Probability Theory

Probability Theory

- Probability theory is a mathematical framework for representing uncertain statements.
- It provides a means of quantifying uncertainty and axioms for deriving new uncertain statements.

Probability theory and information theory

- Probability theory allows us to make uncertain statements and reason in the presence of uncertainty.
- information theory allows us to quantify the amount of uncertainty in a probability distribution.

Machine learning

- Machine Learning must always deal with uncertain quantities, and sometimes may also need to deal with stochastic (non-deterministic) quantities.

Sources of uncertainty

1. Inherent stochasticity in the system being modeled.
2. Incomplete observability: Even deterministic systems can appear stochastic when we cannot observe all of the variables that drive the behavior of the system.
3. 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

- In the case of the doctor diagnosing the patient,
- we use probability to represent a **degree of belief,**
- **1 indicating absolute certainty that the patient has the flu** and
- 0 indicating absolute certainty that the patient does not have the flu.

- **frequentist probability:** related directly to the rates at which events occur,
- **Bayesian probability:** related to qualitative levels of certainty,

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

Random Variables

- A **random variable is a variable that can take on different values randomly.**
- Denote the random variable itself with a lower case letter in plain typeface,
- the values it can take on with lower case script letters.
- For example, *x1 and x2* are both possible values that the random variable x can take on.
- 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

Random variables :

- Discrete: A discrete random variable is one that has a finite or countably infinite number of states. States are not necessarily the integers
- Continuous: is associated with a real value.

Probability Distributions

- A **probability distribution is a description of how likely a random variable or** set of random variables is to take on each of its possible states.

Discrete Variables and Probability Mass Functions

- **Probability mass function (PMF):** A probability distribution over discrete variables may be described using PMF.
- Denoted by capital P .
- The probability that x = x is denoted as P (x) or P (x = x).
- P(x) is usually not the same as P(y).

Joint probability distribution

- **Joint probability distribution:** A Probability mass functions can act on many variables at the same time.
- **P (x = x, y = y) denotes the probability that x = x and y = y** simultaneously.
- Denoted by P(x, y)

Probability Mass Function Properties

1. The domain of P must be the set of all possible states of x.
2. $\forall x \in x, 0 \le P(x) \le 1$. An impossible event has probability 0 and no state can be less probable than that. Likewise, an event that is guaranteed to happen has probability 1, and no state can have a greater chance of occurring.
3. $\Sigma_{x \in x} P(x) = 1$. This property is known as **normalized. Without** this property, we could obtain probabilities greater than one by computing the probability of one of many events occurring

- Consider a single discrete random variable x with *k* *different* states.
- For a **uniform distribution on x**—states equally likely
- $P(x = x_i) = 1/k$

Continuous Variables and Probability Density Functions

- When working with continuous random variables, we describe probability distributions using a **Probability Density Function (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 we do not require $p(x) \leq 1$.
  - $\int p(x)dx = 1$.

- PDF $p(x)$ does not give the probability of a specific state directly
- Probability of landing inside an infinitesimal region with volume δx is given by $p(x)\delta x$
- Probability that x lies in the interval [a, b] is given by
- $\int_{[a,b]} p(x)dx$.

Uniform distribution

uniform distribution on an interval of the real numbers

- u(x; a, b),

where,

a and b are the endpoints of the interval, with $b > a$.

- The ";" notation means parametrized by";
- x is argument of the function,
- a and b are parameters that define the function.
- $u(x; a, b) = 1/(a-b)$ for all $x \in [a, b]$.
- $x \sim U(a, b)$: x follows the uniform distribution on [a, b]

Marginal Probability

- The probability distribution over the subset is known as the **marginal probability distribution:**
- For **discrete random variables** x and y, *P(x, y)*.
  - *P(x) with the* ***sum rule:***
    - $\forall x \in x, P(x = x) = \Sigma_y P(x = x, y = y)$

- For **continuous variables**, we need to use integration instead of summation
  - p(x)= $\int P(x,y)\, dy$

Conditional Probability

- Probability of some event, given that some other event has happened.
- *P(y = y | x = x).*

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

The Chain Rule of Conditional Probabilities

- Any joint probability distribution over many random variables may be decomposed into conditional distributions over only one variable:

$$P(\mathrm{x}^{(1)}, \ldots, \mathrm{x}^{(n)}) = P(\mathrm{x}^{(1)}) \Pi_{i=2}^{n} P(\mathrm{x}^{(i)} \mid \mathrm{x}^{(1)}, \ldots, \mathrm{x}^{(i-1)}).$$

Independence

- Two random variables x and y are **independent if their probability distribution** can be expressed as a product of two factors, one involving only x and one involving only y:
- $\forall \mathbf{x} \in \mathbf{x}, \mathbf{y} \in \mathbf{y}$, $\mathbf{p(x = x, y = y) = p(x = x)*p(y = y).}$
- $\mathbf{x \perp y}$ means that x and y are independent

Conditional Independence

- Two random variables x and y are **conditionally independent** given **a random variable z** if the **conditional probability distribution over x and y factorizes in this way for every value of z:**
- $\forall x \in$ x, $y \in$ y, $z \in$ z,
  p(x = x, y = y | z = z) = p(x=x | z = z) * p(y=y | z = z).
- **x⊥y | z** means that x and y are conditionally independent given z.

Expectation

- The **expectation or expected value of some function f(x) with respect to** a probability distribution P (x) **is the average or mean value that f takes on, when x is drawn from P.**
- **Discrete variables:**
  - $E_{x \sim P} [f(x)] = \Sigma_x P(x) f(x)$
- **Continuous variables:**
  - $E_{x \sim p} [f(x)] = \int p(x) f(x)\, dx$
- **Expectations are linear:**
  - $E_x[\alpha f(x) + \beta g(x)] = \alpha E_x[f(x)] + \beta E_x[g(x)]$
  - when *α and β are not dependent on x.*

Variance

- The **variance gives a measure of how much the values of a function of a random variable x vary,** as we sample different values of x from its probability distribution:
- $\mathrm{Var}(f(x)) = E[\ (f(x) - E[f(x))^2\ ]$
- When the variance is low, the **values of f (x) cluster near their expected value.**
- The square root of the variance is known as the **standard deviation**

Covariance

- The **covariance deals with how much two values are linearly related** to each other, as well as the scale of these variables:
- $\mathrm{Cov}(f(x), g(y)) = E\,[(f(x) - [E[(f(x)]])\ \ (g(y) - E\,[g(y)])]$.
- **High absolute values** of the covariance mean that the **values change very much** and are **far from their respective means.**
- If the sign of the **covariance is positive**, then both variables tend to take on **relatively high values simultaneously**.
- Two variables that are **independent** have **zero covariance**
- Two variables that have **non-zero covariance** are **dependent.**

Correlation

- It **normalize the** contribution of each variable in order to measure how much the variables are related.

Covariance matrix

- A covariance matrix is a square matrix giving the covariance between each pair of elements of a given random vector.
- $Cov(x)_{i,j} = Cov(x_i, x_j)$
- The diagonal elements of the covariance give the variance:
  - $Cov(x_i, x_i) = Var(x_i)$.

Common Probability Distributions

Bernoulli Distribution

- It **is a distribution over a single binary random variable.**
- It is controlled by a single parameter $\varphi \in [0, 1]$, which gives the probability of the random variable being equal to 1.
- **Properties**:
  - $P(x=1) = \varphi$
  - $P(x=0) = 1 - \varphi$
  - $P(x = x) = \varphi^x (1 - \varphi)^{1-x}$
  - $E_x[x] = \varphi$
  - $Var_x(x) = \varphi(1 - \varphi)$

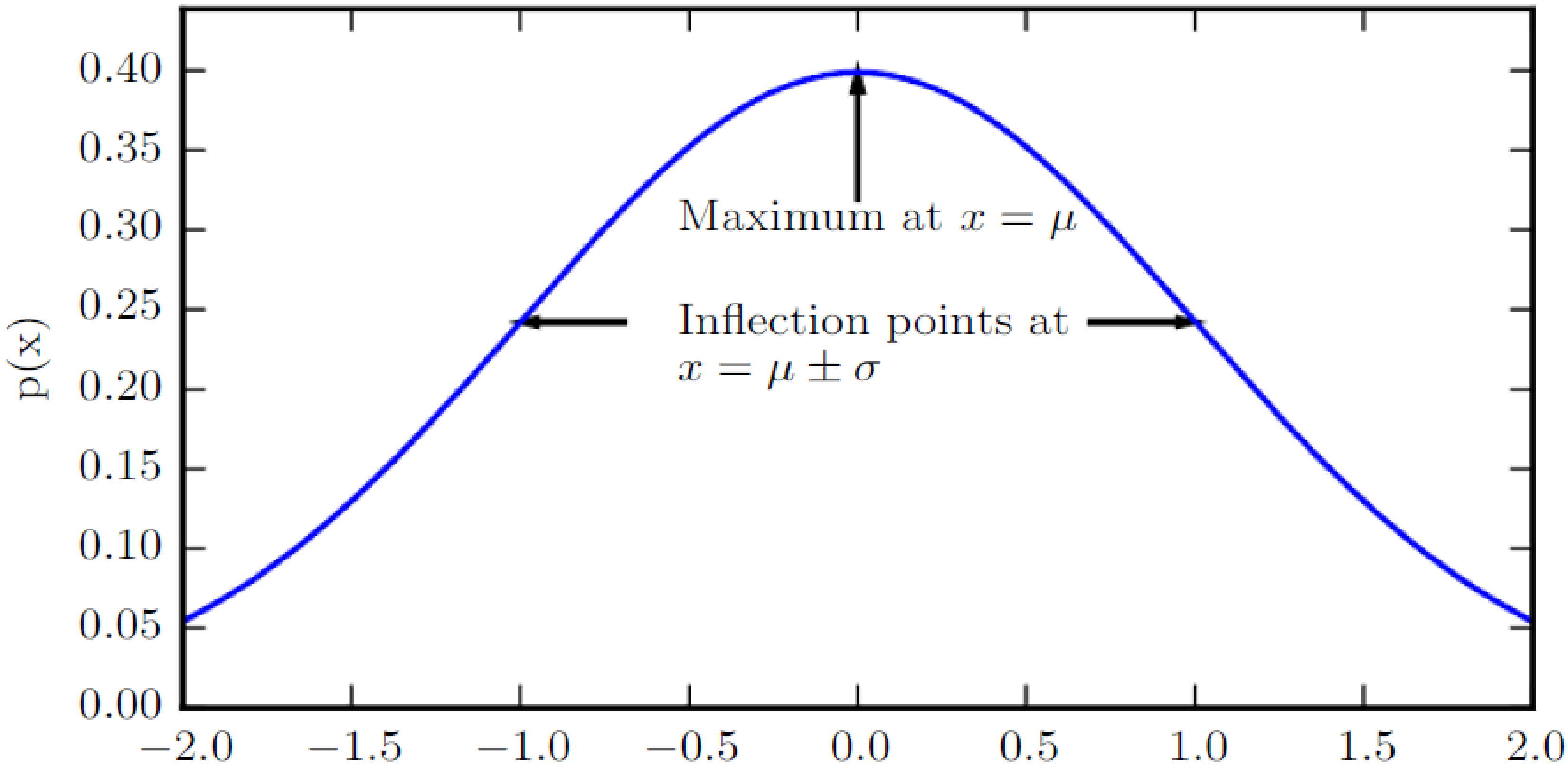
Gaussian Distribution

- The **most commonly** used distribution over real numbers is the **normal distribution.**
- Also known as the **Gaussian distribution**

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

- The parameter μ is the mean of the distribution, $E[x] = \mu$.
- The standard deviation is given by σ, and the variance by σ^2 .
- It **inserts the least amount of prior knowledge into a model.**

The normal distribution



It 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.

The standard normal distribution, with $\mu = 0$ and $\sigma = 1$.

Parameterized by **precision**

- When we need to frequently evaluate the PDF with different parameter values, a more efficient way of parametrizing the distribution is to use a parameter $\beta \in (0, \infty)$ *to control the* **precision or inverse variance of the distribution**

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

Absence of prior knowledge

The normal distribution is a default choice in the absence of prior knowledge for two major reasons.

1. The **central limit theorem shows that the sum of many independent** random variables is approximately normally distributed
2. Out of all possible probability distributions with the same variance, the normal distribution **encodes the maximum amount of uncertainty over the real numbers**.

Multivariate Normal Distribution

- Parameterized by covariance matrix Σ:

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

  – Σ gives the covariance matrix of the distribution

- Parameterized by **precision matrix *β*:**

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

Exponential and Laplace Distributions

- **Exponential distribution:** It is used when probability distribution with a sharp point at $x = 0$ is required **:**
  - $p(x; \lambda) = \lambda 1_{x \geq 0} \exp(-\lambda x)$
- The indicator function $1_{x \geq 0}$ is used to assign **probability zero to all negative values of x.**

Laplace Distribution

- It places a sharp peak of probability mass at an arbitrary point μ.

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

Dirac Distribution

- It specifies that **all of the mass in a probability distribution clusters** around a **single point.**
- Dirac delta function, $\delta(x)$:
  - $p(x) = \delta(x - \mu)$.
- It is **zero-valued** everywhere **except 0**, and integrates to 1

Empirical Distribution

- A common use of the Dirac delta distribution is as a component of an **empirical distribution**

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

- It puts probability mass 1/m on each of the m points x(1) , . . . , x(m) forming a given dataset.
- The Dirac delta distribution is only necessary to define the empirical distribution over continuous variables.

Mixtures of Distributions

- A mixture distribution is made up of several component distributions
  - $P(x) = \Sigma_i \; P(c=i)P(x \mid c=i)$.
  - P(c) is the distribution over component identities

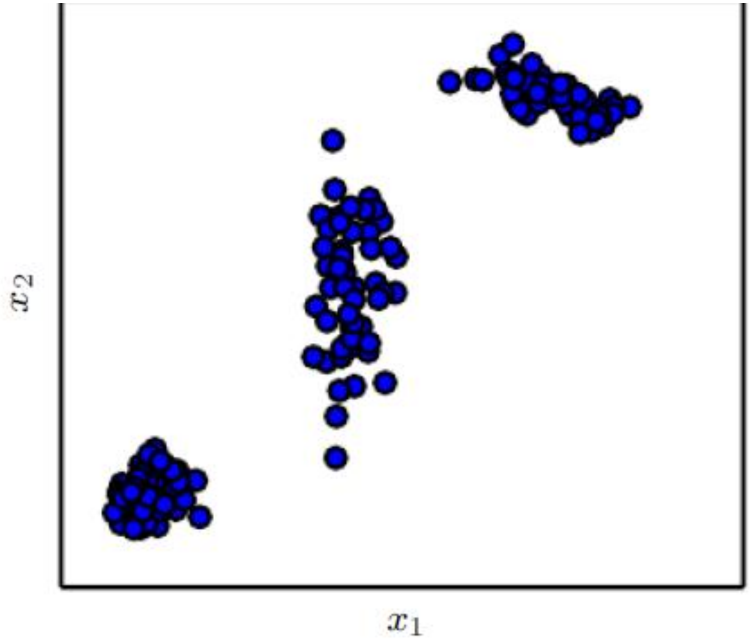
Gaussian Mixture model

- **Powerful and common type** of mixture model
- components $p(x \mid c = i)$ are Gaussians.
- Each component has a separately parametrized mean $\mu(i)$ and covariance $\Sigma(i)$.

Gaussian mixture model

Three components:

- It has the same amount of variance in each direction. (isotropic covariance matrix)
- It can control the variance separately along each axis-aligned direction. (diagonal covariance matrix)
- It has a full-rank covariance matrix, allowing it to control the variance separately along an arbitrary basis of directions.

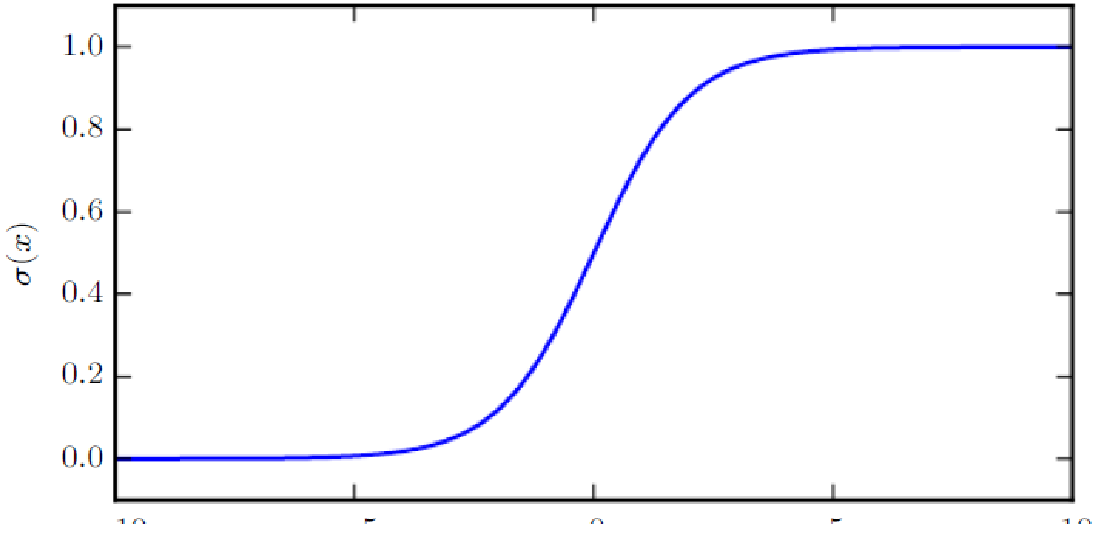


Logistic Sigmoid

- Commonly used to produce the φ parameter in Bernoulli distributions

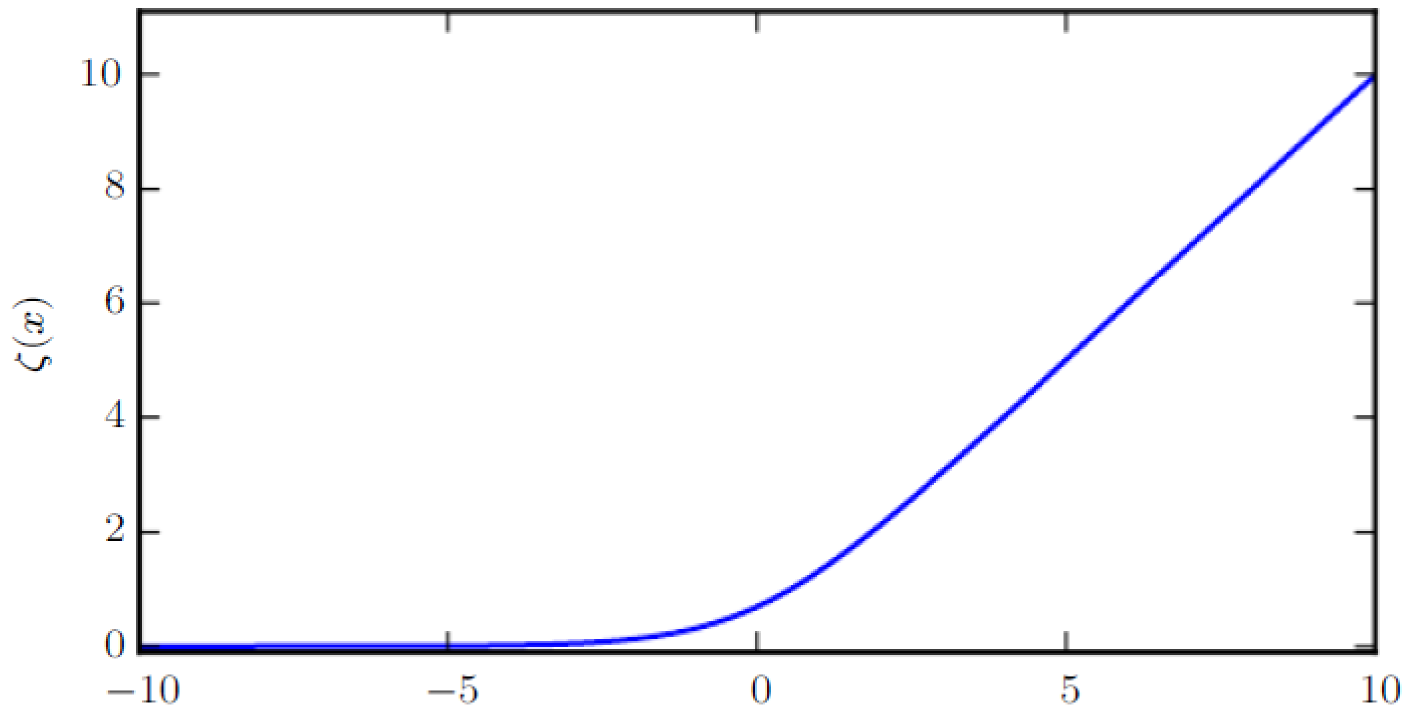
$$\sigma(x) = \frac{1}{1+\exp(-x)}$$

- Range➔(0,1)
- It **saturates** when its argument is **very positive or very negative**, i.e. insensitive to small changes in its input.



Softplus function

- $\zeta(x) = \log(1 + \exp(x))$
- Range➔ $(0, \infty)$.
- $x^+ = \max(0, x)$



Some useful properties

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

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

$$1 - \sigma(x) = \sigma(-x)$$

$$\log \sigma(x) = -\zeta(-x)$$

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

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

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

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

$$\zeta(x) - \zeta(-x) = x$$