Probabilistic Machine Learning

Course Overview

Ralf Herbrich

Course Overview

- **Context**: Probabilistic machine learning …
  - is highly data-efficient
  - **can** be highly scalable (approximate approximate inference!)
  - offers explainable prediction and uncertainty quantification
  - has gained a lot of practical relevance over the past 15 years
  - allows practitioners to easily incorporate domain expertise
- **Goal**:
  1. Introduce the *basic ideas* and learning & prediction algorithms of probabilistic machine learning
  2. Familiarize you with *developing* algorithms for probabilistic machine learning.
- **Requirements**:
  - **Theory**: Linear algebra and calculus
  - **Practice**: Programming in Julia

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

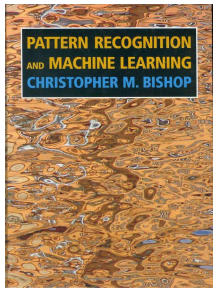
Probabilistic Machine Learning

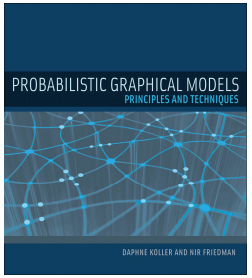
Course Structure

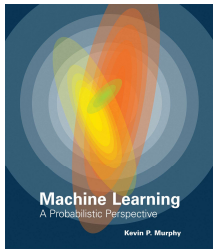
- **Introduction**
  1. Course Overview
  2. What is Machine Learning?
  3. The Role of Probability in Machine Learning
  4. Introduction to Probability Theory
  5. Probability Distributions
- **Graphical Models**
  6. Bayesian Networks
  7. Factor Graphs
- **Bayesian Learning Algorithms**
  8. Bayesian Ranking (TrueSkill)
  9. Bayesian Linear Regression
  10. Gaussian Processes
  11. Bayesian Classification
  12. Classification
- **Real World Data**
  13. Modelling Data

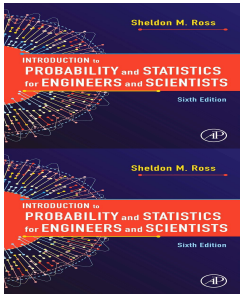
Literature

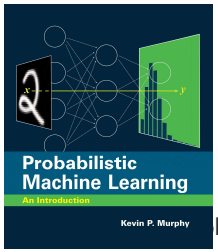
1. Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.
2. Koller, D., and N. Friedman. (2009). *Probabilistic Graphical Models: Principles and Techniques*. MIT Press.
3. Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. MIT Press.
4. Ross, S. (2021). *Introduction to Probability and Statistics for Engineers and Scientists*. Academic Press.











Lecturer

- **Short CV**
  - **1992-1997**: Diploma in Computer Science (major in AI & Economics)
  - **1997-2000**: PhD in Statistics in 2000
  - **2000-2011**: PostDoc + Researcher at Microsoft Research
    - TrueSkill™ (Halo 3), Drivatars™ (Forza Motorsport), adPredictor
  - **2011-2012**: Software Engineer at Facebook (Ads Optimization)
  - **2012-2020**: Director for Core ML/AI at Amazon
    - (Neural) Machine Translation of product catalog
    - Food quality prediction for produce items
  - **2020-2022**: Senior Vice President for AI at Zalando
  - **Since 2022**: Professor for AI & Sustainability at Hasso-Plattner Institute

Hasso Plattner Institut
Digital Engineering · Universität Potsdam

Have Fun Learning Probabilistic Machine Learning!

Probabilistic Machine Learning

What is Machine Learning?

Ralf Herbrich

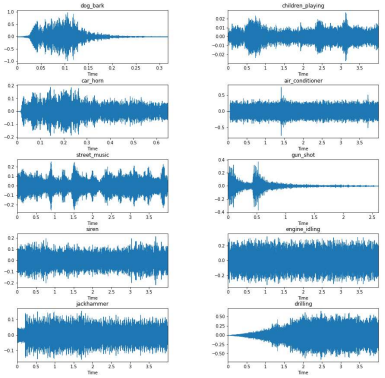
History of Machine Learning

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: $l(\hat{y}, y) = \boldsymbol{C}_{\hat{y},y} \in \mathbb{R}^{K \times K}$

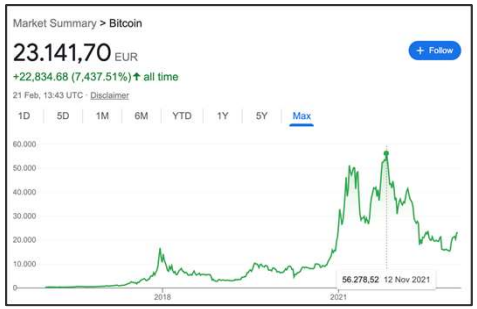


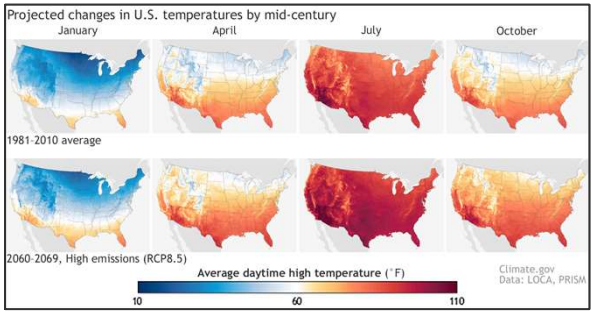
Probabilistic Machine Learning

What is Machine Learning?

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: $l(\hat{y}, y) = h(\hat{y} - y)$





Hasso Plattner Institut
Digital Engineering · Universität Potsdam

Have Fun Learning Probabilistic Machine Learning!

Probabilistic Machine Learning

The Role of Probability in Machine Learning

Ralf Herbrich

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 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!

Rules of Probability

- **Mathematical Definition**. *A number $P(A) \in [0,1]$ assigned to an event A that indicates how likely A is to occur.*

- **Rules**:
  - **Monotonicity**: If $A \subseteq B$ then $P(A) \leq P(B)$
  - **Complement Rule**: $P(A^c) = 1 - P(A)$ for all A
  - **Sum Rule**: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ for all A, B
  - **Product Rule**: $P(A \cap B) = \underbrace{\frac{P(A \cap B)}{P(B)}}_{P(A|B)} \cdot P(B)$

- History in gambling (ca. 1600), experimentation (ca. 1900) and AI (ca. 2000)!

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* $A \subseteq S$
  2. $P(S) = 1$
  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?*

  $$P\big(\text{Accuracy}\big(\mathcal{A}(D)\big) < \varepsilon\big) \leq \delta$$

- **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?*

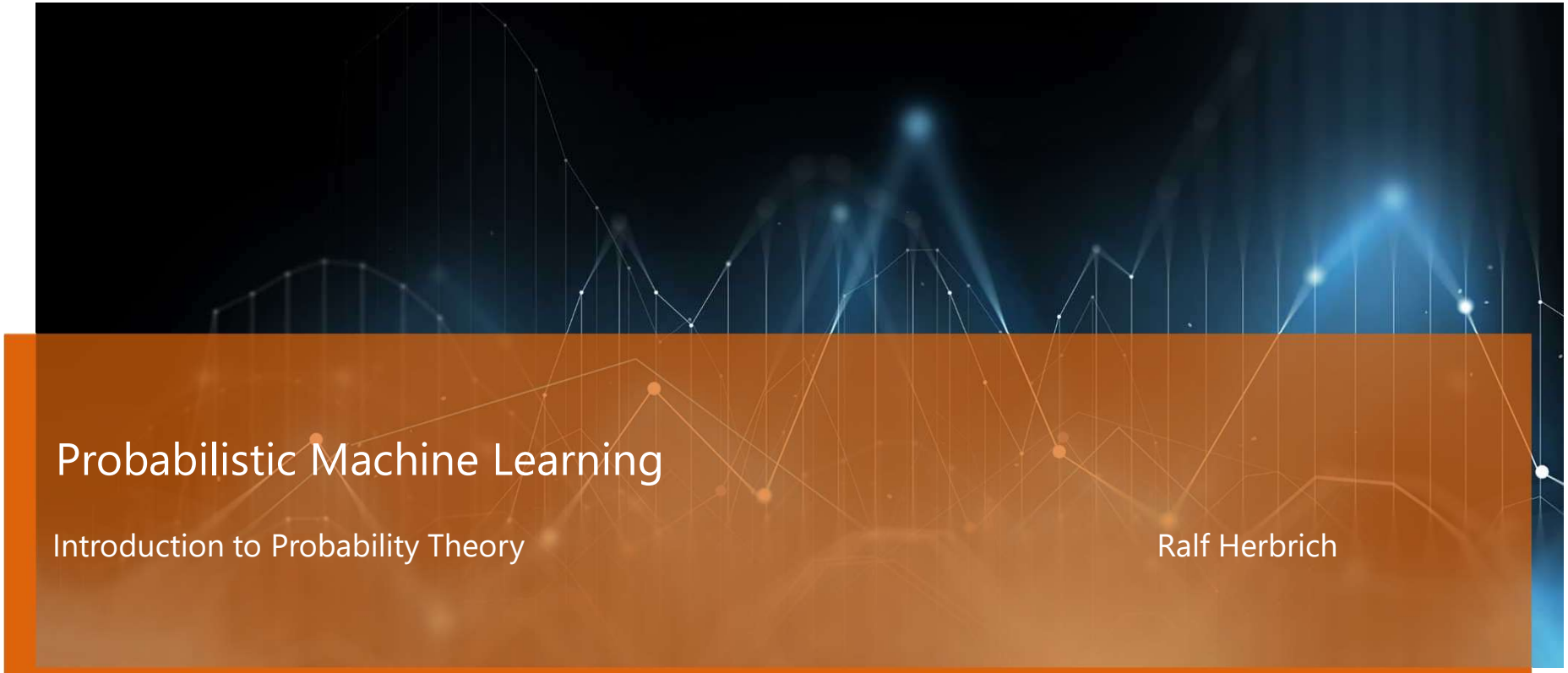
  $$P(f|D) = \frac{P(D \wedge f)}{P(D)} = \frac{P(D|f)P(f)}{P(D)}$$

- **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)

Hasso Plattner Institut
Digital Engineering · Universität Potsdam

Thank You!



Probabilistic Machine Learning

Introduction to Probability Theory

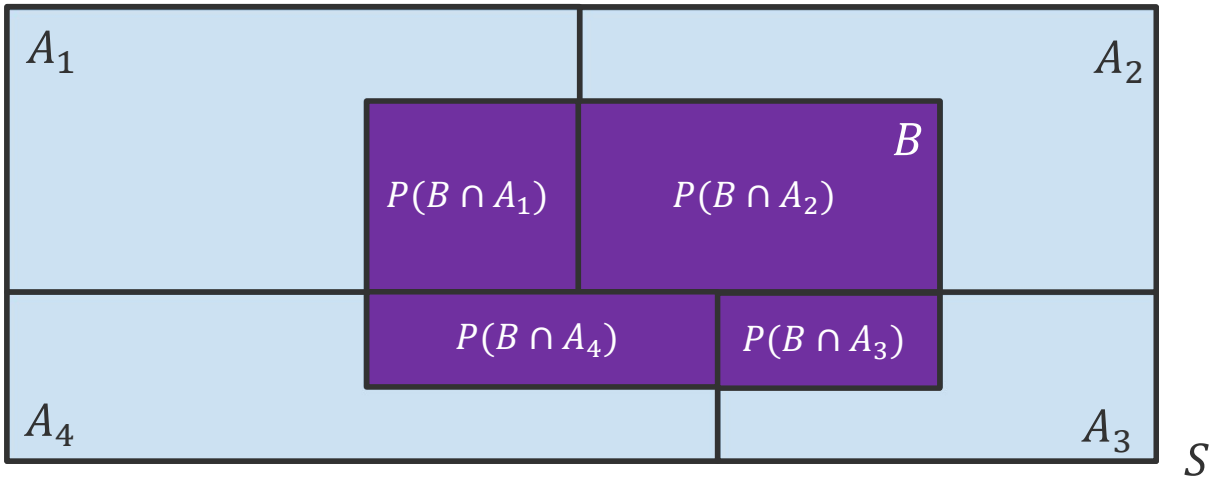
Ralf Herbrich

Probability Theory: Sum Rule (Marginalization)

- **Total Probability 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*

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

- **Geometric Proof**



Probability Theory: Product Rule (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 $P(B) > 0$*

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

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

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

=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

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

posterior — likelihood — prior

Bayes Rule: False-Positive Puzzle

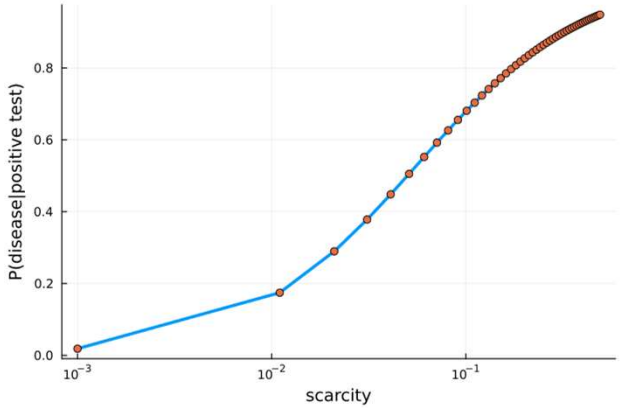
- **Situation**: A test for rare disease is assumed to be correct 95% of the time (i.e., the probability that the test shows the disease or lack thereof is 95%). It's a rare disease that occurs in 0.1% of the population. If you have a positive test outcome, what is the probability that you have the diseases?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

$$A = \text{"Person has the disease"}$$
$$B = \text{"Test result is positive"}$$

likelihood: $P(B|A)$ — prior: $P(A)$ — posterior: $P(A|B)$

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

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



- **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 $A_1, A_2, \dots, A_n$ are independent if*

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

- **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 machine learning:
  1. Knowing one training example provides no information about the probability of any other training example (realistic?!)
  2. Knowing the noise in one training example provides no information about the noise of another 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

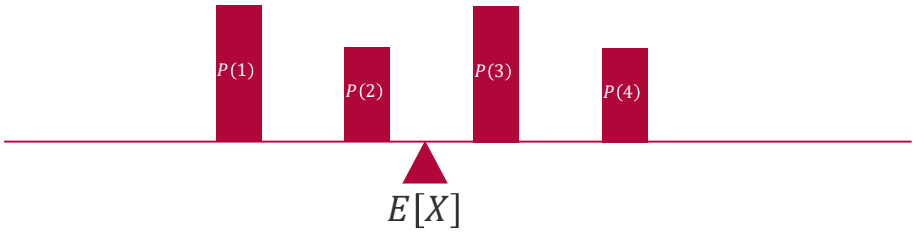
$$p(0) = P(\text{tail}, \text{tail})$$
$$p(1) = P(\text{head}, \text{tail}) + P(\text{tail}, \text{head})$$
$$p(2) = P(\text{head}, \text{head})$$

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* $\text{var}[X]$ *of a random variable* X *is defined by*

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

Hasso Plattner Institut
Digital Engineering · Universität Potsdam

Thank You!

Probabilistic Machine Learning

Probability Distributions

Ralf Herbrich

Recap: 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

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

Families of Probability Distributions

- In computational statistics some classes of probability distributions have emerged whose distributions can be fully described with a small number of parameters $\boldsymbol{\theta} \in \mathbb{R}^d$
  - Bernoulli Distribution
  - Normal Distribution

- **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 phenomenon in real data
  2. Function classes often not closed under Bayes' rule (conjugancy!)

Probability Distributions: Conjugacy

- **Bayes Rule for Random Variables**. *For any probability distribution p over two random variables X and Θ, it holds*

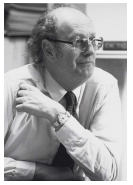
$$\text{Posterior} \quad p(\theta|x) = \frac{\overbrace{p(x|\theta) \cdot p(\theta)}^{p(x,\theta)}}{p(x)} \quad \text{Likelihood} \quad \text{Prior}$$

- **Conjugacy**. *A family $\{p(x,\theta)\}_{x,\theta}$ is conjugate if the posterior $p(\theta|x)$ is part of the same family as the prior $p(\theta)$ for any value of x.*

Likelihood $p(x\|\theta)$	Model Parameter	Conjugate Prior $p(\theta)$
$\text{Ber}(x;\pi)$	π	$\text{Beta}(\pi;\alpha,\beta)$
$\text{Bin}(x;n,\pi)$	π	$\text{Beta}(\pi;\alpha,\beta)$
$\mathcal{N}(x;\mu,\sigma^2)$	μ,σ^2	$\mathcal{N}(\mu;m,s^2)$

- **Big Advantage**: Computing the exact posterior is computationally efficient!

Hasso Plattner Institut
Digital Engineering · Universität Potsdam



Howard Raiffa
(1924 – 2016)

Robert Osher Schlaifer
(1914 – 1994)

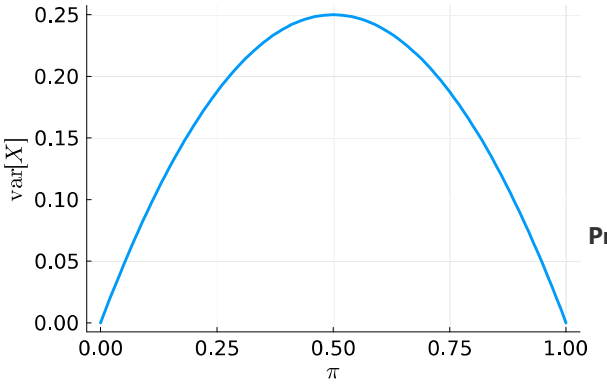
Probability Distributions: Bernoulli

- **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 *if*

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

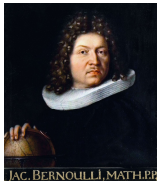
- **Machine Learning**: Distribution that is used for modelling classes of objects
- **Properties**:

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



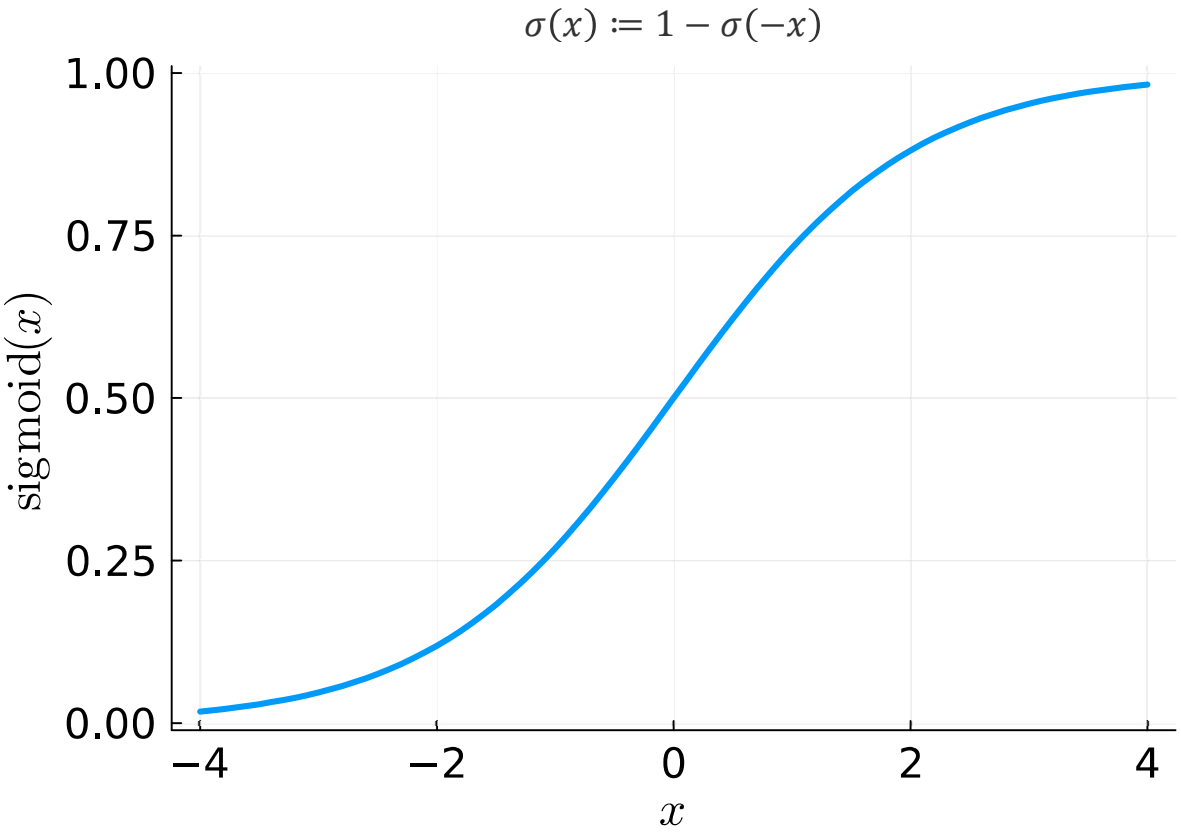
- In **machine learning**, the parameter π is often **parameterized** by a function of the inputs x

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



Jacob Bernoulli (1655 – 1705)

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



Probability Distributions: Normal

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

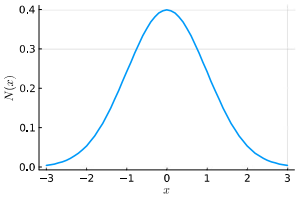
$$p_X(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- **Properties**:

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

- **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 prior 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)



Normal Distribution: Representations

- **Two Parameterizations (for different purposes)**:
  - **Scale-Location Parameters**

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

  - **Natural Parameters**

$$\mathcal{G}(x;\tau,\rho) = \sqrt{\frac{\rho}{2\pi}}\cdot\exp\left(-\frac{\tau^2}{2\rho}\right)\cdot\exp\left(\tau\cdot x-\rho\cdot\frac{x^2}{2}\right)$$

- **Conversions**

Two divisions only!

$$\mathcal{N}(x;\mu,\sigma^2) = \mathcal{G}\left(x;\frac{\mu}{\sigma^2},\frac{1}{\sigma^2}\right). \qquad \mathcal{G}(x;\tau,\rho) = \mathcal{N}\left(x;\frac{\tau}{\rho},\frac{1}{\rho}\right)$$

- **Two Special Cases**
  1. **Constant function**: $c(x) = 1 = \exp(0) = \lim_{\sigma^2\to\infty}\exp\left(-\frac{x^2}{\sigma^2}\right) = \frac{\mathcal{G}(x;0,0)}{\mathcal{N}(0;0,0)}$
  2. **Dirac Delta**: $\delta(x) = \lim_{\sigma^2\to 0}\mathcal{N}(x;0,\sigma^2)$

Normal Distributions and the Product Rule

- **Theorem (Multiplication)**. *Given two one-dimensional Gaussian distributions* $\mathcal{G}(x;\tau_1,\rho_1)$ *and* $\mathcal{G}(x;\tau_2,\rho_2)$ *we have*

$$\mathcal{G}(x;\tau_1,\rho_1,\gamma_1)\cdot\mathcal{G}(x;\tau_2,\rho_2,\gamma_2)=\mathcal{G}(x;\tau_1+\tau_2,\rho_1+\rho_2)\cdot\mathcal{N}(\mu_1;\mu_2,\sigma_1^2+\sigma_2^2)$$

Gaussian density

Additive updates!

- **Theorem (Division)**. *Given two one-dimensional Gaussian distributions* $\mathcal{G}(x;\tau_1,\rho_1)$ *and* $\mathcal{G}(x;\tau_2,\rho_2)$ *we have*

$$\frac{\mathcal{G}(x;\tau_1,\rho_1,\gamma_1)}{\mathcal{G}(x;\tau_2,\rho_2,\gamma_2)}=\mathcal{G}(x;\tau_1-\tau_2,\rho_1-\rho_2)\cdot\frac{1}{\mathcal{N}\left(\frac{\tau_1-\tau_2}{\rho_1-\rho_2};\frac{\tau_2}{\rho_2},\frac{1}{\rho_1-\rho_2}+\frac{1}{\rho_2}\right)}$$

Gaussian density

Subtractive updates!

Hasso Plattner Institut
Digital Engineering · Universität Potsdam

Thank You!