Introduction to Probabilistic Machine Learning

Ralf Herbrich

Hasso Plattner Institut
Digital Engineering · Universität Potsdam

Probability

Nightline Potsdam

Egal was ist, du kannst uns anrufen!

Wir führen unsere Gespräche wertschätzend und unvoreingenommen.
Du bleibst dabei **anonym**.

dienstags, mittwochs, donnerstags und sonntags
von 21 bis 24 Uhr
unter 0331 977 1834 oder im Chat

https://nightline-potsdam.de

Weitere Infos:

Course Setup

- **Goal**: Stimulate interest in method development for machine learning algorithms
  - We will pick the pace that helps you to get excited; please interrupt and ask questions!
- **Format**: We have one topic per week with a lecture and tutorial
  - **Lecture**: Monday, 11:00am – 12:30pm (HS1)
  - **Tutorial**: Tuesday: 3:15pm – 4:45pm (HS1)
- **Assignment**: Six 2-weeks assignments to solve them (groups of two). They account for 30% of all points (5 points each)
  - Handed out every other week on Monday (starting 2nd week, April 15)
  - Each assignment has a theory and a practice part
- **Tutorial**: Supporting the material of the lecture and the assignments
  - In the tutorial, Rainer and Alex will solve similar exercises to the assignments with you
  - They will answer questions you have with the actual assignments
- **Exam (70 points)**: Counts for 70% of all points; 90-120 minutes long (w/c July 22)

Course Material

- **Books**: All our material and communication will happen over Moodle
  - Bishop, C. Pattern Recognition and Machine Learning. Springer. 2006.
  - MacKay, D. Information Theory, Inference, and Learning Algorithms. CUP. 2003
- **Moodle**: Share our lecture slides, tutorials, solutions
  - **Location**: https://moodle.hpi.de/course/view.php?id=755
  - **Announcements**: https://moodle.hpi.de/mod/forum/view.php?id=25738
- **GitHub Repository**: Supporting material as well as code samples
  - **Location**: https://github.com/HPI-Artificial-Intelligence-Teaching/pml-sose2024
  - If you find mistakes, please submit issues and pull requests
- **GitHub Classrooms**: Used for all our assignments
  - If you do not have a GitHub account, please create one now
  - Find a team member as assignments are solved in groups of two
  - More details tomorrow in the first tutorial

Course Structure

1. Probability Theory (Unit 1)
2. Inference & Decision Theory (Unit 2)
3. Graphical Models: Independence (Unit 3)
4. Graphical Models: Inference (Unit 4)
5. Bayesian Ranking (Unit 5)
6. Linear Basis Function Models (Unit 6)
7. Bayesian Regression (Unit 7)
8. Non-Bayesian Classification (Unit 8)
9. Bayesian Classification (Unit 9)
10. Gaussian Processes (Unit 10)
11. Information Theory (Unit 11)
12. Real-World Applications of Probabilistic Machine Learning (Unit 12)

Julia

- 2012 developed by Jeff Bezanson, Alan Edelman, Stefan Karpinski and Viral B. Shah at MIT
- Used for numerical and scientific computing with high performance
  - Execution speed is similar to C and FORTRAN
  - Hierarchical and parameterized type system as well as method overloading („multiple dispatching“) as central concepts
  - Native calls from C-(compiled) code possible (without wrappers)
- Unicode is efficiently supported (e.g., UTF-8)
- Alongside C, C++ and FORTRAN, the only programming language that has entered the “PetaFlop Club“

Jeff Bezanson (1981–)

Alan Edelman (1963 –)

Stefan Karpsinski (1981–)

Viral Shah

Overview

1. History of Machine Learning
2. Probability in Machine Learning
3. Probability Theory
4. Probability Distributions

Overview

1. **History of Machine Learning**
2. Probability in Machine Learning
3. Probability Theory
4. Probability Distributions

History of Machine Learning

Hasso Plattner Institut
Digital Engineering • Universität Potsdam

1950: Turing Test

1957: Perceptrons

1967: Nearest Neighbors

1986: Neural Networks

1995: Support Vector Machines

2000: Graphical Models

2012: Deep Neural Networks

Alan Turing

Frank Rosenblatt

Thomas Cover

Geoffrey Hinton

Vladimir Vapnik

Michael Jordan

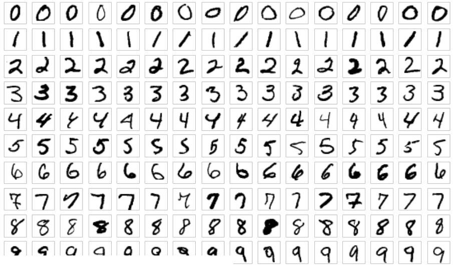
Yann LeCun, Geoffrey Hinton, Yoshua Bengio

Machine Learning: Definition

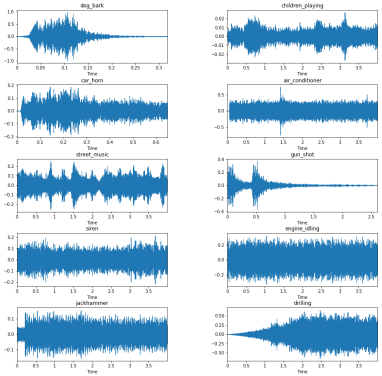
- **Tom Mitchell (1997)**. *A computer program is said to* ***learn*** *from experience* E *with respect to some class of tasks* T *and performance measure* P*, if its performance at tasks in* T*, as measured by* P*, improves with experience* E*.*
  - **Performance measures** are often called *loss functions*
  - **Experience** is often called *training data*
  - **Task** is also called a *prediction* by a computer program
- **Temporal Definition**. *A computer program is said to* ***learn*** *from data* D *recorded* ***in the past*** *if the accuracy of predictions made* ***in the future*** *improves over time.*
  - **Accuracy**: Performance measure against which an ML algorithm is judged
  - **Past Data**: Training data
  - **Future Data**: Test data

Machine Learning: Classification

- **Task**: Assigning examples to one of K **pre-defined** classes
  - **Examples**:
    - Digits classification to 10 classes based on pixel images
    - Phoneme classification
    - Auto-correct models for text input



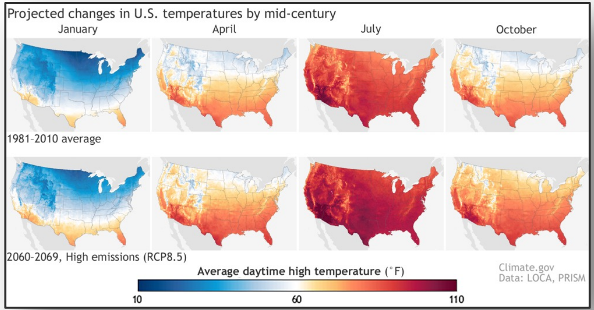
- **Performance**: Cost of misclassifying an example
  - **Examples**:
    - Symmetric loss: $l(\hat{y}, y) = \begin{cases} 1 & \hat{y} \neq y \\ 0 & \text{otherwise} \end{cases}$
    - Non-symmetric loss: $l(\hat{y}, y) = \boldsymbol{C}_{\hat{y},y} \in \mathbb{R}^{K \times K}$



Machine Learning: Regression

- **Task**: Assigning examples to a real value
  - **Examples**:
    - Price prediction of a good/service (Product Pricing)
    - Temperature prediction (Weather Forecast)
    - Effect of medication on health metrics (Digital Health)

- **Performance**: Cost of missing the true target y by $\widehat{\Delta} = \hat{y} - y$
  - **Examples**:
    - Symmetric loss: $l(\hat{y}, y) = h(|\hat{y} - y|)$ with h being monotonic
    - Non-symmetric loss: $l(\hat{y}, y) = h(\hat{y} - y)$



Overview

1. History of Machine Learning
2. **Probability in Machine Learning**
3. Probability Theory
4. Probability Distributions

What is Probability?

- **Weather forecast**: A meteorologist says

  „Tomorrow, it is going to rain in Bangalore with 60%"

- **Two interpretations**:
  1. The meteorologist has analyzed all regions which have similar environmental conditions than Bangalore today. His (**objective**) estimate based on past data is that the procedure which predicts rain tomorrow is correct 60% of the time.
  2. The meteorologist *believes* that it is more likely that it rains tomorrow in Bangalore (than it is to not rain tomorrow). 60% is the quantification of the (**subjective**) belief of the meteorologist.

Frequentist vs. Subjectivist Interpretation

- **Frequentist Interpretation**
  - Probability is a property of the event (“it rains tomorrow in Bangalore”)
  - Is operationalized by repeated experiments
  - Typically used by scientists and engineers

- **Subjective Interpretation**
  - Probability is an expression of belief of the person that makes a statement
  - Is subjective and people-dependent: Two people with identical data can come to different probabilities
  - Typically used by philosophers and economists

1. Probability is not a physical measure but a thought model for randomness!
2. The mathematical rules for probability are **identical** for both interpretations!

History of Probability

- **BC**:
  - Chance games were highly popular in ancient Greece & Rome
  - No mathematical analysis of chance (missing algebraic framework)
- **16th century**:
  - Girolamo Cardano published first book on methods to calculate the probability of card games and game of dice
- **17th century**:
  - Pierre de Fermat und Blaise Pascal exchange important questions about probability and motivate the first scientific studies of probability

Gerolamo Cardano
(1501 – 1575)

Pierre de Fermat
(1607 – 1665)

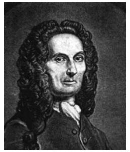
Blaise Pascal
(1623 – 1662)

History of Probability (ctd)

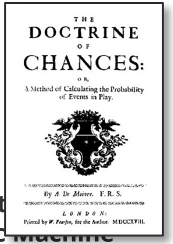
- **18th century**:
  - Jacob Bernoulli investigates random coin tosses and proves the Law of Large Numbers
  - Thomas Bayes investigates conditional probabilities and formulates Bayes' rule (published post-hum)
  - Abraham de Moivre introduces the normal distribution as the limit distribution of an infinite sum of Bernoulli random variables (weak form of central limit theorem)
- **19th century**:
  - Carl Friedrich Gauss introduces the least-square method and proves that the distribution of independent random variables converges to a normal distribution
  - Pierre-Simon Laplace publishes *Théorie analytique des probabilités* where he united probability theory and statistics and introduces hypothesis tests

Jacob Bernoulli (1655 – 1705)

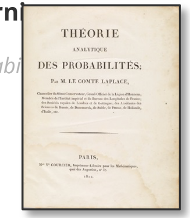
Thomas Bayes (1701 – 1761)



Abraham de Moivre (1667 – 1754)



Pierre-Simon Laplace (1749 – 1827)

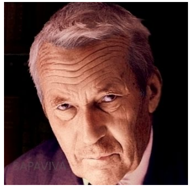


History of Probability (ctd)

HPI Hasso Plattner Institut
Digital Engineering · Universität Potsdam

- **20th century**:
  - Andrey Kolmogorov introduces three axioms that imply the rules of probability theory for relative frequencies (frequentist interpretation)
  - Richard Threlkeld Cox introduces three (other) axioms that imply the rules of probability theory for degrees of belief (subjectivist interpretation)

- **21st century**:
  - **1974**: Vladimir Vapnik introduces probability theory as the foundation of the theory of machine learning
  - **1980**: Judea Pearl and Phil Dawid introduce graphical models that allow to operationalize probability theory for complex and causal processes
  - **2000**: Probability theory and statistics enter every aspect of modern science and artificial intelligence



Andrey Kolmogorov
(1903 – 1987)


A. Kolmogoroff
Grundbegriffe der Wahrscheinlichkeitsrechnung
Springer-Verlag
Berlin Heidelberg New York

Richard Threlkeld Cox
(1898 – 1991]Introd


AMERICAN JOURNAL of PHYSICS
Probability, Frequency and Reasonable Expectation

Probabilistic Machine
arni
babi

Vladimir Vapnik
(1936 –)

Judea Pearl
(1936–)

Philip Dawid
(1946–)

Rules of Probability

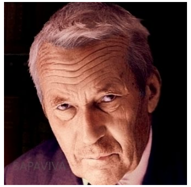
- **Mathematical Definition**. *A number* $P(A) \in [0,1]$ *assigned to an* event *or* statement A *that indicates how* likely A *is to occur.*
- **Set Theory**. We model events and statements via set theory and assume
  - A countably infinite total set $\Omega \supseteq A$
  - If $A(x)$ is a 1st order logic statement, then $\mathrm{A} := \{x \mid A(x)\}$ and
    - $\mathrm{A} \subseteq \mathrm{B} \equiv \forall x: A(x) \rightarrow B(x)$ and $\mathrm{A}^c \equiv \forall x: \neg A(x)$
    - $\mathrm{A} \cup \mathrm{B} \equiv \forall x: A(x) \vee B(x)$ and $\mathrm{A} \cap \mathrm{B} \equiv \forall x: A(x) \wedge B(x)$
- **Rules**: For all $A, B \subseteq \Omega$
  - **Monotonicity**: If $A \subseteq B$ then $P(A) \leq P(B)$
  - **Complement Rule**: $P(A^c) = 1 - P(A)$
  - **Sum Rule**: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
  - **Product Rule**: $P(A \cap B) = \underbrace{\frac{P(A \cap B)}{P(B)}}_{P(A|B)} \cdot P(B)$

Frequentist vs. Subjective Probabilities

- **Kolmogorov (1933)**: *The rules of probability for **sets** follow from the following 3 axioms*
    1. $P(A) \geq 0$ *for all* $A \subseteq \Omega$
    2. $P(\Omega) = 1$
    3. $P(\bigcup_i A_i) = \sum_i P(A_i)$ *if for all* $i \neq j: A_i \cap A_j = \emptyset$

- **Cox (1944)**: *The rules of probability for **logic** follow from the following 3 axioms*
    1. $P(A) \in [0,1]$ *for all logical statements* A
    2. $P(A)$ *is independent of how the statement is represented*
    3. *If* $P(A|C') > P(A|C)$ *and* $P(B|A \wedge C') = P(B|A \wedge C)$ *then*

$$P(A \wedge B|C') \geq P(A \wedge B|C)$$



Andrey Kolmogorov (1903 – 1987)

Richard Threlkeld Cox (1898 – 1991)

The Role of Probability in Machine Learning

- **Theory**: *How likely is it, that the accuracy of a predictor $\mathcal{A}(D)$ learned from training data D is small?*

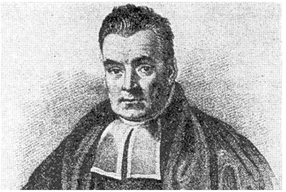
  $$\boxed{P}\left(\text{Accuracy}\left(\mathcal{A}(D)\right) < \varepsilon\right) \le \delta$$

- **Typical Assumptions**
  1. Independent identically distributed data (IID)
  2. Accuracy is an expected performance measure on the next test example
- **Frequentist view on probability**: Over N applications of the learning algorithm and draws of random training data D, for how many is the learned predictor accurate?

- **Practice**: *What can we say about the plausibility of a single predictor f in light of training data D?*

  $$\boxed{P(f|D) = \frac{P(D \wedge f)}{P(D)} = \frac{P(D|f)P(f)}{P(D)}}$$

- **Typical Assumptions**
  - Independent identically distributed data (IID)
  - Known conditional dependence of data and predictor
- **Subjectivist view on probability**: Given the certain and known training data, what is the remaining uncertainty over the right predictor for (future) data?



(Rev) Thomas Bayes (1701 – 1761)

Overview

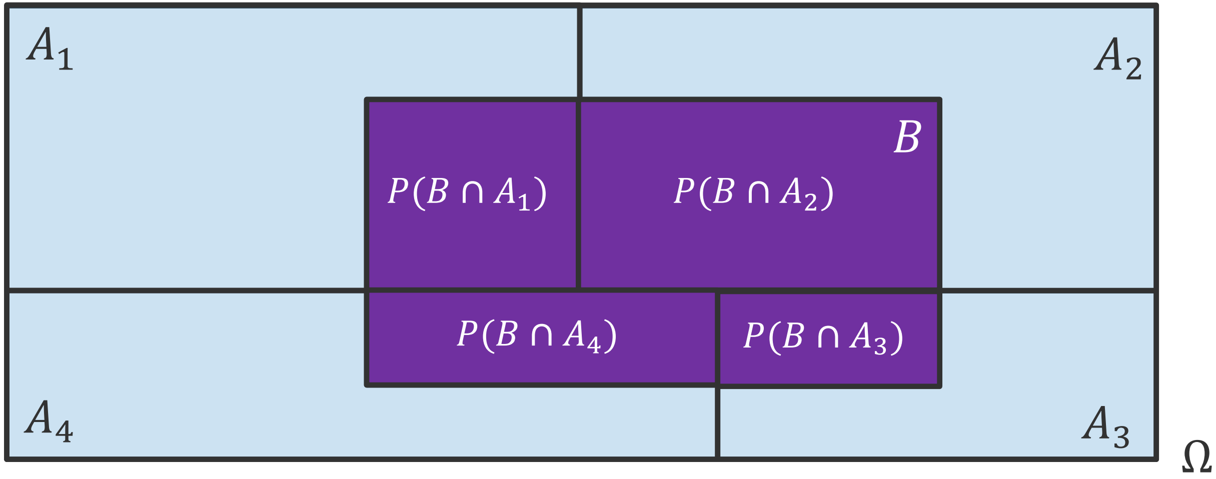
1. History of Machine Learning
2. Probability in Machine Learning
3. **Probability Theory**
4. Probability Distributions

Probability Theory: Sum Rule

- **Total Probability Theorem**. *Let $A_1, A_2, \dots, A_n \subseteq \Omega$ be disjoint events that form a partition of the sample space Ω and $P(A_i) > 0$ for all A_i. Then, for any event $B \subseteq \Omega$*

$$P(B) = \sum_{i=1}^{n} P(B \cap A_i) = \sum_{i=1}^{n} P(B|A_i) \cdot P(A_i)$$

- **Geometric Proof**



Probability Theory: Bayes Rule

- **Bayes' Theorem**. *Let $A_1, A_2, \ldots, A_n$ be disjoint events that form a partition of the sample space S and $P(A_i) > 0$ for all A_i. Then, for any event B with $P(B) > 0$*

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B)} = \frac{P(B|A_i)P(A_i)}{\sum_{j=1}^{n} P\left(B|A_j\right) \cdot P(A_j)}$$

- **Proof**. Follows from the definition of conditional probability and "multiply-by-1"

$$P(A_i \cap B) \cdot \boxed{\frac{P(B)}{P(B)}} = P(A_i \cap B) \cdot \boxed{\frac{P(A_i)}{P(A_i)}}$$

=1 (by definition $P(A_i) > 0$ and $P(B) > 0$)

$$P(A_i|B) \cdot P(B) = P(B|A_i) \cdot P(A_i)$$

(by definition of conditional probability)

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B)}$$

- **Simplified view** when looking at the probabilities as functions of A_i

$$\underbrace{\boxed{P(A_i|B)}}_{\text{posterior}} \propto \underbrace{\boxed{P(B|A_i)}}_{\text{likelihood}} \cdot \underbrace{\boxed{P(A_i)}}_{\text{prior}}$$

Bayes Rule: False-Positive Puzzle

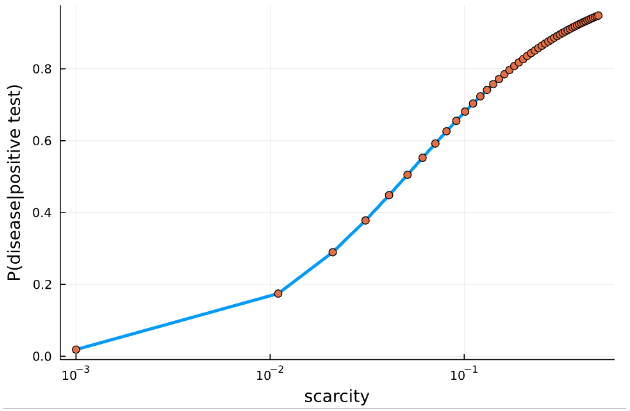
- **Situation**: A test for a rare disease is assumed to be correct 95% of the time (i.e., the probability that the test shows the disease or lack thereof is 95%). It's a rare disease that occurs in 0.1% of the population. If you have a positive test outcome, what is the probability that you have the disease?
- **Solution**:

$$A = \text{"Person has the disease"}$$

$$B = \text{"Test result is positive"}$$

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B|A) \cdot P(A) + P(B|\neg A) \cdot P(\neg A)}$$

$$P(A|B) = \frac{0.95 \cdot 0.001}{0.95 \cdot 0.001 + 0.05 \cdot 0.999} \approx 0.0187$$



- **Counterintuitive**: According to *The Economist* (February 20, 1999), 80% of leading American hospital staff guessed the probability to be 95%!

Probability Theory: Independence

- **Independence**. *We say that the events* $A_1, A_2, \ldots, A_n$ *are independent if*

$$P\left(\bigcap_{i \in I} A_i\right) = \prod_{i \in I} P(A_i), \qquad \text{for all subsets } I \text{ of } \{1, \ldots, n\}$$

- **Intuition**. Knowledge of an event A does not provide information about the probability of an independent event B

$$\underbrace{P(A \cap B)}_{P(B|A) \cdot P(A)} = P(B) \cdot P(A) \Leftrightarrow \boldsymbol{P(B|A) = P(B)}$$

- **Important modelling assumption** (often implicitly) used in machine learning when making assumptions about training and test data generation: knowing one training example provides no information about the probability of any other training example (realistic?!)
- **Counterintuitive geometry**: If A and B are disjoint, they are **not** independent!

Probability Theory: Random Variable

- **Random Variable**. *A random variable is a real-valued function of the outcome of the experiment. A function of a random variable defines another random variable.*
  - **Examples**:
    - Tossing a coin N times, the **number** of heads
    - Given an image, the **pixel intensity** of the top-left pixel (in 8-bit)
- **Probability Mass Function**. *The probability mass function $p(x)$ assigns each value x the probability that the random variable takes the value x.*
  - **Example**: Coin toss: If $N = 2$ then

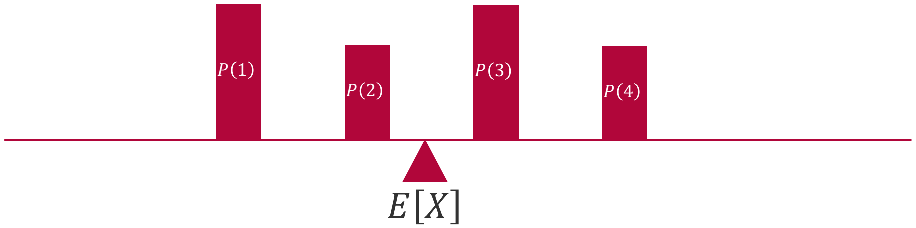
$$p(0) = P(\text{tail}, \text{tail})$$
$$p(1) = P(\text{head}, \text{tail}) + P(\text{tail}, \text{head})$$
$$p(2) = P(\text{head}, \text{head})$$

Probability Theory: Expectation and Variance

- **Expected Value**. *The expected value* $E[X]$ *(also called* expectation*) of a random variable* X *is defined by*

$$E[X] := \sum_x x \cdot p(x)$$

- **Intuition**. Center of gravity when placing the weight $p(x)$ at position x on a straight line



- **Variance**. *The variance* $\mathrm{var}[X]$ *of a random variable* X *is defined by*

$$\mathrm{var}[X] := \sum_x (x - E[X])^2 \cdot p(x) = E[(X - E[X])^2]$$

Overview

Probability Distributions

- Only defined for **random variables**, *not* for events of logic statements!
  - **Discrete random variables**: $p: \mathbb{Z} \mapsto [0,1]$ and $\sum_x p(x) = 1$
  - **Continuous random variables**: $p: \mathbb{R} \mapsto \mathbb{R}^+$ and $\int p(x)dx = 1$
    - Note that, by definition, they are only a **model** for real data!
- In computational statistics some classes of probability distributions have emerged whose distributions can be fully described with a small number of parameters $\boldsymbol{\theta} \in \mathbb{R}^d$
  - **Advantages**:
    1. **Storage Efficiency**: Only d real numbers for whole function!
    2. **Compute Efficiency**: Only $O(d)$ computation for rules of probability!
  - **Disadvantages**:
    1. Too restrictive to represent true phenomena in real data
    2. Function classes often not closed under Bayes' rule

Bayes' Rule for Random Variables

$$\boxed{p(\theta|x) = \frac{p(x|\theta) \cdot p(\theta)}{p(x)}}$$

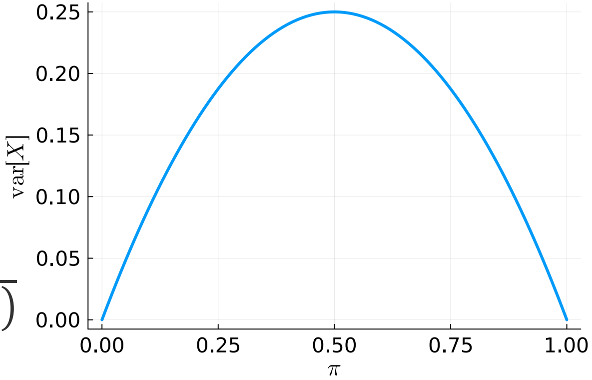
Probability Distributions: Bernoulli

- **Bernoulli Distribution**. *A random variable which only takes the values* 0 *and* 1 *is said to have a* Bernoulli *distribution parameterized by the probability* π *of the outcome* 1

$$p_X(k) = \begin{cases} \pi & \text{if } k = 1 \\ 1 - \pi & \text{if } k = 0 \end{cases}$$

- **Machine Learning**: Distribution that is used for modelling classes of objects
- **Properties**:

$$E[X] = \pi$$
$$\mathrm{var}[X] = \pi(1 - \pi)$$



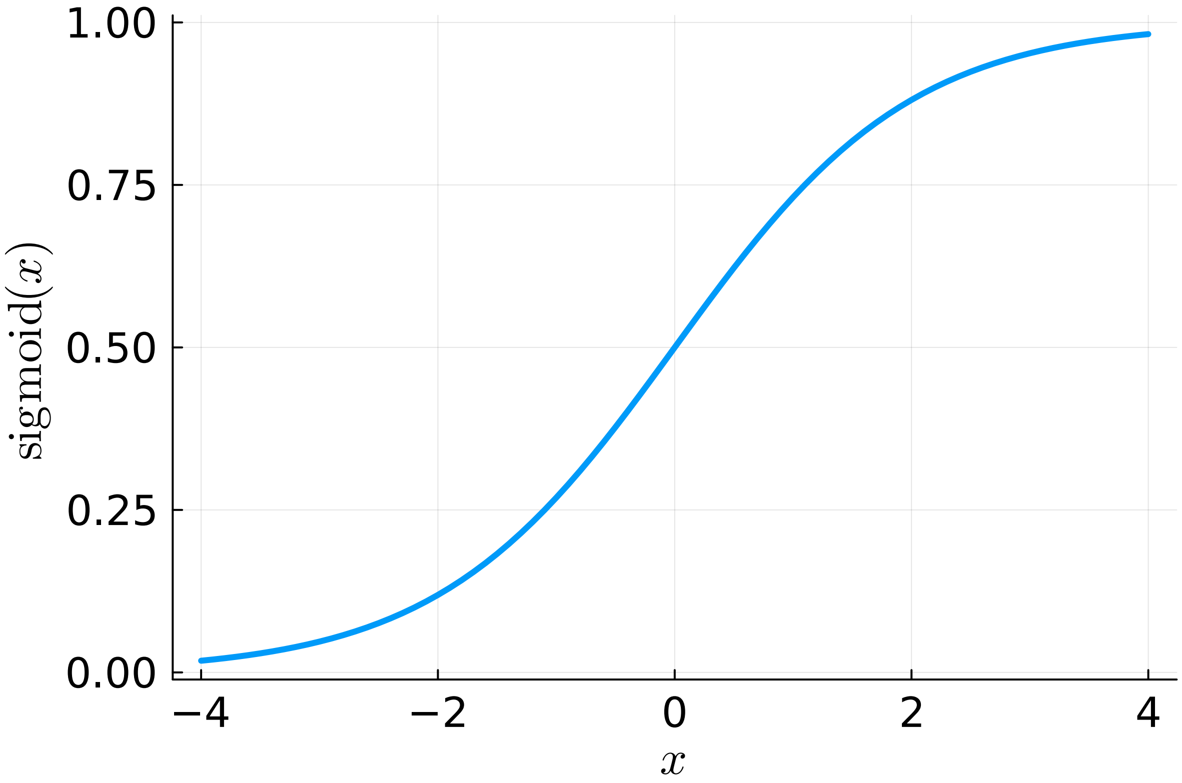
- In **machine learning**, the parameter π is often **parameterized** by a function of the inputs x

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

Jacob Bernoulli (1655 – 1705)

Probability Distributions: Logistic Function $\sigma(x)$

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



Introduction to Probabilistic Machine Learning

Unit 1 - Probability

Probability Distributions: Binomial & Beta Distribution

- **Binomial Distribution**. *The sum of* n *independent Bernoulli random variables with the same success probability* π *has a* Binomial distribution *with*

$$\forall k \in \{0,1,\ldots,n\}: \qquad p_X(k) = \binom{n}{k} \pi^k (1-\pi)^{n-k}$$

- **Rarely** used in Machine Learning practice but in Machine Learning theory (for modelling the distribution of the *number* of prediction errors)

- **Properties**:

$$E[X] = n\pi$$
$$\mathrm{var}[X] = n\pi(1-\pi)$$

Jacob Bernoulli (1655 – 1705)

Probability Distributions: Normal

- **Normal Distribution**. *A continuous random variable X is said to have a standard normal distribution if the density is given by*

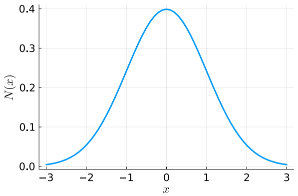
$$p(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- **Properties**:

$$E[X] = \mu$$
$$\mathrm{var}[X] = \sigma^2$$

- **Importance**. The Normal distribution plays a fundamental role in ML!
  - **Data Modelling**: The limit distribution for the sum of a large number of indepedent and identically distributed random variables.
  - **Machine Learning**: The most common belief distribution for the parameters of prediction functions!
  - **Information Theory**: The distribution function with the most uncertainty ("entropy") when fixing mean and variance of the random variable.

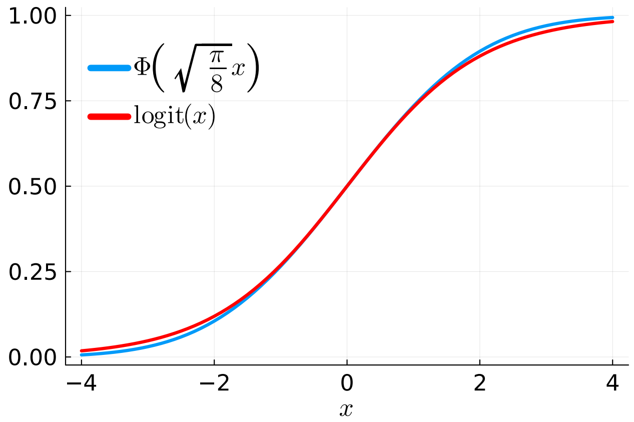
Carl Friedrich Gauss (1777 - 1855)



Probability Distribution: Normal (ctd)

- **Cumulative Density Function (CDF).** The Normal CDF is defined by

$$\Phi(x;\mu,\sigma^2) := \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right) dt$$



- Approximation using the Erf function: $\Phi(x) = \frac{1+\mathrm{Erf}(x/\sqrt{2})}{2}$ (Erf function is implimented in many math libraries)
  - **Numerical Recipes**. $\mathrm{Erf}(x) \approx 1 - t \cdot \exp\left(-x^2 + \sum_{i=0}^{9} a_i t^i\right), t = \left(1 + \frac{1}{2}|x|\right)^{-1}$

Introduction to Probabilistic Machine Learning

Unit 1 - Probability

Summary

- **History of Machine Learning**
  - Machine learning is a 70-year-old field of research
  - Key step in artificial intelligence improving on a task based on data
- **Probability in Machine Learning**
  - Probability is not a physical quantity but a mathematical model of uncertainty
  - Two different axiomatic justifications of the same math: one for data and one for parameters!
- **Probability Theory**
  - Two key rules of probability theory: Sum rule & Product (Bayes') rule
  - Independence is a concept of probability; it does not require random variables!
  - A random variable is a real-valued function of the outcome of the experiment
- **Probability Distributions**
  - The Bernoulli distribution with a sigmoid link function is key for classification learning
  - The normal distribution is centrally important in probabilistic machine learning

See you next week!