Probability and Statistics for Machine Learning

Francesco Pugliese, PhD

pugliese05@gmail.com

Table of contents

Table of contents (cont..)

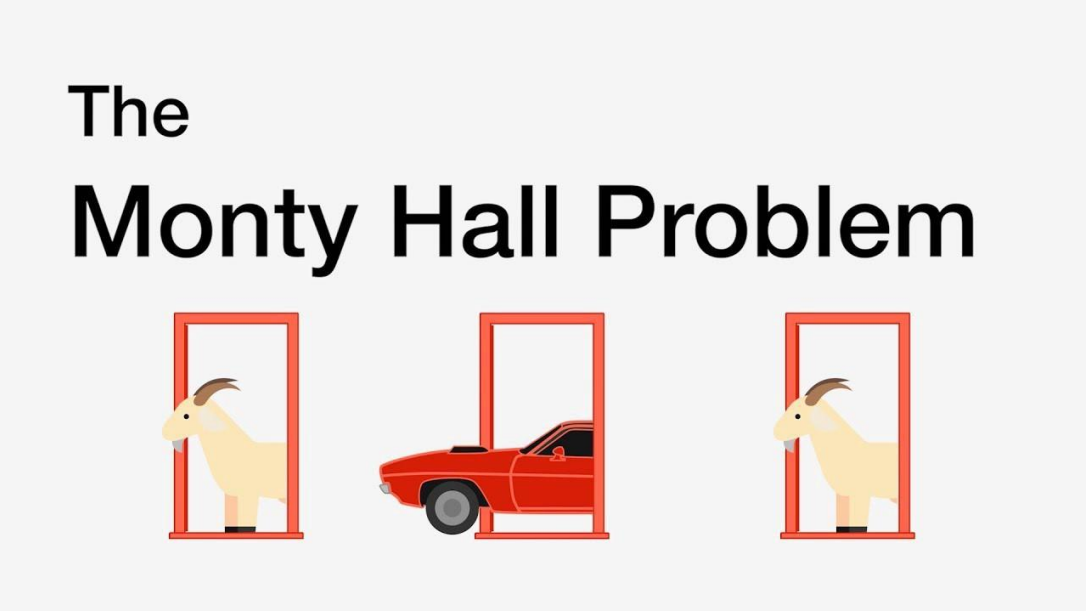
Probability and Information theory

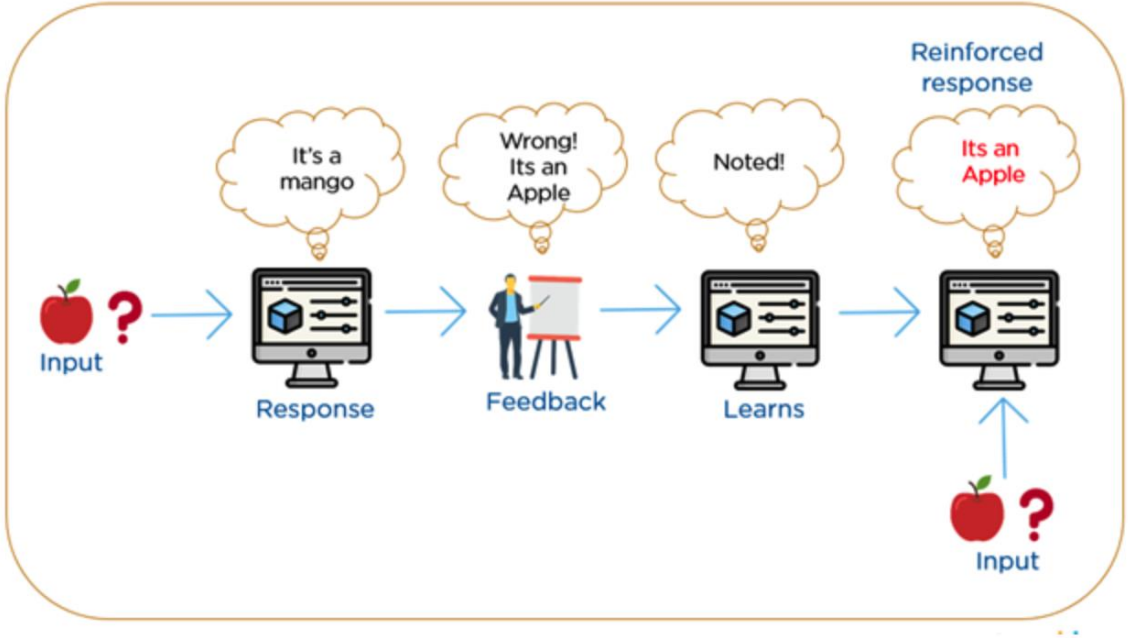
- Probability theory is a mathematical framework for representing **uncertain** events.
- In artificial intelligence applications, we use probability theory in two major ways.
  - Laws of probability tell us how AI should **reason**
  - We can use probability theory to analyze the **behavior** and decisions of AI systems
    - **Example:** Robotics
- **Probability theory** allows us to make uncertain statements and to **reason in the presence of uncertainty**
- **Information theory** enables us to **quantify the amount of uncertainty** in a probability distribution.

Sources of Uncertainty

- Machine learning must always deal with **uncertain** quantities and sometimes **stochastic** (nondeterministic) quantities. Uncertainty and stochasticity can arise from many sources.
- There are three possible sources of uncertainty:
  - **Inherent stochasticity** in the system being modeled.
    - **Example**: Quantum mechanics
  - **Incomplete observability**: Even deterministic systems can appear stochastic when we **cannot observe** all the variables that drive the behavior of the system.
    - Monty Hall problem
    - Reinforcement learning. **Example**: Alpha Go
    - Economic models like **stock** markets
    - Image recognition
  - **Incomplete modeling**: When we use a model that must **discard** some of the information we have observed, the discarded information results in uncertainty in the model's **predictions**.
    - **Example**: A simplified model of brain since we don't have computational resources to simulate entire brain

Sources of Uncertainty





Why probability ?

- In many cases, it is more practical to use a **simple** but uncertain rule rather than a complex but certain one, even if the true rule is deterministic and our modeling system has the fidelity to accommodate a complex rule.
- For example, the simple rule "**Most birds fly**" is **cheap** to develop and is broadly useful, while a rule of the form, "Birds fly, except for very young birds that have not yet learned to fly, sick or injured birds that have lost the ability to fly, flightless species of birds including the cassowary, ostrich and kiwi. . ." is **expensive** to develop, maintain and communicate and, after all this effort, is still **fragile** and prone to **failure**.
- 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

Types of probability

- Frequentist probability:
  - **Frequency** of events
  - Example: The chance of drawing a certain hand in poker
  - Fixed model, different data (We run the same experiments each time with different data)

- Bayesian probability:
  - A degree of **belief**
  - Example: A doctor saying a patient has a 40 percent chance of having a flu
  - Fixed data and different models (We use the same belief to check the uncertainty of different models and update our beliefs)
  - Based on **Bayes rule** which we talk about later in the presentation

Frequentist

[repeat repeat repeat]

Baysesian

[observe, guess, experiment]

Random variables

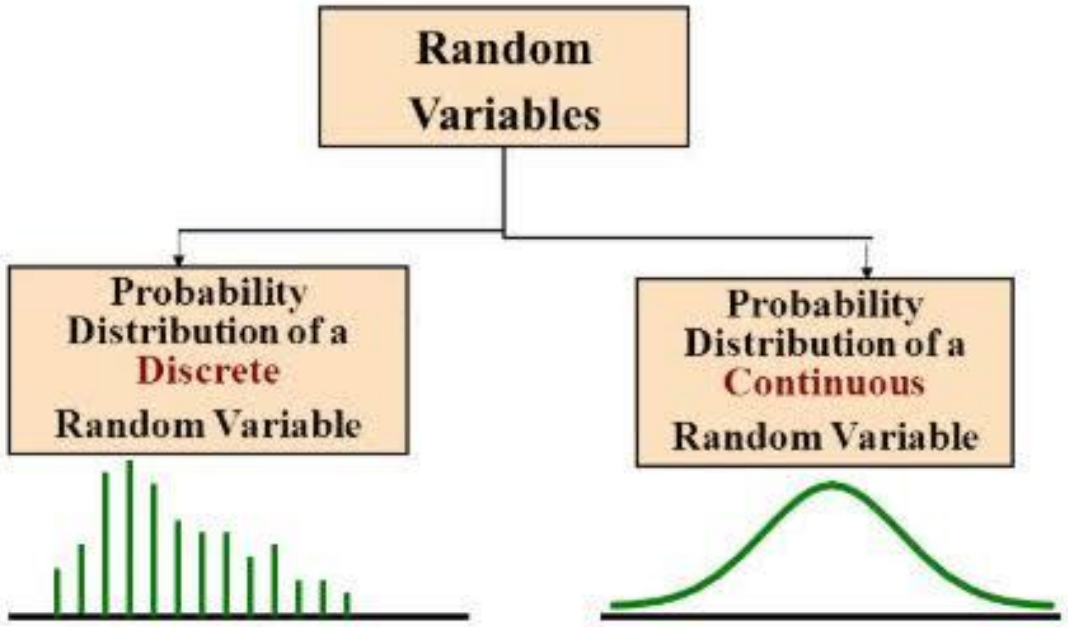
- A random variable is a variable that can take on **different** values randomly.
- On its own, a random variable is just a **description** of the states that are possible; it must be coupled with a **probability distribution** that specifies how **likely** each of these states are.
- Random variables may be discrete or continuous.
    - A **discrete** random variable is one that has a finite or countably infinite number of states. Note that these states are not necessarily the integers; they can also just be named states that are not considered to have any numerical value.
    - A **continuous** random variable is associated with a real value
- **Probability distribution**: description of how **likely** a random variable or set of random variables is to take on each of its **possible states**. The way we describe probability distributions depends on whether the variables are discrete or continuous.

Types of random variables

- Discrete random variable:
  - Finite number of states, not necessarily integers they can also be a **named states** (that are not considered to have any numerical value)
  - **Example**: Coin toss (2 states), Throwing a dice (6 states), Drawing a card from a deck of cards (52 states) etc..,

- Continuous random variable:
  - Must be associated with a **real** value
  - **Example**: Rainfall on a given day (in centimeters), Stock price of a company, Temperature of a given day

Random variables

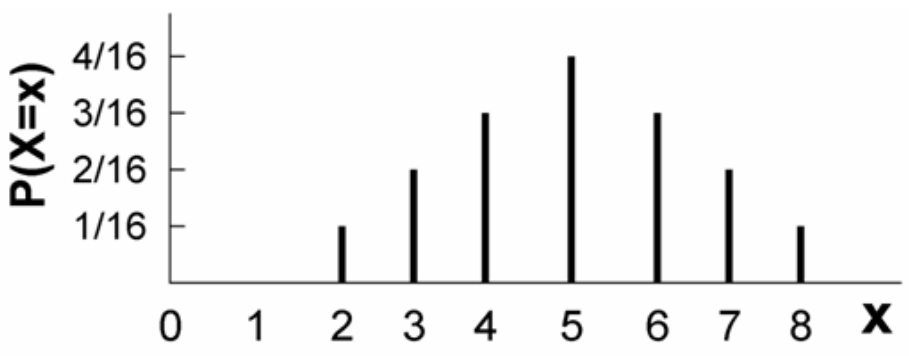


Probability mass function (PMF)

- Probability distribution over discrete random variables is referred to as a **probability mass function**(PMF)
- Maps from a state of a random variable to the probability of that random variable taking on that state.
- The probability that x=x is denoted as P(x), with a probability of 1 indicating that x=x is **certain** and a probability of 0 indicating that x=x is **impossible**.
- Criterion for being a PMF:
  - The domain of P must be the set of all possible states of x
  - $0 \leqq P(x) \leqq 1$
  - Summation of all possible states of P(x) = 1

Probability mass function (PMF)

x	P(x)
2	1/16
3	2/16
4	3/16
5	4/16
6	3/16
7	2/16
8	1/16



Types of Distributions

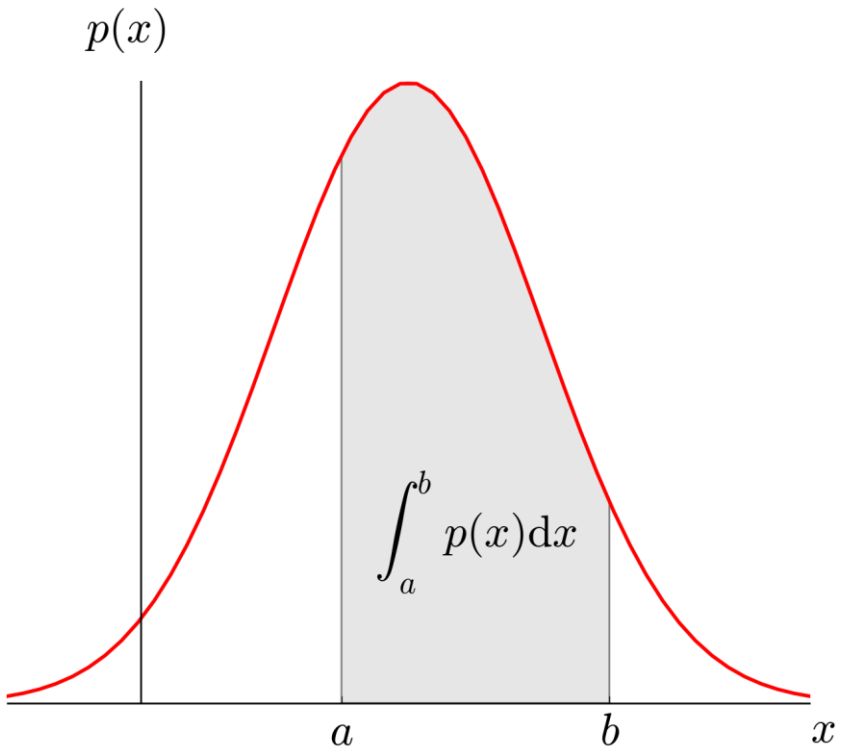
- **Joint probability distribution**:
  - Probability mass function that can act on many variables at the same time.
  - Such a probability distribution over many variables is known as a joint probability distribution.
  - **Example**: P(x=x, y=y) denotes the probability that x=x and y=y simultaneously.
  - We may also write P (x, y) for brevity

- **Uniform distribution**:
  - Each state of the distribution is equally likely
  - P(x = xi) = 1/k where, k is the total number of possible states
  - Completely normalized equal distribution with equally likely states

Probability density function (PDF)

- When working with continuous random variables, we describe probability distributions using a probability density function (PDF) rather than a probability mass function(PMF).
- Statistical expression used in probability theory as a way of representing the range of possible values of a continuous random variable.
- To be a probability density function, 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 we do not require $p(x) \leq 1$.
  - $\int p(x)dx = 1$
- In other words a PDF, p(x) does not give the probability of a specific state directly; instead the probability of landing inside an **infinitesimal region** with volume δx is given by $p(x)\delta x$.
- We can integrate the density function to find the actual probability mass of a set of points. Specifically, the probability that x lies in some set S is given by the integral of p(x) over that set.

Probability density function (PDF)



PDF and Machine Learning

- The Probability Density Function works by **conceptualizing** the probabilities of a continuous random event occurring by defining a **range**, or **interval**.
- For example, if one wanted to calculate the probability that a **specific** temperature, say 70 degrees, will be reached, they may turn to a PMF, as the variable is defined in discrete terms. However, if one wanted to calculate the probability that a temperature **between** 70-75 degrees will be reached, they may use a PDF, as the variable is defined as a **range** with infinite discrete values.
- Since the PDF defines probabilities with intervals, the **probability of a single discrete value is defined as zero**, since it does not have a range.
- A Probability Density Function is a tool used by machine learning algorithms and neural networks that are trained to calculate probabilities from continuous random variables.
- For example, a neural network that is looking at **financial markets** and attempting to guide **investors** may calculate the probability of the **stock market rising 5-10%**. To do so, it could use a PDF in order to calculate the total probability that the continuous random variable range will occur.

Marginal probability

- The probability distribution over a subset of all variables
- Oftentimes we will be working with Marginal probability distributions in AI applications since we don't have all the variables or data points that is one of the sources of uncertainty that we talked about earlier.
- With discrete random variables:
  - If we know P(x,y), we can find P(x) with the **sum rule**, more on this in the next slide
  - P(x=x) = $\sum_y$ P(x=x, y=y) (so P(x) is the summation over all possible y values)
- With continuous random variables:
  - p(x) = ∫ p(x,y) dy

Marginal probability

$f(x, y)$ is defined by the following table

	$x = 0$	$x = 1$	$x = 2$
$y = 1$	0.3	0.2	0.1
$y = 2$	0.1	0.1	0.2
$f_x(x)$	0.4	0.3	0.3

$$f_x(x) = \begin{cases} 0.4 & x = 0 \\ 0.3 & x = 1 \\ 0.3 & x = 2 \end{cases}$$

$$f_y(y) = \begin{cases} 0 & y = 1 \\ & y = 2 \end{cases}$$

$$f_y(1) = P(Y = 1) = 0.6$$

Conditional probability

- The probability of some event, given that some other event has happened

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

- Conditional probability that y = y given x = x
- Widely used in Bayesian inference to form the **beliefs**.
- Its very important to understand marginal and conditional probability because in general we are not going to have **pristine** probability distributions. We are going to be working either with the **subset** of variables or we are going to be adding criterion which will essentially be saying what is the probability of y given x or given a certain criterion

Expectation (of a random variable)

- The expectation of some function f(x) (f(x) is the random variable for x) with respect to some probability distribution P(x) is the mean value f takes on when it is drawn from P.
- Mean or average value
- For discrete random variables:

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

- For continuous random variables:

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

- It can be used to find the **variance** of the distribution and consequently **standard deviation**

Variance (of a random variable)

- The expectation of **squared deviance** of a random variable from its mean is referred to as variance (σ^2).
- Measures how **far** random numbers drawn from a probability distribution P(x) are spread out from their average value(Expectation value).

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

- **Standard deviation**(***σ***): Square root of the variance
- Variance and Standard deviation are two measures of **spread** which are widely used in machine learning. They come all the time because in machine learning we want to know what kind of distributions that our input variables have. so we will find these 2 and understand **patterns** in our data at a single glance.
- These 2 are also a great measure to generating new probability distributions of our data.
    - For example, We can add Gaussian noise to the images to make the AI system generalize the new images in a better way.

Covariance (of a random variable)

- Measure of how much two variables are linearly related to each other.

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

- **High absolute value –** values are both far from their respective means at the same time.
- **Positive** - both variables take on large values simultaneously.
- **Negative** – variables take on large values at different times.
- The notion of covariance and dependence are related but distinct concepts.
- **Independence** is different from covariance because it also includes non linear relationships. It is possible that two variables are dependent but have 0 covariance since covariance only measures linear correlation between two variables it doesn't account for non linear correlation.
- Covariance is effected by scale, so the larger your variables are the larger the covariance is going to be.
- In machine learning, we can exploit the property of covariance to either **compress** your data or in getting better results

Covariance matrix

- Covariance matrix of a random vector x is n x n matrix, such that

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

- The diagonal elements of the covariance matrix given the variance:

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

- While applying machine learning algorithms to our data, almost all of our input, weights, activations and outputs are going to be **vectors** and **matrices**. So often times we will be applying covariance. So we will be creating the covariance matrix so that it can applied to our analysis. Because we will be working exclusively with **matrix forms and notations** exclusively for all the steps of a machine learning project.

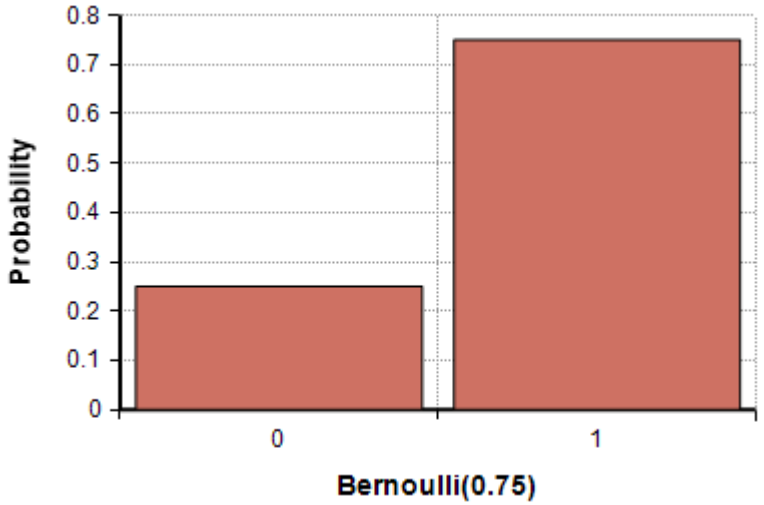
Common probability distributions

- Probability distributions are useful in many contexts in machine learning.
- Some of them are as follows –
  - Bernoulli
  - Normal
  - Poisson
  - Binomial

Bernoulli

- The Bernoulli distribution is a 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.
- It has the following properties

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



Normal

- The most commonly used distribution over real numbers, also known as **Gaussian** distribution
- The two parameters $\mu \in \mathbf{R}$ and $\sigma \in (0, \infty)$ control the normal distribution.
- The parameter μ gives the coordinate of the central peak.
- This is also the mean of the distribution: $\mathbf{E[x] = \mu}$. The standard deviation of the distribution is given by σ, and the variance by $\sigma 2$
- Normal distributions are a **sensible** choice for many applications.
- The central limit theorem shows that the sum of many independent random variables is approximately normally distributed. This means that in practice, many **complicated** systems can be modeled successfully as normally.
- Normal distribution **encodes** the maximum amount of **uncertainty** over the real numbers. We can thus think of the normal distribution as being the one that inserts the **least** amount of prior knowledge into a model

Normal

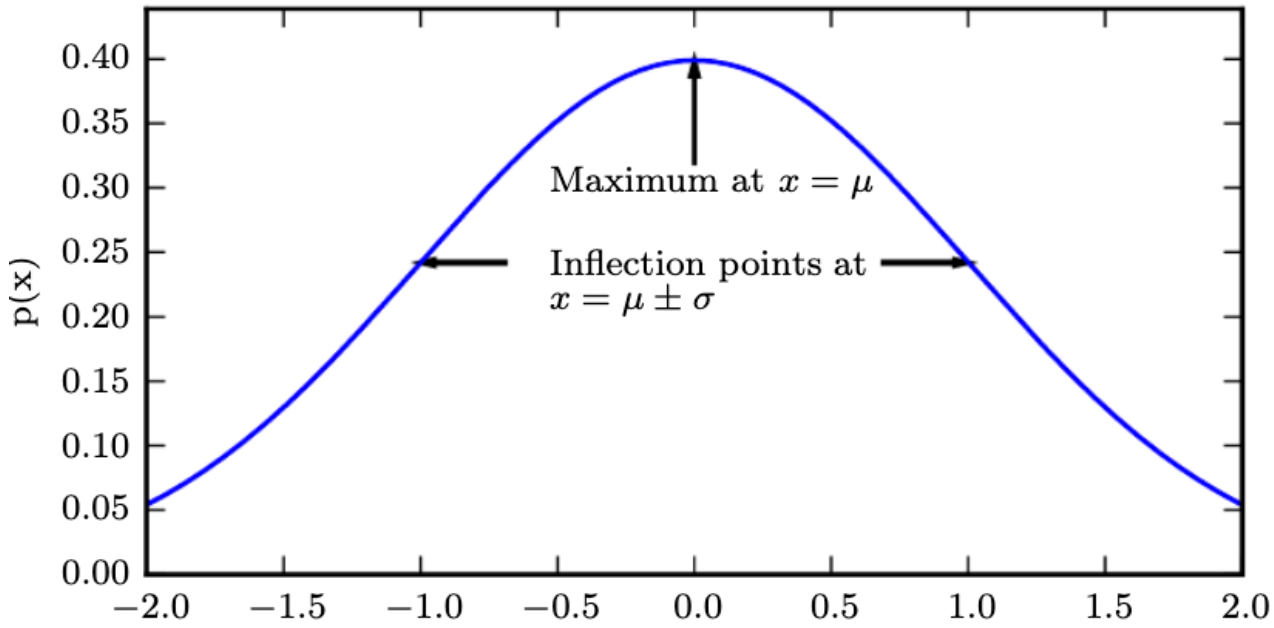
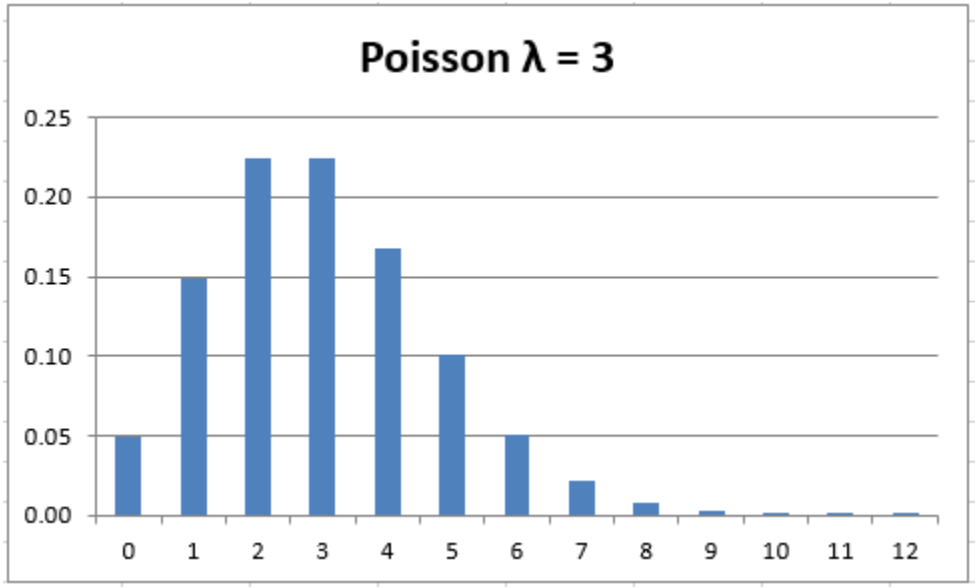


Figure: The normal distribution.

Note: The normal distribution N(x;μ, σ2) exhibits a classic "bell curve" shape, with the x coordinate of its central peak given by μ, and the width of its peak controlled by σ. In this example, we depict the **standard normal distribution**, with **μ = 0** and **σ = 1**

Poisson

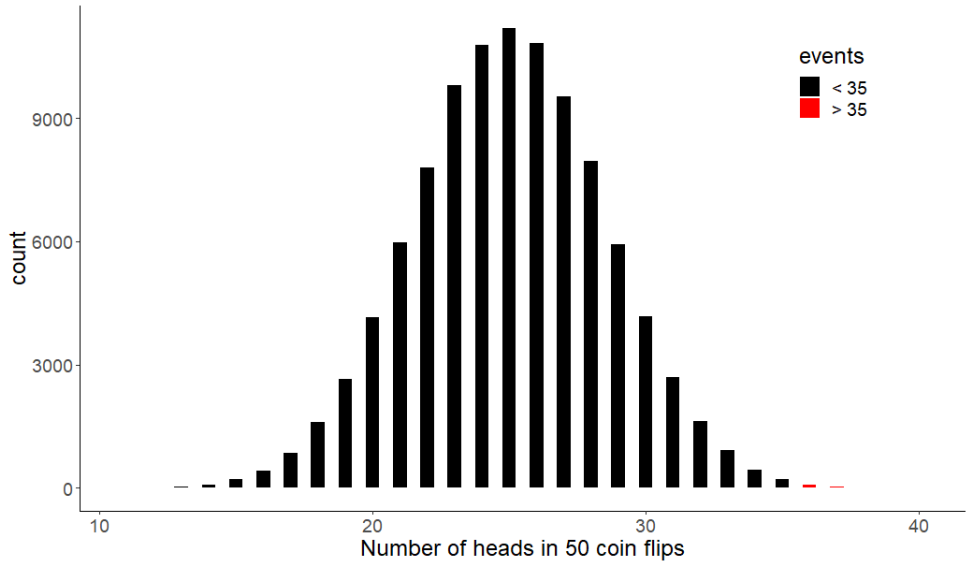
- Is a discrete probability distribution that expresses the probability of a given number of events occurring in a **fixed interval of time or space**.
- Some examples that may follow a Poisson distribution include,
  - Number of phone calls received by a call center per hour
  - Number of patients arriving in an emergency room between 10 and 11 pm



Binomial

- The binomial distribution with parameters n and p is the **discrete** probability distribution of the number of **successes** in a sequence of n independent experiments,
- The binomial distribution is frequently used to model the number of successes in a sample of size n drawn with replacement from a population of size N.
- Binomial probability distributions help us to understand the likelihood of rare events and to set probable expected ranges.

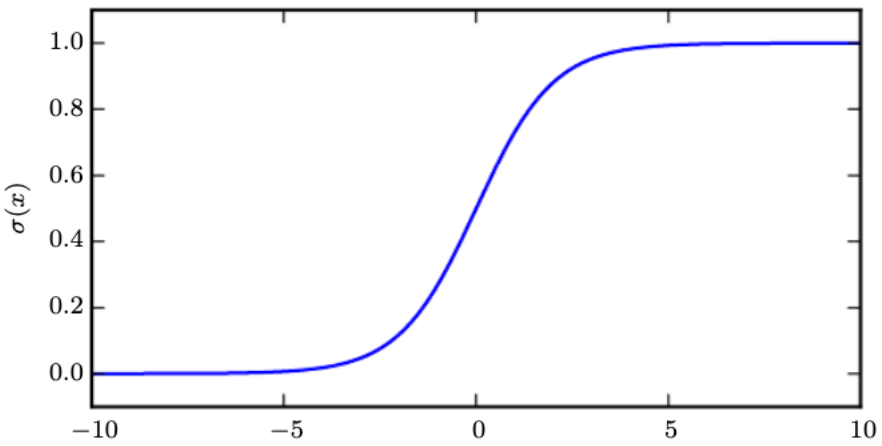
$$\Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$



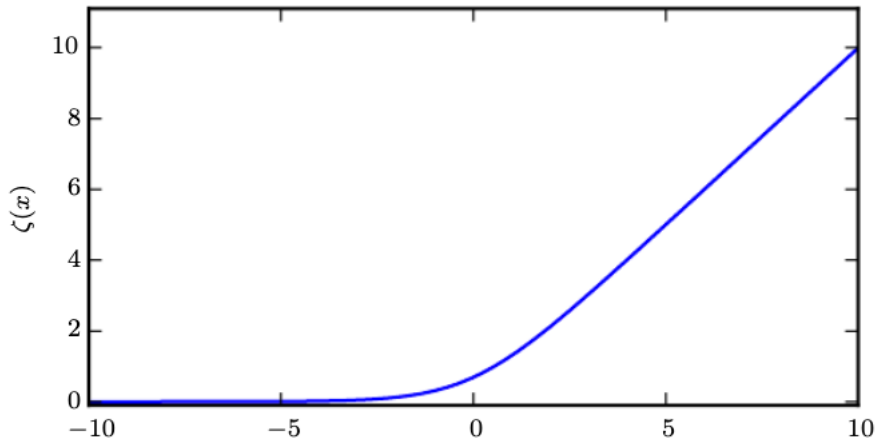
Useful properties of common distributions

- Certain functions arise often while working with probability distributions, especially the probability distributions used in deep learning models.
- Some of them are:
  - Sigmoid
  - Softmax
- **Sigmoid**:
  - The logistic sigmoid is commonly used to produce the Θ parameter of a Bernoulli distribution because its range is (0,1), which lies within the valid range of values for the Θ parameter.
  - The sigmoid function saturates when its argument is very positive or very negative, meaning that the function becomes very flat and insensitive to small changes in its input.
- **Softmax**:
  - The softmax function can be useful for producing the β or σ parameter of a normal distribution because its range is $(0, \infty)$.
  - It also arises commonly when manipulating expressions involving sigmoids.

Useful properties of common distributions



Sigmoid function



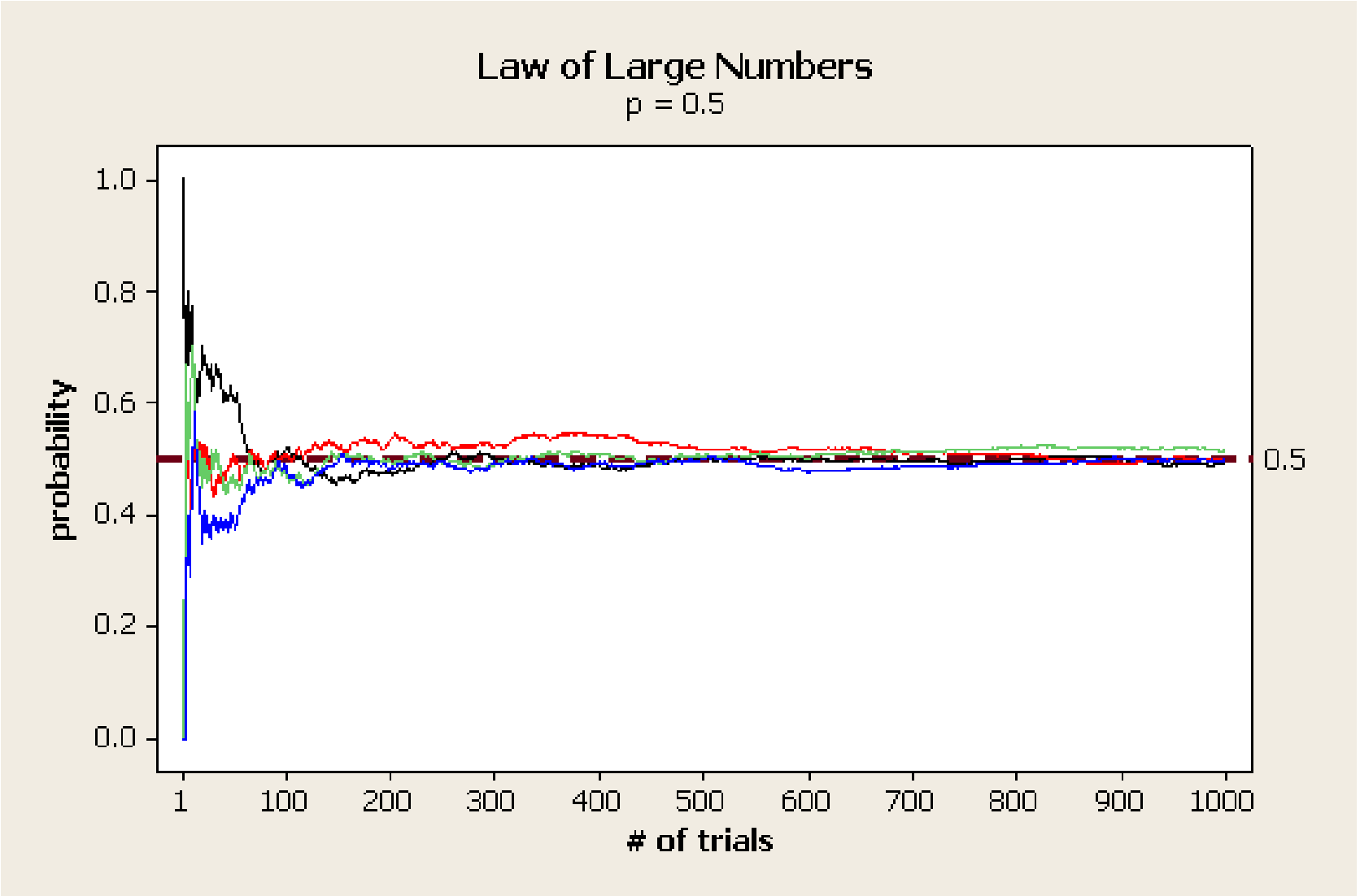
Softmax function

Law of large numbers

- The law of large numbers is a theorem from probability and statistics that suggests that the average result from repeating an experiment multiple times will better approximate the true or expected underlying result.
- **The law of large numbers explains why casinos always make money in the long run**.

  — Page 79, Naked Statistics: Stripping the Dread from the Data, 2014.

- We have an intuition that **more observations** are better. This is the same intuition behind the idea that if we collect **more data**, our sample of data will be more representative of the problem domain.
- Has important implications in applied machine learning.
- The law of large numbers is critical for understanding the selection of training datasets, test datasets, and in the evaluation of model skill in machine learning.
- States that the mean of a large sample is close to the mean of the distribution.
- The law reminds us to repeat the experiment in order to develop a large and representative sample of observations before we start making inferences about what the result means.


Law of Large Numbers
p = 0.5
probability
1.0
0.8
0.6
0.4
0.2
0.0
0.5
1
100
200
300
400
500
600
700
800
900
1000
# of trials

Law of large numbers (Implications in Machine Learning)

- The law of large numbers has important implications in applied machine learning.
- Let's take a moment to highlight a few of these implications –
- **Training data**:
  - The data used to train the model must be representative of the observations from the domain. This really means that it must contain enough information to generalize to the true unknown and underlying distribution of the population.
  - This is easy to conceptualize with a single input variable for a model, but is also just as important when you have multiple input variables.
  - There will be unknown relationships or dependencies between the input variables and together, the input data will represent a multivariate distribution from which observations will be drawn to comprise your training sample. Keep this in mind during data collection, data cleaning, and data preparation.
  - You may choose to exclude sections of the underlying population by setting hard limits on observed values (e.g. for outliers) where you expect data to be too sparse to model effectively.

Law of large numbers (Implications in Machine Learning)

- **Test data**:
  - The thoughts given to the training dataset must also be given to the test dataset.
  - This is often neglected with the blind use of 80/20 spits for train/test data or the blind use of 10-fold cross-validation, even on datasets where the size of 1/10th of the available data may not be a suitable representative of observations from the problem domain.
- **Model skill evaluation**:
  - Consider the law of large numbers when presenting the estimated skill of a model on unseen data.
  - It provides a defense for not simply reporting or proceeding with a model based on a skill score from a single train/test evaluation.
  - It highlights the need to develop a sample of multiple independent (or close to independent) evaluations of a given model such that the mean reported skill from the sample is an accurate enough estimate of population mean.

The central limit theorem

- The central limit theorem describes the shape of the distribution of sample means as a Gaussian or Normal distribution
- The theorem states that as the size of the sample increases, the distribution of the mean across multiple samples will approximate a Gaussian distribution.
- It demonstrates that the distribution of errors from estimating the population mean fit a normal distribution.
- This estimate of the Gaussian distribution will be more accurate as the size of the samples drawn from the population is increased. This means that if we use our knowledge of the Gaussian distribution in general to start making inferences about the means of samples drawn from a population, that these inferences will become more useful as we increase our sample size.
- The central limit theorem does not state anything about a single sample mean(like law of large numbers); instead, it is broader and states something about the shape or the distribution of sample means.

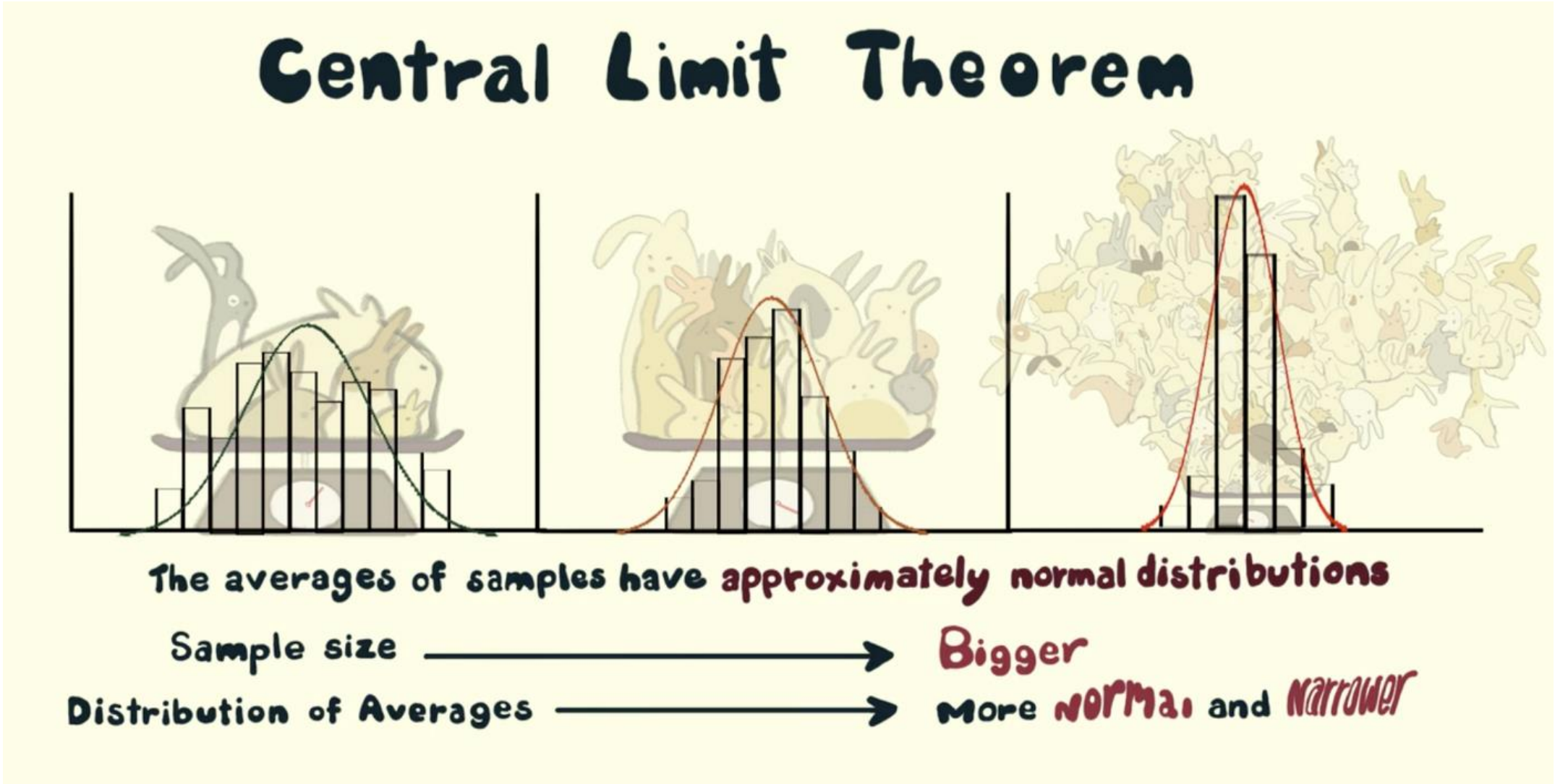


Image Credits: Casey Dunn & Creature Cast on Vimeo

The central limit theorem (Implications in Machine Learning)

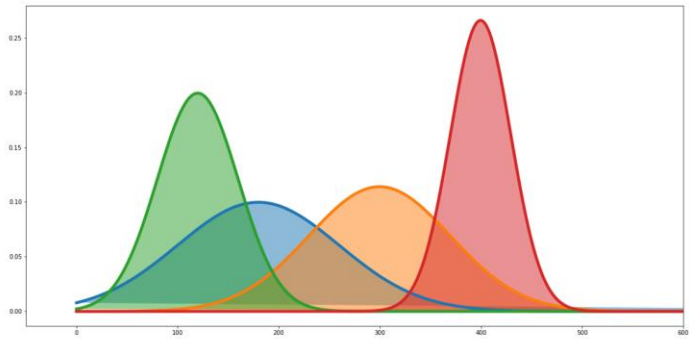
- The central limit theorem has important implications in applied machine learning. The theorem does inform the solution to linear algorithms such as linear regression, but not exotic methods like artificial neural networks that are solved using numerical optimization methods.
- **Significance Tests**:
  - In order to make inferences about the skill of a model compared to the skill of another model, we must use tools such as statistical significance tests.
  - These tools estimate the likelihood that the two samples of model skill scores were drawn from the same or a different unknown underlying distribution of model skill scores.
  - If it looks like the samples were drawn from the same population, then no difference between the models skill is assumed, and any actual differences are due to statistical noise.
  - The ability to make inference claims like this is due to the central limit theorem and our knowledge of the Gaussian distribution and how likely the two sample means are to be a part of the same Gaussian distribution of sample means.

The central limit theorem (Implications in Machine Learning)

- **Confidence Intervals**:
  - Once we have trained a final model, we may wish to make an inference about how skillful the model is expected to be in practice.
  - The presentation of this uncertainty is called a confidence interval.
  - We can develop multiple independent (or close to independent) evaluations of a model accuracy to result in a population of candidate skill estimates.
  - The mean of these skill estimates will be an estimate (with error) of the true underlying estimate of the model skill on the problem.
  - With knowledge that the sample mean will be a part of a Gaussian distribution from the central limit theorem, we can use knowledge of the Gaussian distribution to estimate the likelihood of the sample mean based on the sample size and calculate an interval of desired confidence around the skill of the model.

Maximum Likelihood

- The goal of maximum likelihood is to **fit** an **optimal** statistical distribution to some data.
- This makes the data **easier** to work with, makes it more general, allows us to see if **new** data follows the same distribution as the previous data, and lastly, it allows us to classify **unlabeled** data points.
- The reason you want to fit a distribution to your data is it can be easier to work with and it is also more general – it applies to every experiment of the same type
- To maximize the likelihood of the event of interest in our analysis.
- In everyday conversation, the probability and likelihood mean the same thing. However in statistical analysis, ”likelihood” refers to finding the optimal value for the mean or standard deviation for a distribution given a bunch of observed measurements. This is how we fit a distribution to data
- MLE (Maximum likelihood estimate) tells us which curve has the highest likelihood of fitting our data.



Bayes rule

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

Here,

P(A/B) is posterior as the **conditional probability** of event A given event B.

P(A) is our **prior**, or the initial **belief** of the probability of event A

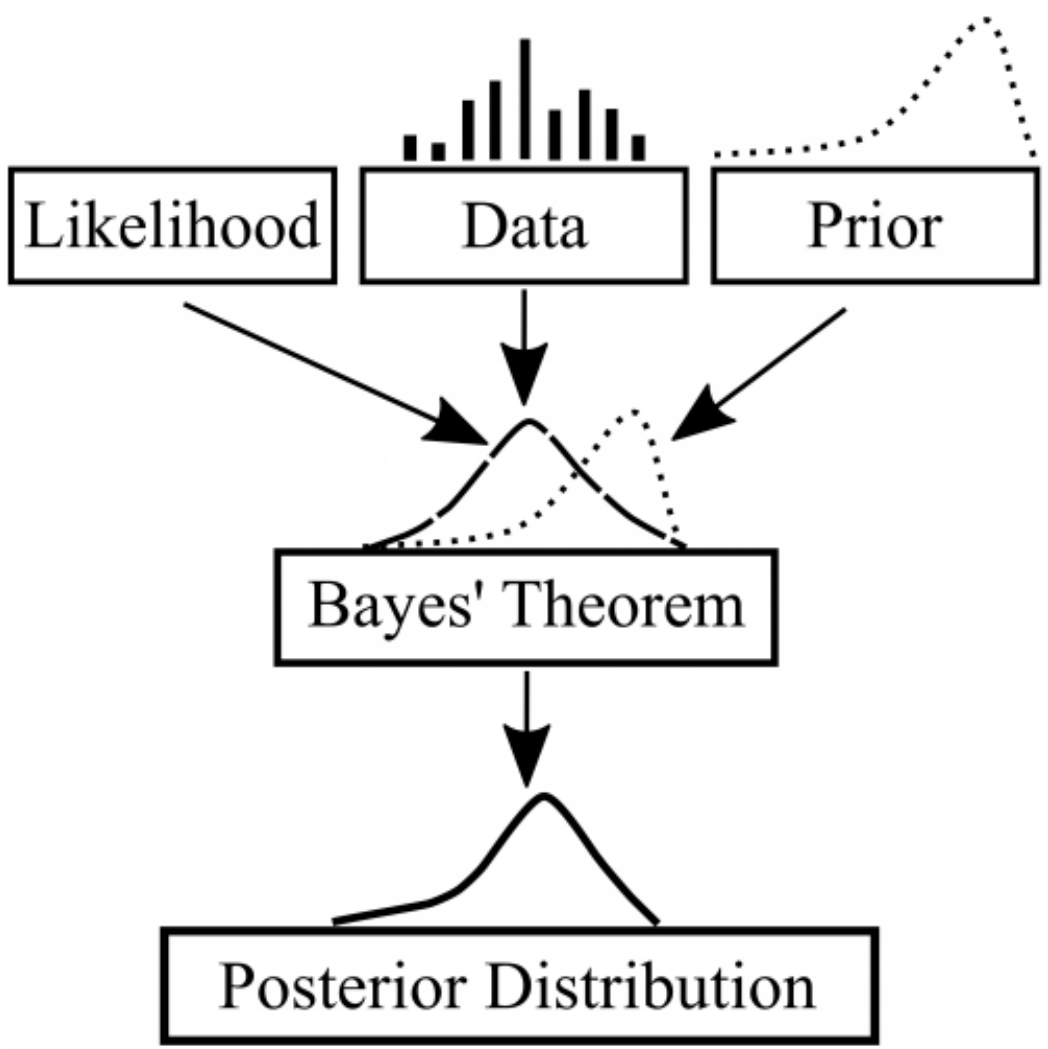
P(B|A) is the **likelihood** (also a conditional probability),
which we **derive** from our **data**, and

P(B) is a **normalization** constant to make the probability distribution sum to 1.

The general form of Bayes' Rule in statistical language is the
posterior probability equals the likelihood times the prior
divided by the normalization constant.

This short equation leads to the entire field of Bayesian Inference,
an effective method for reasoning about the world.

Bayesian inference



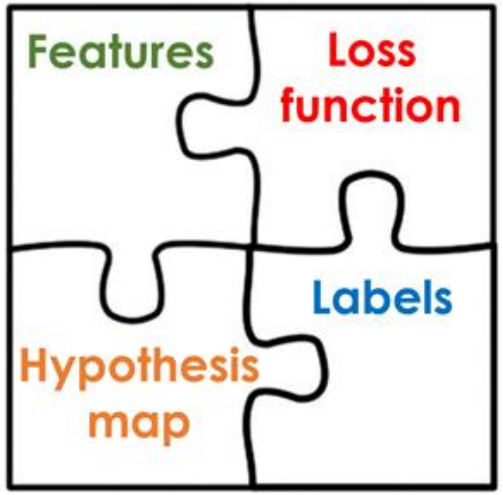
Bayesian inference

- Bayesian statistics is a mathematical procedure that applies probabilities to statistical problems. It provides people the tools to update their **beliefs** in the evidence of new data.
- A Bayesian is one who, vaguely expecting a **horse**, and catching a glimpse of a **donkey**, strongly believes he has seen a **mule**
- The fundamental idea of Bayesian inference is to become "**less wrong**" with **more data**.
- Bayesian data analysis is based on the following two principles:
  - Probability is interpreted as a measure of uncertainty, whatever the source. Thus, in a Bayesian analysis, it is standard practice to **assign** probability distributions not only to unseen data, but also to parameters, models, and hypotheses.
  - **Uncertainty** is quantified both before and after the collection of data and Bayes' formula is used to update our beliefs in light of the new data.

- Bayesian analysis, a method of statistical inference that allows one to combine **prior** information about a **population** parameter with **evidence** from information contained in a sample to **guide** the statistical inference process

Bayes theorem in Machine learning

- Bayes' theorem can be used in both regression, and classification.
- Generally, in Supervised Machine Learning, when we want to train a model the main building blocks are
  - a set of data points that contain **features** (the attributes that define such data points),
  - the **labels** of such data point (the numeric or categorical tag which we later want to predict on new data points), and
  - a **hypothesis** function or model that links such features with their corresponding labels.
  - We also have a **loss** function, which is the difference **between** the predictions of the model and the real labels which we want to reduce to achieve the best possible results.



Bayes theorem in Machine learning

- **Regression:**

  Using frequentist approach -

$$y = \theta_0 + \theta_1 x$$

Equation describing a linear model

  - After having trained the model with the available data we would get a value for both of the θs. This training can be performed by using an iterative process like gradient descent or another probabilistic method like Maximum Likelihood. In any way, we would just have ONE single value for each one of the parameters.

  Using Bayes approach –

  - When we use Bayes' theorem for regression, instead of thinking of the parameters (the θs) of the model as having a single, unique value, we represent them as parameters having a certain **distribution**: the **prior** distribution of the **parameters**.
  - What this means is that our parameter set (the θs of our model) is **not constant**, but instead has its own **distribution**. Based on previous knowledge (from experts for example, or from other works) we make a first hypothesis about the distribution of the parameters of our model. Then as we train our models with more data, this distribution gets updated and grows more **exact** (in practice the **variance** gets **smaller**)

References and further reading

- http://www.deeplearningbook.org/
- https://deepai.org/machine-learning-glossary-and-terms/probability-density-function
- https://towardsdatascience.com/an-intuitive-real-life-example-of-a-binomial-distribution-and-how-to-simulate-it-in-r-d72367fbc0fa
- https://towardsdatascience.com/bayes-rule-applied-75965e4482ff
- https://towardsdatascience.com/central-limit-theorem-in-action-1d4832599b7f
- https://en.wikipedia.org/wiki/List_of_probability_distributions
- https://towardsdatascience.com/probability-learning-ii-how-bayes-theorem-is-applied-in-machine-learning-bd747a960962
- https://towardsdatascience.com/probability-learning-iii-maximum-likelihood-e78d5ebea80c