Probabilistic Models

- ML algorithms involve probability distributions over a large number of random variables
- These probability distributions involve direct interactions between few variables
- Using a single function to describe the entire joint probability distribution can be very inefficient (computationally and statistically)
- Instead we can split a probability distribution into many factors that we multiply together

- For example, suppose three rv: a, b and c such that a influences the value of b and b influences the value of c but $a \perp c|b$
- We can represent probability distribution over all three variables as product of probability distribution over two variables

$$p(a,b,c) = p(a)p(b|a)p(c|b)$$

- Factorization greatly reduces the number of parameters needed to describe the distribution and hence the cost of representing a distribution
- Such factorization can be represented using **graphs**

- Graph is a set of vertices connected to each other with edges
- When factorization of a probability distribution is represented through graph, it is called a Structured Probabilistic Model (SPM) or graphical model
- Types of SPM: directed and undirected
- Both type use a graph $\mathcal{G}$ in which each node corresponds to a random variable, and an edge means probability distribution to represent interaction between those two random variables
- Directed models represent factorization into conditional probability distributions
- A directed model contains one factor for every random variable X_i in the distribution, and that factor consist of conditional distribution over X_i given the parents of X_i, denoted as $Pa_{\mathcal{G}}(X_i)$

$$p(X) = \prod_i p(X_i | Pa_{\mathcal{G}}(X_i))$$

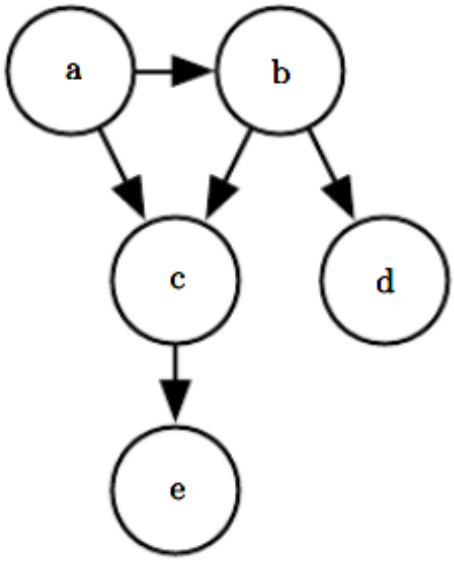


Figure 1 : Graph corresponds to probability distribution $p(a,b,c,d,e) = p(a)p(b|a)p(c|a,b)p(d|b)p(e|c)$

- Undirected models represent factorization into a set of functions but not probability distribution of any kind
- Set of nodes that are all connected to each other in $\mathcal{G}$ are called a clique
- Each clique $\mathcal{C}^{(i)}$ is associated with a factor $\phi^{(i)}(\mathcal{C}^{(i)})$
- These factors are functions not probability distribution
- The output of each factor must be non-negative

- Probability of a configuration of random variable is proportional to the product of all these factors
- Product may exceed 1 so that we need to normalize it

$$p(X) = \frac{1}{Z} \prod_i \phi^{(i)}(\mathcal{C}^{(i)})$$

where Z is sum or integral over all states of the product of ϕ functions

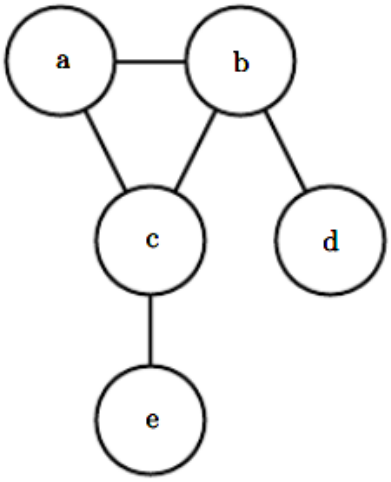


Figure 2 : Undirected model corresponding to probability distribution that can be factored as $p(a,b,c,d,e) = \frac{1}{Z}\phi^{(1)}(a,b,c)\phi^{(2)}(b,d)\phi^{(3)}(c,e)$