# Chapter 3 Probability and Information Theory

Probability Theory is a mathematical framework for representing uncertain statements.

3 possible sources of uncertainty:

1. Inherent stochasticity in the system being modeled.
2. Incomplete observability - the Monty Hall problem.
3. Incomplete modeling

More practical to use simple but uncertain tool rather than a more complex but certain one.

We use a degree of belief with 1 indicating absolute certainty of occuring and 0 indicates absolute certainty of not occurring.

Probability can be seen as the extension of logic to deal ē uncertainty.

Random variables can be discrete or continuous.

Probability Distributions - depends on use of continuous or discrete variables

# Probability Mass Function PMF

A PMF describes the probabilities distribution with the use of discrete variables.

Joint Probability distribution - a distribution over many variables at one time

To be a PMF on a random variable $x$, a function $P$ must satisfy the following properties:

- The domain of $P$ must be the set of all possible states of $x$
- $\forall x \in X, 0 \leq P(x) \leq 1$
- $\sum_{x \in X} P(x) = 1$, this is referred to being normalized

# Probability Density Function PDF

A PDF describes probability distribution with the use of continuous variables.

To be a PDF a function $p$ must satisfy the following properties:

- The domain of $p$ must be the set of all possible states of $x$.
- $\forall x \in X, p(x) \geq 0$, note that $p(x) \leq 1$ is not required
- $\int p(x)\,dx = 1$ (all of the area under the curve = 1)

A PDF of $p(x)$ does not give a probability of a specific state, but rather the probability of landing inside an infinitesimal region. a volume of $\delta x$ is given by $p(x)\delta x$. Specifically the probability that $x$ lies in some set $\mathbb{S}$ is given by the integral of $p(x)$ over that set.

$$\int_a^b p(x)\,dx$$

Marginal Probability distribution - The probability distribution over a subset

## Conditional Probability

$P(A \mid B)$ - probability of A given that B has

already occured.

Example: $P(A|B)$ $\quad S = \{1, 2, 3, 4, 5, 6\}$

$A = \{1, 3, 5\}$ $\quad B = \{3, 4, 5\}$

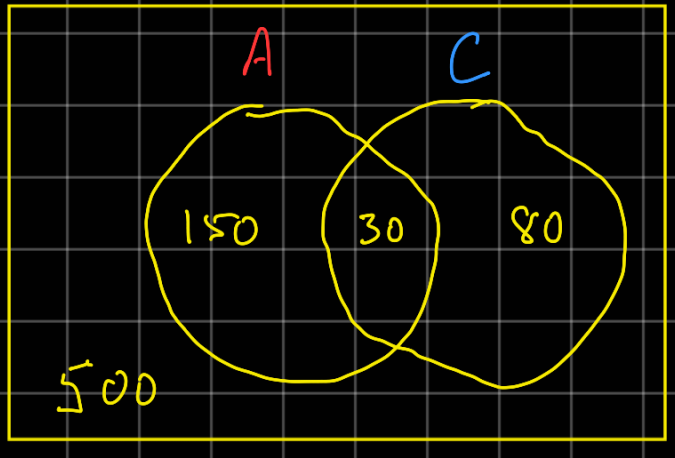
'ask yourself how much of A is in B, 2/3, 2 elements of A that are in 3 elements of B

$P(A|B) =$

$A \cap B = \frac{2}{6}$

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)} = \frac{\frac{2}{6}}{\frac{3}{6}} = \frac{2}{3}$$

---




$P(A) = \frac{150}{500} = \frac{3}{10}$

$P(C) = \frac{80}{500} = \frac{4}{25}$

$P(A|C) = \frac{30}{80}$ or $\frac{\frac{30}{500}}{\frac{80}{500}} = \frac{30}{80}$

$P(C|A) = \frac{30 \text{ of C in A}}{150}$ or $\frac{\frac{30}{500}}{\frac{150}{500}} = \frac{30}{150}$

C and A = 30

Probability = $\frac{30}{500}$

$$P(C|A) = \frac{P(C \text{ and } A)}{P(A)} = \frac{150}{500}$$

Bayes Theorem $P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \qquad P(B|A) = \frac{P(B \cap A)}{P(A)}$$

$$P(A \cap B) = P(B \cap A)$$

$$\frac{P(A|B)\,\cancel{P(B)}}{\cancel{P(B)}} = \frac{P(B|A)\,P(A)}{P(B)}$$

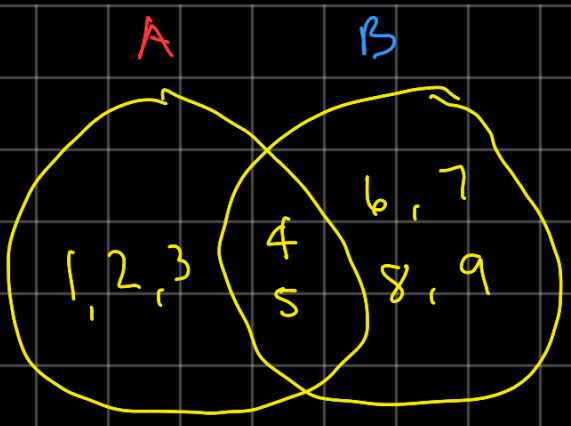
$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

Example:

$A = \{1, 2, 3, 4, 5\}$ $\quad B = \{4, 5, 6, 7, 8, 9\}$

$S = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$

$$P(A|B) = \frac{2}{6} = \frac{1}{3}$$




$$P(A) = \frac{5}{9}$$

$$P(B) = \frac{6}{9}$$

$$P(B|A) = \frac{2}{5}$$

How much of B is in A? →

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)} = \frac{\frac{2}{5} \cdot \frac{5}{9}}{\frac{6}{9}} = \frac{\frac{2}{9}}{\frac{6}{9}} = \frac{1}{3}$$

# Independence and Conditional Independence

although two conditions, A and B, may be thought of as independent (2 separate coin tosses), a third variable C (whether a biased coin is used) may lead to a conditional independent situation.

Independent $P(A) = P(A|B)$ $\quad A \perp B$

Conditional Independent $P(A|C) = P(A|B, C)$

$A \perp B \mid C$

## Expectation, Variance, Covariance

Expectation (expected value) is the average value of some random variable X. It is the calculated probability weighted sums of values that can be drawn.

$E[X] = sum(X_1 * P_1, X_2 * P_2, X_3 * P_3 \dots X_n * P_n)$

If the probability is likely for all values:

$E[X] = sum(X_1, X_2, X_3 \dots X_n) * \frac{1}{n}$

For discrete variables

$$E_{x \sim P}[f(x)] = \sum_x P(x) f(x)$$

For continuous variables

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

Variance - the variance of some random variable $X$ is a measure of how much values in the distribution vary on average with respect to the mean. Variance is calculated as the average squared difference from the expected value

$$Var[X] = E[(X - E[X])^2]$$

$$Var(f(x)) = E[(f(x) - E(f(x))^2]$$

When variance is low, the values of $f(x)$ cluster near their expected value.

Standard Deviation - The square root of variance

Covariance - the measure of the joint probability of two random variables. It describes how the two variables change together. It gives some sense of how much two values are linearly related to each other, as well as the scale of the variables

$$Cov(X, Y) = E[(X - E[X]) * (Y - E[Y])]$$

where $E[X]$ is the expected value for $X$ and $E[Y]$ is the expected value for $Y$

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

High covariance values mean that values change very much together, if covariance is positive, then both variables take on relatively high values simultaneously, if covariance is negative, then one variable takes on a relatively high value while the other takes on a relatively low value.

Covariance and dependence are related but are distinctively different. It is possible for 2 variables to be dependent but have zero covariance.

Independence is a stronger requirement than zero covariance, zero covariance means no linear relationships, independence also means no nonlinear relationships.

## Common Probability Distributions

### Bernoulli Distribution

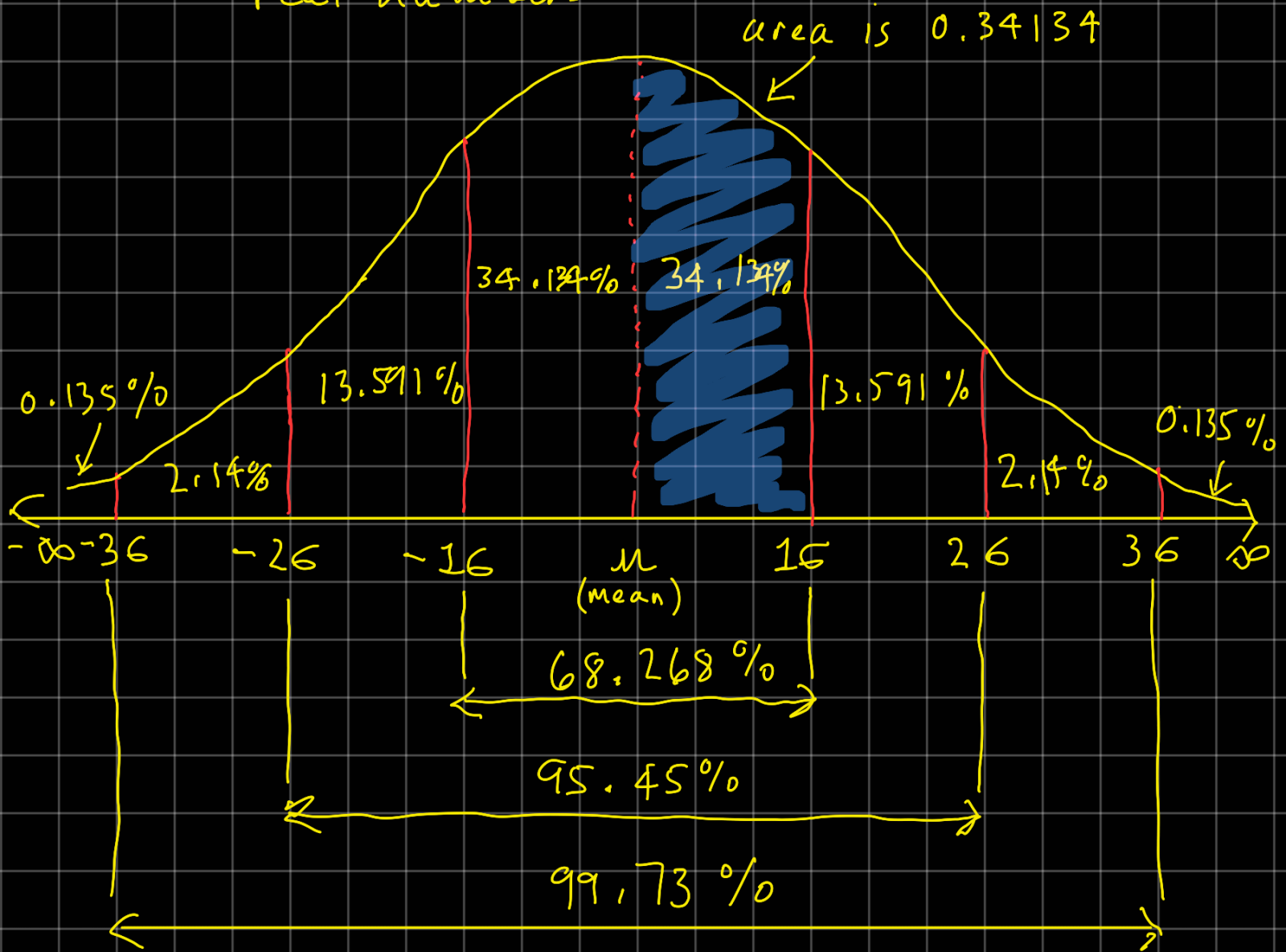
distribution over a single binary random variable

## Multinoulli Distribution (Categorical distribution)

distribution over a single discrete variable with k different states, where k is finite.

## Gaussian Distribution (normal distribution)

the most commonly used distribution over real numbers




$\mu$ = population mean = $E[X]$

$\sigma$ = Standard deviation

$Area_{total} = 1$ 100%

$$P(a \leq x \leq b) = \int_a^b \frac{e^{-(x-\mu)^2/2\sigma^2}}{\sigma\sqrt{2\pi}}\, dx$$

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)} \quad \text{for } -\infty < x < \infty$$

## Exponential and Laplace Distribution

In the context of deep learning, we often want to have a probability distribution with a sharp point at $x = \phi$, we use exponential distribution. To place a sharp point (peak) of probability mass at an arbitrary point $\mu$, we use Laplace distribution

## Dirac Distribution and Empirical Distribution

To specify that all of the mass in a probability distribution clusters around a single point, This can be accomplished by defining a PDF using the Dirac Delta function, $\delta(x)$, which is a component of Empirical distribution

# Useful Properties of Common Functions

phi ↓ $\phi$

Logistic sigmoid - commonly used to produce the $\phi$ parameter of Bernoulli distribution because of its range is $(0, 1)$

$$f(x) = \frac{1}{1 + e^{-x}}$$

- It matches the feature space into probability functions
  - when $x$ is really large (towards $\infty$), output will be close to 1
  - when $x$ is really small (towards $-\infty$), output will be close to 0
  - when $x$ is 0, output will be $\frac{1}{2}$

- a nonlinear relationship ensure that most points will be either close to 0 or 1, no middle ambiguous zone (changes output to binary 0 or 1)

- function is differentiable

Softplus function - useful for producing the $\beta$ or $\sigma$ parameter of a normal distribution because its range is $(0, \infty)$. Softplus is a smooth approximation to the ReLU function and can be used to constrain the output to always be positive.

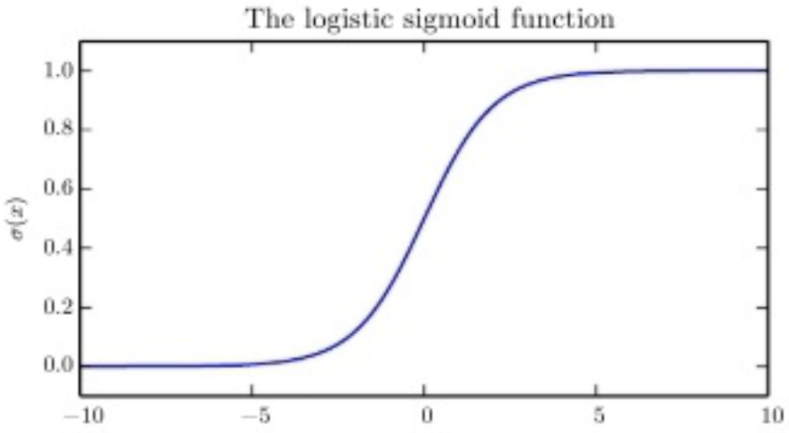



Figure 3.3: The logistic sigmoid function.

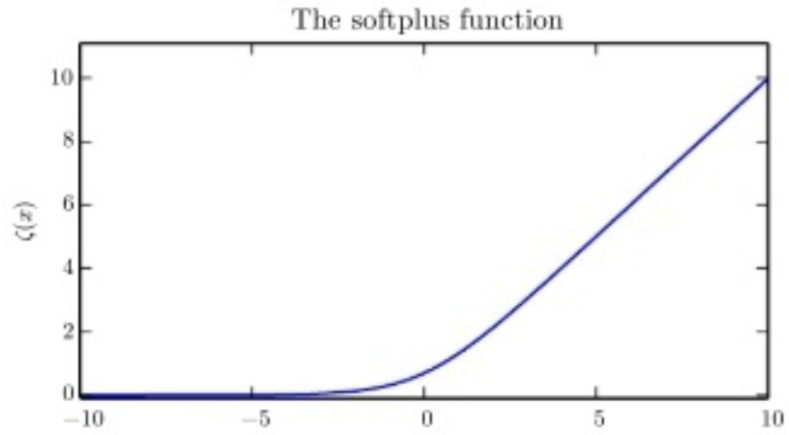



Figure 3.4: The softplus function.

Information Theory - a branch of applied mathematics that revolves around quantifying how much information is present in a signal. It tells us how to design optimal codes and calculate the expected length of messages sampled from specific probability distributions using various encoding schemes.

The basic intuition behind Information Theory is that learning that an unlikely event has occured is more informative than learning that a likely event has occured.

- Likely events should have low information content, and in the extreme case, events that are guaranteed to happen should have no information content.
- Less likely events should have higher information content
- Independent events should have additive information. example: tossed coin 2 heads twice should convey twice as much information than heads come up once.

Self-information of event $X = x$

$$I(x) = -\log P(x)$$

$I(x)$ is using natural log and written as nats
one nat = amount of information gained by observing an event of probability $\frac{1}{e}$

using $\log_2$ → Shannons or bits

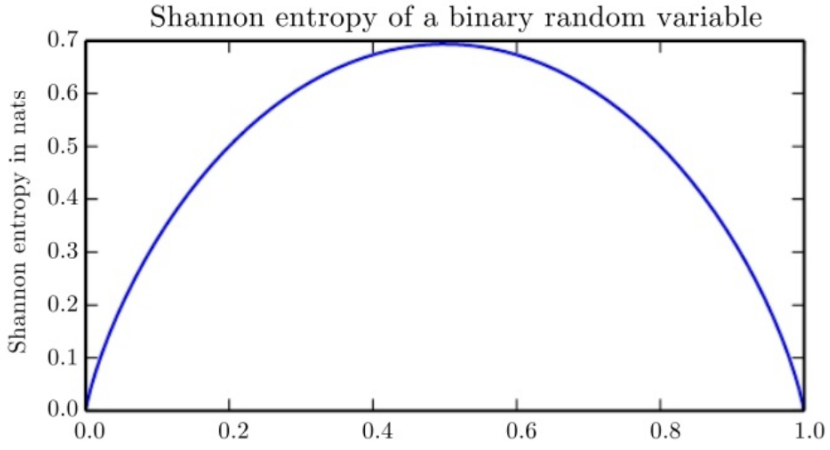



Figure 3.5: This plot shows how distributions that are closer to deterministic have low Shannon entropy while distributions that are close to uniform have high Shannon entropy. On the horizontal axis, we plot $p$, the probability of a binary random variable being equal to 1. The entropy is given by $(p-1)\log(1-p) - p\log p$. When $p$ is near 0, the distribution is nearly deterministic, because the random variable is nearly always 0. When$p$ is near 1, the distribution is nearly deterministic, because the random variable is nearly always 1. When $p = 0.5$, the entropy is maximal, because the distribution is uniform over the two outcomes.

Shannon entropy:

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

Distributions that are nearly deterministic (where the outcome is nearly certain) have low entropy; distributions that are closer to uniform have high entropy.

Structured Probabilistic Models

Instead of using a single function to represent a probability distribution, a probability distribution can be split into many factors that are multiplied together.

These can be represented in a graph

Directed graph:

These represent factorizations into conditional probability distributions

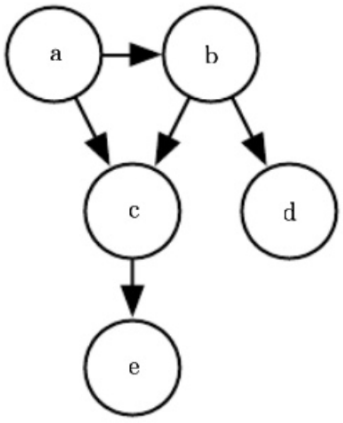



Figure 3.7: A directed graphical model over random variables a, b, c, d and e. This graph corresponds to probability distributions that can be factored as

$$p(\mathrm{a}, \mathrm{b}, \mathrm{c}, \mathrm{d}, \mathrm{e}) = p(\mathrm{a})p(\mathrm{b} \mid \mathrm{a})p(\mathrm{c} \mid \mathrm{a}, \mathrm{b})p(\mathrm{d} \mid \mathrm{b})p(\mathrm{e} \mid \mathrm{c}). \tag{3.54}$$

This graph allows us to quickly see some properties of the distribution. For example, a and c interact directly, but a and e interact only indirectly via c.

Undirected Graph:

These represent factorizations set into functions; unlike directed graphs, these functions are usually not probability distributions of any kind. Any set of nodes that are all connected in G is called a Clique.

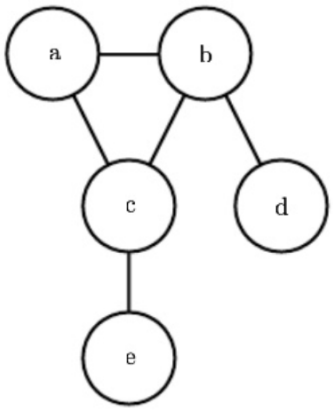



Figure 3.8: An undirected graphical model over random variablesa, b, c, d and e. This graph corresponds to probability distributions that can be factored as

$$p(\mathrm{a}, \mathrm{b}, \mathrm{c}, \mathrm{d}, \mathrm{e}) = \frac{1}{Z} \phi^{(1)}(\mathrm{a}, \mathrm{b}, \mathrm{c}) \phi^{(2)}(\mathrm{b}, \mathrm{d}) \phi^{(3)}(\mathrm{c}, \mathrm{e}). \tag{3.56}$$

This graph allows us to quickly see some properties of the distribution. For example,a and c interact directly, but a and e interact only indirectly via c.