Information theory quantifies amount of uncertainty in a distribution

Uncertainty is caused by inherent stochasticity, incomplete observability (e.g there is one correct answer but it is unknown), and incomplete modeling
It is more practical to use a simple but uncertain rule

Frequentist probability related to rates at which events occur
Bayesian probability is related to qualitative levels of certainty such as degree of belief
Bayesian and frequentist probabilities are treated the same since common sense assumptions imply the same axioms control both kinds of probability

Random Variables
Probability distribution is a description of how likely a random variable is to take on each of its possible states ; description of distributions varies based on whether the variables are discrete or continuous

Probability mass function - distribution over discrete vars ; indicated by capital P
Different variables have different probability mass functions (e.g P(x) != P(y))

The probability that x is *x* is denoted as $P(x)$ where *x* is a particular value
Sometimes we write the name of the random var explicitly P(x = *x*) and sometimes we define a variable first and then use ~ to specify which distribution it follows (e.g x ~ P(x) is x drawn from a particular prob. mass function)

Probability distribution over multiple variables is a joint prob. distribution

$P(\mathrm{x} = x, \mathrm{y} = y)$ denotes the prob. that x = *x* and y = *y* ; we can also write it concisely as P(*x, y*)
To be a probability mass function on a random variable x, the domain of P must be the set of all possible x, and for all x, the probability must be greater than equal to 0 and less than or equal to 1

$\sum_{x \in \mathrm{x}} P(x) = 1$ (total probability cannot be greater than 1) ; this is being normalized

Uniform distribution - equal chance of each state

$$P(\mathrm{x} = x_i) = \frac{1}{k}$$ for all i

Continuous Variables and Probability Dense Functions
When working with continuous random variables, probability distributions are described using probability dense functions (PDFs)

To be a probability dense function, the domain of p must be all possible x, for each x p(x) >= 0 but p(x) is not bounded by 1, and $\int p(x)dx = 1.$
A probability dense function gives the probability of landing inside an infinitesemally small volumetric region with volume δx ; the probability of landing inside this region is given by p(x)* δx

We can integrate the density function to find the probability mass of a set of points (e.g for [a, b] $\int_{[a,b]} p(x)dx.$)
Consider a uniform distribution over an interval of real numbers, as given by the notation $u(x; a, b)$ where a and b are the endpoints and the semicolon indicates that x is parametrized by those endpoints

$u(x; a, b) = 0 \text{ for all } x \notin [a, b]$

$u(x; a, b) = \frac{1}{b-a}$ This integrates to one

We can denote x follows the uniform distribution on [a, b] by writing $\mathrm{x} \sim U(a, b)$

Marginal Probability
Marginal prob. distribution - probability distribution over a subset
If we know P(x, y), we can find P(x), where the only condition is x and y can be anything, with

$$\forall x \in \mathrm{x}, P(\mathrm{x} = x) = \sum_{y} P(\mathrm{x} = x, \mathrm{y} = y).$$

the sum rule

$$p(x) = \int p(x, y)dy.$$

Use integration for continuous variable distributions

Conditional Probability
Condition probability - the probability of one event given another
$P(\mathrm{y} = y \mid \mathrm{x} = x)$ is the probability of y given *x*

$$P(\mathrm{y} = y \mid \mathrm{x} = x) = \frac{P(\mathrm{y} = y, \mathrm{x} = x)}{P(\mathrm{x} = x)}.$$

This can be computed with
Intervention query - computing what would happen if some action were undertaken ; this is the domain of causal modeling

Chain Rule of Conditional Probabilities

Any joint prob. distribution (chance of happening at the same time) over many variables may be decomposed into conditional distributions over one variable

$P(\mathrm{x}^{(1)}, \ldots, \mathrm{x}^{(n)}) = P(\mathrm{x}^{(1)}) \Pi_{i=2}^{n} P(\mathrm{x}^{(i)} \mid \mathrm{x}^{(1)}, \ldots, \mathrm{x}^{(i-1)}).$ is the chain rule of probability

For example, we get

$$\begin{aligned} P(\mathrm{a}, \mathrm{b}, \mathrm{c}) &= P(\mathrm{a} \mid \mathrm{b}, \mathrm{c}) P(\mathrm{b}, \mathrm{c}) \\ P(\mathrm{b}, \mathrm{c}) &= P(\mathrm{b} \mid \mathrm{c}) P(\mathrm{c}) \\ P(\mathrm{a}, \mathrm{b}, \mathrm{c}) &= P(\mathrm{a} \mid \mathrm{b}, \mathrm{c}) P(\mathrm{b} \mid \mathrm{c}) P(\mathrm{c}). \end{aligned}$$

Independence and Conditional Independence

$\forall x \in \mathrm{x}, y \in \mathrm{y},\ p(\mathrm{x} = x, \mathrm{y} = y) = p(\mathrm{x} = x) p(\mathrm{y} = y).$

Two variables x and y are independent if their probability distribution can be expressed as a product of two factors, one involving x, and one involving y

Two random variables x and y are conditionally independent if their conditional probability distribution can be factorized this way for every value of z

$\forall x \in \mathrm{x}, y \in \mathrm{y}, z \in \mathrm{z},\ p(\mathrm{x} = x, \mathrm{y} = y \mid \mathrm{z} = z) = p(\mathrm{x} = x \mid \mathrm{z} = z) p(\mathrm{y} = y \mid \mathrm{z} = z).$

$\mathrm{x} \perp \mathrm{y}$ means independent ; $\mathrm{x} \perp \mathrm{y} \mid \mathrm{z}$ means conditionally independent given z

Expectation, Variance, and Covariance

Expected value of f with respect to P(x) is the average or mean value f takes on when x is drawn from P

$$\mathbb{E}_{\mathrm{x} \sim P}[f(x)] = \sum_{x} P(x) f(x),$$

Or for continuous variables

$$\mathbb{E}_{\mathrm{x} \sim p}[f(x)] = \int p(x) f(x) dx.$$

$\mathbb{E}_{\mathrm{x}}[f(x)].$ gives the average value of f over random values of x

Expectations are linear when alpha and beta are independent of x

$$\mathbb{E}_{\mathrm{x}}[\alpha f(x) + \beta g(x)] = \alpha \mathbb{E}_{\mathrm{x}}[f(x)] + \beta \mathbb{E}_{\mathrm{x}}[g(x)],$$

Variance measures how much values of a function of a random variable x vary as different values of x are sampled from its probability distribution

$$\mathrm{Var}(f(x)) = \mathbb{E}\left[(f(x) - \mathbb{E}[f(x)])^2 \right].$$

When variance is low, values of f cluster near their expected values
Standard deviation is the square root of variance

Covariance gives a sense of how much 2 values are linearly related to each other, and the scale of these variables

$$\operatorname{Cov}(f(x), g(y)) = \mathbb{E}\left[(f(x) - \mathbb{E}\left[f(x)\right])\left(g(y) - \mathbb{E}\left[g(y)\right]\right)\right].$$

High values of covariance mean that the values (f and g) change very much and are far from their means at the same time
If the sign of covariance is positive, both values take on high values simultaneously; if negative, one variable takes on a high value while the other takes on a low value
Correlation normalizes the contribution of each variable to measure how the variables are related without accounting for scale

Covariance and dependence are related but not the same ; independent has zero covariance, and nonzero covariance means two variables are dependent
For two variables to have zero covariance, there must be no linear dependence between them
Independence excludes nonlinear relationships

For example, suppose x is a randomly drawn variable from a uniform distribution from [-1, 1]
s is either 1 or -1, with a 50% chance of being each
y = sx
Despite x and y being dependent, Cov(x, y) = 0

The covariance of a random vector which is a member of R^N is an n by n matrix where

$$\operatorname{Cov}(\mathbf{x})_{i,j} = \operatorname{Cov}(\mathrm{x}_i, \mathrm{x}_j).$$

The diagonal elements of the covariance give the variance

$$\operatorname{Cov}(\mathrm{x}_i, \mathrm{x}_i) = \operatorname{Var}(\mathrm{x}_i).$$

Common Probability Distributions
Bernoulli Distribution is a distribution over a binary variable controlled by a parameter phi which is a member of [0, 1] giving the probability of the random variable being equal to 1

$$P(\mathrm{x} = 1) = \phi$$
$$P(\mathrm{x} = 0) = 1 - \phi$$
$$P(\mathrm{x} = x) = \phi^x(1 - \phi)^{1-x}$$
$$\mathbb{E}_{\mathrm{x}}[\mathrm{x}] = \phi$$
$$\operatorname{Var}_{\mathrm{x}}(\mathrm{x}) = \phi(1 - \phi)$$

The multinoulli or categorical distribution over k finite states is parametrized by a vector p which is a member of $[0, 1]^{k-1}$ (k - 1 elements) where p_i gives the probability of the i-th state

The final k-th state's probability is given by 1 - 1^Tp where $\mathbf{1}^\top \boldsymbol{p} \leq 1.$

Since categorical variables are not associated with numeric values, we don't calculate covariance

Normal / Gaussian distribution is the most common over real numbers

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

mu controls the mean and sigma gives the standard deviation
When evaluating the PDF, we square and invert sigma
We can use a parameter beta to control the precision or inverse variance of the distribution

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

Normal distributions are good defaults

- Independent variables when modelled together and accounting for noise is approxiately normally distributed - central limit theorem
- Out of all probability distributions with the same variance, the normal distribution accounts for the maximum amount of uncertainty

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

Multivariate normal distribution, where epsilon is the positive definite symmetric matrix in R^N

Parameter mu is now vector valued ; epsilon gives the covariance
Since epsilon must be inverted to evaluate the PDF, a more computationally efficient way would be to use a precision matrix beta

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

Covariance matrix often made to be a diagonal matrix
Isotropic gaussian distribution - covariance matrix = scalar times I_N

Exponential distribution has a sharp point at x = 0

$p(x;\lambda) = \lambda\mathbf{1}_{x\geq 0}\exp(-\lambda x)$. $1_{x>=0}$ assigns a probability of 0 to all negative values of x
The Laplace distribution places a sharp peak of probability mass at an arbitrary point mu

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

Mass in the PDF clusters around a single point => Dirac delta function

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

Zero-valued everywhere except x = 0, but it integrates to 1
It is a generalized function defined by its properties when integrated ; limit point of a series of functions that put less mass on non-zero points

It is commonly used as a component of an empirical distribution

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

which puts probability mass 1/m on each of the x points
Dirac delta is only needed to define an empirical distribution for continuous variables
For discrete variables, an empirical distribution is a categorical distribution with a probability function corresponding to the empirical frequency of its inputs in the training set

A mixture distribution is made up of several component distributions
The choice of the particular component distribution generates the sample is determined by sampling a component identity from the multinoulli distribution

$$P(\mathrm{x}) = \sum_{i} P(\mathrm{c} = i) P(\mathrm{x} \mid \mathrm{c} = i)$$

where P(c) is the multinoulli distribution over component identities

The empirical distribution is a mixture distribution with one Dirac component example per training example
Latent variable is a random variable which cannot be directly observed ; latent variables could be related to x through a joint distribution

$$P(\mathrm{x}, \mathrm{c}) = P(\mathrm{x} \mid \mathrm{c}) P(\mathrm{c}).$$

Gaussian mixture - components p(x | c = i) are gaussians ; each component i has a separately paramerized mean and covariance
Some mixtures can have more constraints such as covariances being shared across all components through $\boldsymbol{\Sigma}^{(i)} = \boldsymbol{\Sigma}, \forall i.$

The parameters of a Gaussian mixture specify the prior probability $\alpha_i = P(\mathrm{c} = i)$ given to each component i ; prior means the model's beliefs before it is given x
P(c | x) is a posterior probability, since it is computed after observation of x
Gaussian mixture model is a universal approximator of densities since any smooth density can be approximated with a model with enough components

Useful Properties of Common Functions

Logistic sigmoid - $\sigma(x) = \frac{1}{1 + \exp(-x)}.$ This is commonly used to produce the phi parameter for a Bernoulli distribution since its range is (0, 1)
Logistic sigmoid saturates (barely changes) at very positive or very negative values

Softplus function - $\zeta(x) = \log(1 + \exp(x))$ => a softened version of $x^+ = \max(0, x).$
Used for producting the beta parameter of a normal distribution since its range is (0, inf) and is used when manipulating equations involvling sigmoids

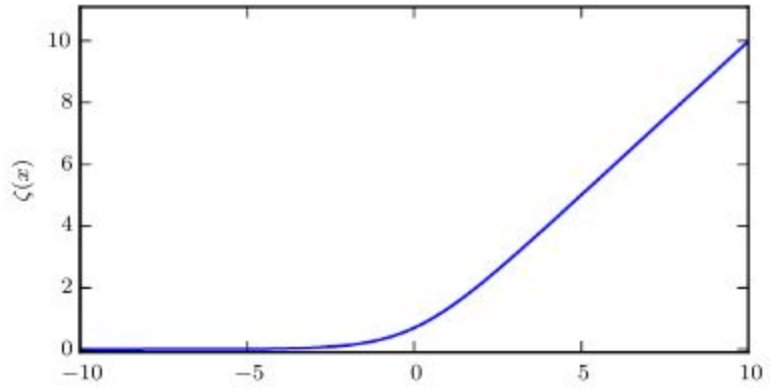

Figure 3.4: The softplus function.

These are the key useful properties

$$\sigma(x) = \frac{\exp(x)}{\exp(x) + \exp(0)} \tag{3.33}$$

$$\frac{d}{dx}\sigma(x) = \sigma(x)(1 - \sigma(x)) \tag{3.34}$$

$$1 - \sigma(x) = \sigma(-x) \tag{3.35}$$

$$\log \sigma(x) = -\zeta(-x) \tag{3.36}$$

$$\frac{d}{dx}\zeta(x) = \sigma(x) \tag{3.37}$$

$$\forall x \in (0, 1),\ \sigma^{-1}(x) = \log\left(\frac{x}{1 - x}\right) \tag{3.38}$$

$$\forall x > 0,\ \zeta^{-1}(x) = \log(\exp(x) - 1) \tag{3.39}$$

$$\zeta(x) = \int_{-\infty}^{x} \sigma(y)dy \tag{3.40}$$

$$\zeta(x) - \zeta(-x) = x \tag{3.41}$$

$\sigma^{-1}(x)$ is called the logit in statistics

Positive part function: $x^+ = \max\{0, x\}$

Since $x^+ - x^- = x$, and $\zeta(x)$ and $\zeta(-x)$ are modelled after the part functions, they share the same properties

Baye's Rule

$P(\mathrm{x} \mid \mathrm{y}) = \frac{P(\mathrm{x})P(\mathrm{y} \mid \mathrm{x})}{P(\mathrm{y})}$. where $P(\mathrm{y}) = \sum_x P(\mathrm{y} \mid x)P(x)$.

Technical Details of Continuous Variables

Continuous variables and PDFs require measure theory

Measure theory tries to characterize sets with paradoxes such as

$p(\boldsymbol{x} \in \mathbb{S}_1) + p(\boldsymbol{x} \in \mathbb{S}_2) > 1$ but $\mathbb{S}_1 \cap \mathbb{S}_2 = \emptyset$. which make use of infinitely precise numbers

Measure theory helps exclude corner cases in R^N

Measure theory defines a set of points which do not meet a certain criteria as measure zero

Measure zero occupies no volume in our dimension

In R^2 a line has measure zero, while a filled polygon has positive measure

Measure theory applies everywhere except for where a set of points has measure zero

Usually the corner cases are negligible

Suppose we have two variables x and y and the relation is given by $\boldsymbol{y} = g(\boldsymbol{x})$ where g is an invertible transformation

The following is *not* true $p_y(\boldsymbol{y}) = p_x(g^{-1}(\boldsymbol{y}))$.

$\mathrm{y} = \frac{\mathrm{x}}{2}$ and $\mathrm{x} \sim U(0, 1)$.

If we say that $p_y(y) = p_x(2y)$, then p_y will be zero everywhere except [0, ½] where it will be 1

This means $\int p_y(y)dy = \frac{1}{2}$ which is an invalid probability distribution

$p(\boldsymbol{x})\delta\boldsymbol{x}$. The probability of x lying in an infinitely small region dx is given by p(x)dx ; however, since g can expand or contract space, the infinitely small volume in x space may be different than in y space

We need to preserve the property

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

from which we derive

$p_y(y) = p_x(g^{-1}(y)) \left|\frac{\partial x}{\partial y}\right|$ and therefore

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

In higher dimensions, the derivative generalizes to theh determinant of the Jacobian matrix, which is defined by the property $J_{i,j} = \frac{\partial x_i}{\partial y_j}$.
Therefore

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

Information Theory
Quantifies how much info is present in a signal
Information theory specifies how to design optimal codes and calculate the length of messages sampled from certain probability distributions
Learning an unlikely event has occurred is more informative than learning a likely event has occurered
To quantify this

- Likely events have lower information content
- Independent events should have additive info

The self information of an event is $I(x) = -\log P(x).$
Log in this case means natural log and therefore the unit is nats
Bits or shannon use log base 2

When x is continuous, some of the discrete properties are lost (e.g an event with unit density has zero info despite the fact it is not guaranteed to occur)
Shannon entropy quantifies amount of uncertainty in an entire probability distribution

$$H(\mathrm{x}) = \mathbb{E}_{\mathrm{x} \sim P}[I(x)] = -\mathbb{E}_{\mathrm{x} \sim P}[\log P(x)].$$

Expected amount of info in an event drawn from that distribution
It gives a lower bound on the number of bits (nat equivalent in base 2) needed to encode symbols from a distributon P

Deterministic distributions with near certain outcomes have low entropy
When x is continuous, Shannon entropy is known as differential entropy

KL divergence measures how different two distributions over the same random variable x are

$$D_{\mathrm{KL}}(P \| Q) = \mathbb{E}_{\mathrm{x} \sim P} \left[\log \frac{P(x)}{Q(x)} \right] = \mathbb{E}_{\mathrm{x} \sim P} \left[\log P(x) - \log Q(x) \right].$$

In terms of discrete variables, it is the extra amount of information needed in nats to send a message contianing symbols from probability distribution P when using a code designed to minimize the length of messages drawn from Q

KL divergence is non-negative and measures "distance"
KL divergence equals 0 if P and Q are the same distribution in the case of discrete variables, or are equal almost everywhere in the case of continuous variables

Cross entropy is similar to KL but lacks the extra term on the left

$$H(P, Q) = H(P) + D_{\mathrm{KL}}(P \| Q),$$

$$H(P, Q) = -\mathbb{E}_{\mathrm{x} \sim P} \log Q(x).$$

Minimizing xentropy with respect to Q is the same as minimizing KL divergence since the extra term in KL is not affected by Q

Expressions of the form 0 log 0 are treated as $\lim_{x \to 0} x \log x = 0.$

Machine learning algorithms use a large amount of random variables but the probability distributions require very few interactions between these variables
A single function to describe the entire joint prob. distribution can be inefficient

We can split a probability distribution into many factors which we can multiply together
Suppose we have 3 random variables: a, b, and c
a influences b and b influences c but a and c are independent given b
We can decompose the joint probability distribution into multiple joint probabilit ydistributions, each considering two variables $p(\mathrm{a}, \mathrm{b}, \mathrm{c}) = p(\mathrm{a}) p(\mathrm{b} \mid \mathrm{a}) p(\mathrm{c} \mid \mathrm{b}).$
Number of parameters is exponentially related to number of variables in the factor
Therefore, multiple smaller factors is much more computationally efficient than one large factor

These kinds of factorizations can be described using graphs, calling it a structured probabilistic model or graphical model
In both directed and undirected structured probabilistic models, each node in a graph corresponds to a random variable and an edge means the prob. distribution is able to represent direct relations between those variables

Directed models use graphs with directed (one-way) edges and they represent factorizations into conditional probability distributions
For every random variable x_i in the distribution, a directed model contains 1 factor consisting of the conditional distribution over x_i given the parents of x_i ($Pa_{\mathcal{G}}(\mathrm{x}_i)$)

$$p(\mathbf{x}) = \prod_i p\left(\mathrm{x}_i \mid Pa_{\mathcal{G}}(\mathrm{x}_i)\right).$$

Undirected models represent factorization into a set of functions which are not probability distributions
Any set of nodes connected to each other in G is called a clique
Each clique $C^{(i)}$ in an undirected model is associated with a factor $\phi^{(i)}(C^{(i)}).$

These factors have to be a non-negative function but they do not have to integrate to 1

Probability of a configuration of random variables is proportional to the product of all factors
Random variable configurations which result in larger factor values are more likely
Divide by a normalizing constant Z which is the sum or integral over all states of the product of the ϕ functions to obtain a normalized prob. distribution

$$p(\mathbf{x}) = \frac{1}{Z} \prod_i \phi^{(i)} \left(\mathcal{C}^{(i)} \right).$$

Any prob. distribution can be described by a directed / undirected graph