*** Applied Machine Learning Fundamentals *** Mathematical Foundations

Daniel Wehner

SAP SE

November 18, 2019

Find all slides on GitHub

Lecture Overview

Unit I	Machine Learning Introduction
Unit II	**Mathematical Foundations**
Unit III	Bayesian Decision Theory
Unit IV	Probability Density Estimation
Unit V	Regression
Unit VI	Classification I
Unit VII	Evaluation
Unit VIII	Classification II
Unit IX	Clustering
Unit X	Dimensionality Reduction

Agenda November 18, 2019

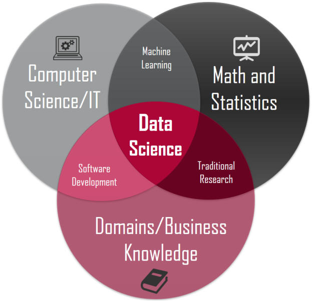
Section:

Introduction

Duale Hochschule
Baden-Württemberg

Introduction

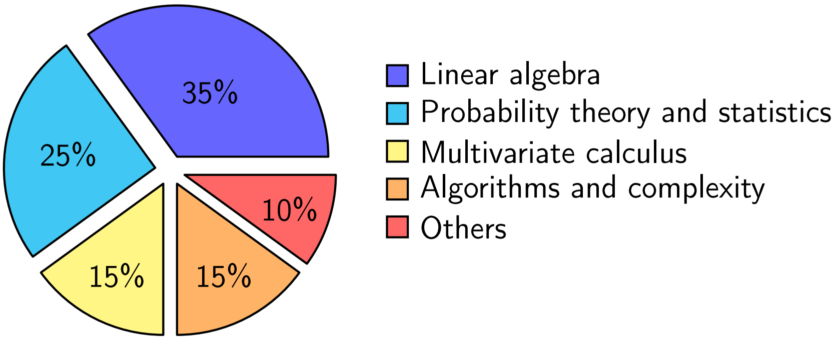
Math is a significant portion of data science / machine learning!



You will need it to understand:

- **Statistical** machine learning
- How optimization for learning / empirical risk minimization works,
- How linear algebra, calculus and statistics are used to make learning and inference more efficient

Math is important!



Section:

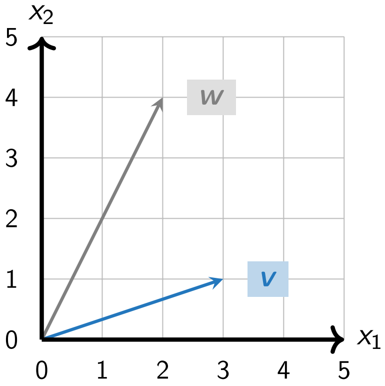
Linear Algebra

What is a Vector?

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$\boldsymbol{v} = \begin{bmatrix} 3 \\ 1 \end{bmatrix}$$

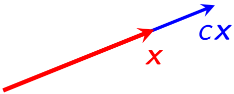
$$\boldsymbol{w} = \begin{bmatrix} 2 \\ 4 \end{bmatrix}$$



Multiplication by a Scalar

$$c\boldsymbol{x} = c\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} cx_1 \\ cx_2 \end{bmatrix}$$

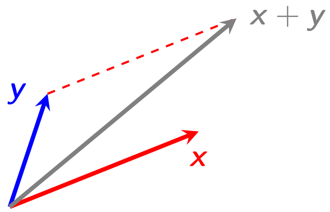
$$2\boldsymbol{v} = 2\begin{bmatrix} 3 \\ 1 \end{bmatrix} = \begin{bmatrix} 6 \\ 2 \end{bmatrix}$$



Addition of Vectors

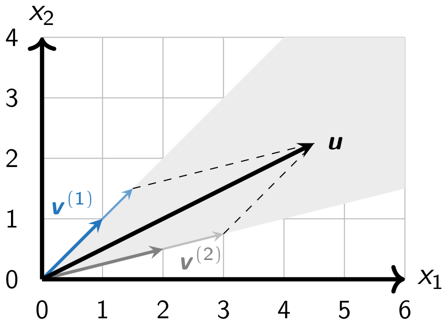
$$\boldsymbol{x} + \boldsymbol{y} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} x_1 + y_1 \\ x_2 + y_2 \end{bmatrix}$$

$$\boldsymbol{v} + \boldsymbol{w} = \begin{bmatrix} 3 \\ 1 \end{bmatrix} + \begin{bmatrix} 2 \\ 4 \end{bmatrix} = \begin{bmatrix} 5 \\ 5 \end{bmatrix}$$



Linear Combination of Vectors

$$\boldsymbol{u} = c_1 \boldsymbol{v}^{(1)} + c_2 \boldsymbol{v}^{(2)} + \cdots + c_n \boldsymbol{v}^{(n)} \tag{1}$$



Vector Transpose, inner and outer Product

- Vector transpose:

$$\boldsymbol{v} = \begin{bmatrix} 3 \\ 1 \end{bmatrix} \qquad \boldsymbol{v}^\intercal = \begin{bmatrix} 3 & 1 \end{bmatrix}$$

- Inner product / dot product / scalar product:

$$\boldsymbol{v} \cdot \boldsymbol{w} \equiv \boldsymbol{v}^\intercal \boldsymbol{w} \equiv \langle \boldsymbol{v}, \boldsymbol{w} \rangle = \sum_{j=1}^{m} v_j w_j \tag{2}$$

$$= \begin{bmatrix} 3 & 1 \end{bmatrix} \begin{bmatrix} 2 \\ 4 \end{bmatrix} = (3 \cdot 2) + (1 \cdot 4) = 10$$

Important

Vector Transpose and inner and outer Product (Ctd.)

- Outer product:

$$\boldsymbol{v}\boldsymbol{w}^{\mathsf{T}} = \begin{bmatrix} 3 \\ 1 \end{bmatrix} \begin{bmatrix} 2 & 4 \end{bmatrix} = \begin{bmatrix} 6 & 12 \\ 2 & 4 \end{bmatrix}$$

The inner product yields a scalar value, the results of an outer product is a matrix!

Length of a Vector

- Length of a vector (Frobenius norm):

$$\|\boldsymbol{x}\| = \sqrt{\boldsymbol{x}^\intercal \boldsymbol{x}} \tag{3}$$

$$\|c\boldsymbol{x}\| = |c| \cdot \|\boldsymbol{x}\| \tag{4}$$

$$\|\boldsymbol{x} + \boldsymbol{y}\| \leqslant \|\boldsymbol{x}\| + \|\boldsymbol{y}\| \tag{5}$$

- Example:

$$\|\boldsymbol{v}\| = \sqrt{3^2 + 1^2} = 10$$

Angle between Vectors

- The angle between two vectors is given by:

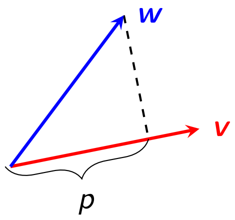
$$\cos \measuredangle(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x} \cdot \boldsymbol{y}}{\|\boldsymbol{x}\| \cdot \|\boldsymbol{y}\|} = \frac{\sum_{j=1}^{m} x_j \cdot y_j}{\sqrt{\sum_{j=1}^{m} (x_j)^2} \cdot \sqrt{\sum_{j=1}^{m} (y_j)^2}} \tag{6}$$

$$\cos \measuredangle(\boldsymbol{v}, \boldsymbol{w}) = \frac{\boldsymbol{v} \cdot \boldsymbol{w}}{\|\boldsymbol{v}\| \cdot \|\boldsymbol{w}\|} = \frac{10}{\sqrt{10} \cdot \sqrt{20}} \approx 0.71$$

- Inner product: $\boldsymbol{x} \cdot \boldsymbol{y} = \|\boldsymbol{x}\| \cdot \|\boldsymbol{y}\| \cdot \cos \measuredangle(\boldsymbol{x}, \boldsymbol{y})$

Projection of Vectors

- How is the projection of $\boldsymbol{x}$ onto $\boldsymbol{y}$ defined?
- Formally, we have:



$$\begin{aligned} p &= \|\boldsymbol{v}\| \cos \measuredangle(\boldsymbol{v}, \boldsymbol{w}) \\ &= \|\boldsymbol{v}\| \frac{\boldsymbol{v} \cdot \boldsymbol{w}}{\|\boldsymbol{v}\| \cdot \|\boldsymbol{w}\|} \\ &= \frac{\boldsymbol{v} \cdot \boldsymbol{w}}{\|\boldsymbol{w}\|} \end{aligned} \tag{7}$$

- Note that p is **not a vector!**

What is a Matrix?

General case ($\mathbb{R}^{n \times m}$):

$$\boldsymbol{X} = \begin{bmatrix} X_{11} & X_{12} & \dots & X_{1m} \\ X_{21} & X_{22} & \dots & X_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \dots & X_{nm} \end{bmatrix}$$

$$\boldsymbol{M} = \begin{bmatrix} 3 & 4 & 5 \\ 1 & 0 & 1 \end{bmatrix} \qquad \mathbb{R}^{2 \times 3}$$

$$\boldsymbol{N} = \begin{bmatrix} 3 & 0 & 0 \\ 0 & 7 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad \mathbb{R}^{3 \times 3}$$

$$\boldsymbol{P} = \begin{bmatrix} 10 & 1 \\ 11 & 2 \end{bmatrix} \qquad \mathbb{R}^{2 \times 2}$$

Matrix Transpose and Addition

- Transpose of a matrix:

$$\boldsymbol{M}^{\intercal} = \begin{bmatrix} 3 & 4 & 5 \\ 1 & 0 & 1 \end{bmatrix}^{\intercal} = \begin{bmatrix} 3 & 1 \\ 4 & 0 \\ 5 & 1 \end{bmatrix} \tag{8}$$

- Addition of matrices:

$$\boldsymbol{X} + \boldsymbol{Y} = \begin{bmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{bmatrix} + \begin{bmatrix} Y_{11} & Y_{12} \\ Y_{21} & Y_{22} \end{bmatrix} = \begin{bmatrix} X_{11} + Y_{11} & X_{12} + Y_{12} \\ X_{21} + Y_{21} & X_{22} + Y_{22} \end{bmatrix} \tag{9}$$

Matrix Multiplication

- Multiplication by scalars:

$$c\boldsymbol{X} = c \begin{bmatrix} X_{11} & X_{12} & X_{13} \\ X_{21} & X_{22} & X_{23} \end{bmatrix} = \begin{bmatrix} c \cdot X_{11} & c \cdot X_{12} & c \cdot X_{13} \\ c \cdot X_{21} & c \cdot X_{22} & c \cdot X_{23} \end{bmatrix} \tag{10}$$

- Matrix-vector multiplication:

$$\boldsymbol{z} = \boldsymbol{X}\boldsymbol{y} = \begin{bmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} X_{11} \cdot y_1 + X_{12} \cdot y_2 \\ X_{21} \cdot y_1 + X_{22} \cdot y_2 \end{bmatrix} \tag{11}$$

Matrix Multiplication (Ctd.)

- Matrix-matrix multiplication:

$$
\begin{aligned}
Z &= XY \\
&= \begin{bmatrix} X_{11} & X_{12} & X_{13} \\ X_{21} & X_{22} & X_{23} \end{bmatrix} \begin{bmatrix} Y_{11} & Y_{12} \\ Y_{21} & Y_{22} \\ Y_{31} & Y_{32} \end{bmatrix} \\
&= \begin{bmatrix} X_{11}Y_{11} + X_{12}Y_{21} + X_{13}Y_{31} & X_{11}Y_{12} + X_{12}Y_{22} + X_{13}Y_{32} \\ X_{21}Y_{11} + X_{22}Y_{21} + X_{23}Y_{31} & X_{21}Y_{12} + X_{22}Y_{22} + X_{23}Y_{32} \end{bmatrix} \qquad (12)
\end{aligned}
$$

Matrix Inversion

- Matrix inversion is defined for **square matrices** $\boldsymbol{X} \in \mathbb{R}^{n \times n}$
- A matrix $\boldsymbol{X}$ multiplied by its inverse $\boldsymbol{X}^{-1}$ gives the identity matrix:

$$\boldsymbol{X}^{-1}\boldsymbol{X} = \boldsymbol{X}\boldsymbol{X}^{-1} = \boldsymbol{I} \tag{13}$$

$$\boldsymbol{I} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix} \tag{14}$$

- If $\boldsymbol{X}^{-1}$ exists, we say that $\boldsymbol{X}$ is non-singular

Matrix Inversion (Ctd.)

- It holds that ($\boldsymbol{C}$ is the cofactor matrix):

$$\boldsymbol{X}^{-1} = \frac{1}{\det(\boldsymbol{X})}\boldsymbol{C}^{\mathsf{T}} \tag{15}$$

- A condition for invertability is that **the determinant has to be different than zero**
- **Example:**

$$\boldsymbol{X} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \qquad \det(\boldsymbol{X}) = 0 \qquad \boldsymbol{X}^{-1} = ?$$

Matrix Inversion Example

$$\boldsymbol{X} = \begin{bmatrix} 1 & 1/2 \\ -1 & 1 \end{bmatrix} \qquad \boldsymbol{X}^{-1} = \begin{bmatrix} 2/3 & -1/3 \\ 2/3 & 2/3 \end{bmatrix}$$

Please verify!

$$\boldsymbol{X}\boldsymbol{X}^{-1} = \boldsymbol{I} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \boldsymbol{X}^{-1}\boldsymbol{X}$$

Use for example the Gauss-Jordan algorithm to find the inverse!

Matrix Pseudoinverse

- **Question:** How can we invert a matrix $\boldsymbol{X} \in \mathbb{R}^{n \times m}$ which is not squared?
- **Left pseudoinverse** $\boldsymbol{X}^{\#}\boldsymbol{X}$:

$$\boldsymbol{X}^{\#}\boldsymbol{X} = \underbrace{(\boldsymbol{X}^{\intercal}\boldsymbol{X})^{-1}\boldsymbol{X}^{\intercal}}_{\text{left-multiplied}}\boldsymbol{X} = \boldsymbol{I}_m \tag{16}$$

- **Right pseudoinverse** $\boldsymbol{X}\boldsymbol{X}^{\#}$:

$$\boldsymbol{X}\boldsymbol{X}^{\#} = \boldsymbol{X}\underbrace{\boldsymbol{X}^{\intercal}(\boldsymbol{X}\boldsymbol{X}^{\intercal})^{-1}}_{\text{right-multiplied}} = \boldsymbol{I}_n \tag{17}$$

Eigenvectors and Eigenvalues

- Some vectors $\boldsymbol{v}$ only change their length when multiplied by a matrix $\boldsymbol{X}$
- **Example:**

$$\begin{bmatrix} 4 & -1 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix} = 2 \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

- These vectors are called eigenvectors, the scaling factors are known as eigenvalues
- More general:

$$\boldsymbol{W}\boldsymbol{v} = \lambda \boldsymbol{v} \tag{18}$$

Eigenvectors form a Basis

- Let us assume that there are n eigenvectors with corresponding eigenvalues:

$$\boldsymbol{v}_1, \boldsymbol{v_2}, \ldots, \boldsymbol{v}_n$$

$$\lambda_1, \lambda_2, \ldots, \lambda_n$$

- **Theorem:**
  - For an $n \times n$ matrix with eigenvectors $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n$, if they correspond to **distinct** eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$, then the set $\{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n\}$ is linearly independent
  - Hence, any vector can be expressed as a linear combination of eigenvectors:

$$\boldsymbol{v} = c_1 \boldsymbol{v}_1 + c_2 \boldsymbol{v}_2 + \cdots + c_n \boldsymbol{v}_n$$

Symmetric Matrices

- A squared $n \times n$ matrix $\boldsymbol{X}$ is **symmetric**, iff

$$\forall i, j: \quad X_{ij} = X_{ji} \tag{19}$$

$$\boldsymbol{X} = \boldsymbol{X}^\intercal \tag{20}$$

- Some properties:
  - The inverse $\boldsymbol{X}^{-1}$ is also symmetric
  - **Eigen-decomposition:** $\boldsymbol{X}$ can be decomposed into $\boldsymbol{X} = \boldsymbol{Q}\boldsymbol{D}\boldsymbol{Q}^\intercal$, where the columns of $\boldsymbol{Q}$ are the eigenvectors of $\boldsymbol{X}$, and $\boldsymbol{D}$ is a diagonal matrix whose entries are the corresponding eigenvalues

Positive (semi-)definite Matrices

- A **squared symmetric** matrix $\boldsymbol{X}^{n\times n}$ is positive definite, iff for any vector $\boldsymbol{y} \in \mathbb{R}^n$:

$$\boldsymbol{y}^\intercal \boldsymbol{X} \boldsymbol{y} > 0 \tag{21}$$

- Or positive semi-definite, iff $\boldsymbol{y}^\intercal \boldsymbol{X} \boldsymbol{y} \geqslant 0$

Such matrices are important in machine learning. For instance, the covariance matrix is always positive semi-definite.

Section:

Statistics

Duale Hochschule
Baden-Württemberg

Random Variables

- What is a random variable?

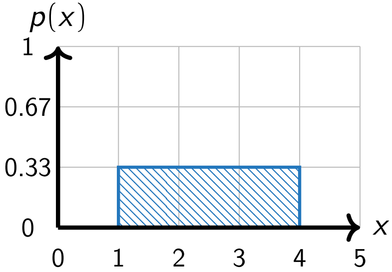
Random Variables

- What is a **random variable**?
  - It's a random number determined by chance (according to a distribution)
  - Random variables in machine learning: input data, output data, noise
- What is a **probability distribution**?

Random Variables

- What is a **random variable**?
  - It's a random number determined by chance (according to a distribution)
  - Random variables in machine learning: input data, output data, noise
- What is a **probability distribution**?
  - Describes the probability that a random variable is equal to a certain value
  - It can be given by the physics of an experiment (e. g. throwing dice)
  - **Discrete** vs. **continuous** distributions

Uniform Distribution



Every outcome is equally probable within a bounded region $\mathcal{R}$

$$p(x) = {}^{1}\!/_{\mathcal{R}} \tag{22}$$

Discrete Distributions

The random variables take on **discrete values**

Examples:

- When throwing a die, the possible values are given by a countably finite set:

$$x_i \in \{1, 2, 3, 4, 5, 6\}$$

- The number of sand grains at the beach (countably infinite set):

$$x_i \in \mathbb{N}$$

Discrete Distributions (Ctd.)

- All probabilities sum up to 1:

$$\sum_i p(x_i) = 1$$

- Discrete distributions are particularly important in classification
- A discrete distribution is described by a **probability mass function** (also called frequency function)

Bernoulli Distribution

- A **Bernoulli random variable** only takes on two values (e. g. 0 and 1):

$$x \in \{0, 1\} \tag{23}$$

$$p(x = 1|\mu) = \mu \tag{24}$$

$$\mathrm{Bern}(x|\mu) = \mu^x (1 - \mu)^{1-x} \tag{25}$$

$$\mathbb{E}\{x\} = \mu \tag{26}$$

$$\mathrm{var}\{x\} = \mu(1 - \mu) \tag{27}$$

- The only parameter is μ, i. e. the distribution is completely defined by this parameter

Binomial Distribution

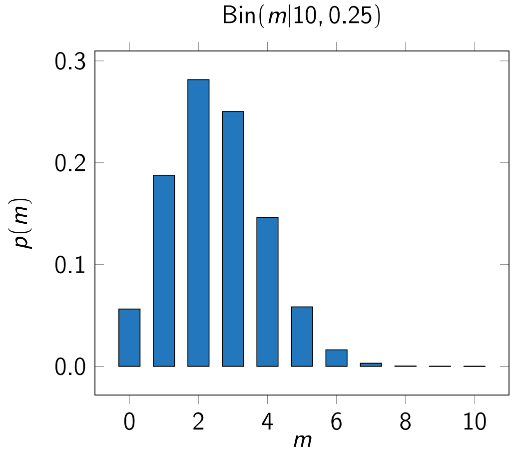
- **Binomial variables** are a sequence of n repeated Bernoulli variables
- **Example:** What is the probability of getting $m \in \mathbb{N}$ heads in n trials?

$$\mathrm{Bin}(m|n, \mu) = \binom{n}{m} \mu^m (1 - \mu)^{n-m} \tag{28}$$

$$\mathbb{E}\{m\} = n\mu \tag{29}$$

$$\mathrm{var}\{m\} = n\mu(1 - \mu) \tag{30}$$

Binomial Distribution (Ctd.)



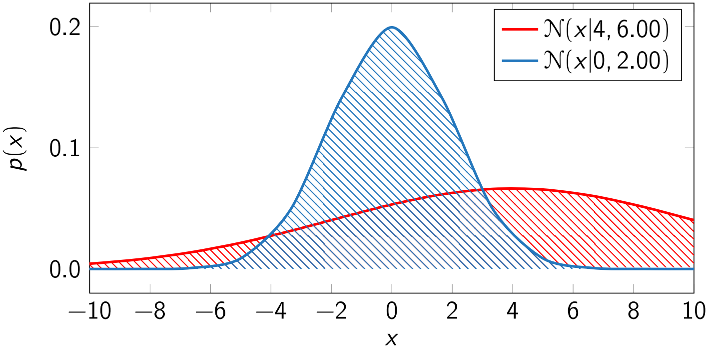
Continuous Distributions

The random variables take on **continuous values**

- Continuous distributions are discrete distributions where the **number of discrete values goes to infinity** while the **probability of each value goes to zero**
- It's described by a probability density function which integrates to 1:

$$\int_{-\infty}^{+\infty} p(x)\,\mathrm{d}x = 1$$

Gaussian Distribution



$$p(x) = \mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \tag{31}$$

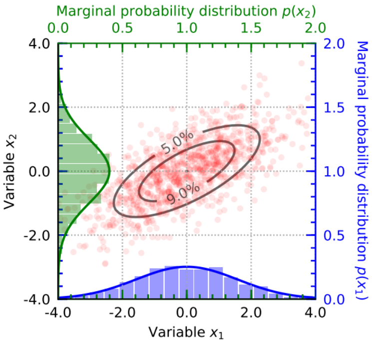
Central Limit Theorem

Central Limit Theorem:
The distribution of the sum of n i. i. d. (independent and identically distributed) random variables becomes increasingly Gaussian as n increases.

- The Gaussian distribution is one among the most important distributions
- Gaussians are often a good model
- Working with Gaussians leads to **analytical solutions for complex operations**

Multivariate Gaussian Distribution

$$p_D(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^D|\boldsymbol{\Sigma}|}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\intercal \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\} \qquad (32)$$



For clarification: $\boldsymbol{x}$ and $\boldsymbol{\mu}$ are vectors while $\boldsymbol{\Sigma}$ is a matrix. The probability given by $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in [0; 1]$ is still a scalar value!

Basic Rules of Probability

- **Joint distribution:**

$$p(x, y) \tag{33}$$

- **Marginal distribution:**

$$p(y) = \int_x p(x, y) \, \mathrm{d}x \tag{34}$$

- **Conditional distribution:**

$$p(y|x) = \frac{p(x, y)}{p(x)} \tag{35}$$

Basic Rules of Probability (Ctd.)

- **Probabilistic independence:**

$$p(x, y) = p(x)p(y) \tag{36}$$

- **Chain rule of probabilities:**

$$\begin{aligned} p(x_1, \ldots, x_n) &= p(x_1|x_2, \ldots, x_n)p(x_2, \ldots, x_n) \\ &= p(x_1|x_2, \ldots, x_n)p(x_2|x_3, \ldots, x_n) \ldots p(x_{n-1}|x_n)p(x_n) \end{aligned} \tag{37}$$

- **Bayes' rule:**

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)} \tag{38}$$

Expectation

$$\mathbb{E}_{x \sim p(x)}\{f(x)\} = \mathbb{E}_x\{f\} = \mathbb{E}\{f\} = \sum_x p(x) f(x) \qquad \text{discrete case} \tag{39}$$

$$= \int_x p(x) f(x)\, \mathrm{d}x \qquad \text{continuous case} \tag{40}$$

Approximate expectation:

$$\mathbb{E}\{f\} = \int_x p(x) f(x)\, \mathrm{d}x \approx \frac{1}{n} \sum_{i=1}^{n} f(x_i) \tag{41}$$

Expectation (Ctd.)

- Some rules of expectations:
  - $\mathbb{E}\{a\boldsymbol{x}\} = a\mathbb{E}\{\boldsymbol{x}\}$
  - $\mathbb{E}\{\boldsymbol{x} + \boldsymbol{y}\} = \mathbb{E}\{\boldsymbol{x}\} + \mathbb{E}\{\boldsymbol{y}\}$
  - $\mathbb{E}\{\boldsymbol{x}\boldsymbol{y}\} = \mathbb{E}\{\boldsymbol{x}\}\mathbb{E}\{\boldsymbol{y}\}$ (if $\boldsymbol{x}$ and $\boldsymbol{y}$ are independent)
  - $\mathbb{E}\{\sum_i a_i x_i\} = \sum_i a_i \mathbb{E}\{x_i\}$
- Expectations of functions:
  - $\mathbb{E}\{g(\boldsymbol{x})\} = \int_{\boldsymbol{x}} p(\boldsymbol{x}) g(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}$
  - In general: $\mathbb{E}\{g(\boldsymbol{x})\} \neq g(\mathbb{E}\{\boldsymbol{x}\})$

Variance and Covariance

- Covariances give a measure of correlation: (how much variables change together)
- Scalars:

$$\begin{aligned} \mathrm{cov}\{x, y\} &= \mathbb{E}_{x,y}\{(x - \mathbb{E}_x\{x\})(y - \mathbb{E}_y\{y\})\} \\ &= \mathbb{E}_{x,y}\{xy\} - \mathbb{E}_x\{x\}\mathbb{E}_y\{y\} \end{aligned} \tag{42}$$

- Vector notation:

$$\mathrm{cov}\{\boldsymbol{x}, \boldsymbol{y}\} = \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}\{(\boldsymbol{x} - \mathbb{E}_{\boldsymbol{x}}\{\boldsymbol{x}\})(\boldsymbol{y} - \mathbb{E}_{\boldsymbol{y}}\{\boldsymbol{y}\})^\intercal\} \tag{43}$$

Kullback-Leibler Divergence

- The **Kullback-Leibler (KL) divergence** is a similarity measure between two distributions p and q:

$$\mathrm{KL}(p\|q) = \sum_x p(x) \cdot \log \frac{p(x)}{q(x)} \tag{44}$$

- Some properties:
  - It is not a distance metric: $\mathrm{KL}(p\|q) \neq \mathrm{KL}(q\|p)$
  - It is non-negative: $\mathrm{KL}(p\|q) \geqslant 0$
  - If $\forall x : p(x) = q(x) \Rightarrow \mathrm{KL}(p\|q) = 0$

Section:

Optimization

Motivation

- In every machine learning problem, you will have:
  1. An **objective function** you want to optimize
  2. **Data** you want to learn from
  3. **Parameters** which need to be learned
  4. Assumptions about the problem and the data
- We would like to have general solutions to the problem of learning
- Different algorithms embody different objective functions and assumptions

Every machine learning problem is an optimization problem!

Unconstrained Optimization

You know how to do that, don't you?

Constrained Optimization

Formalization:

$$\begin{aligned} \min_{\boldsymbol{\theta}} \mathcal{J}(\theta) &= \dots && \longleftarrow \text{cost function / objective} \\ \text{s.t. } f(\theta) &= 0 && \longleftarrow \text{equality constraints} \\ g(\theta) &\geqslant 0 && \longleftarrow \text{inequality constraints} \end{aligned}$$

What should an ideal optimization problem, i. e. the cost function and constraints look like?

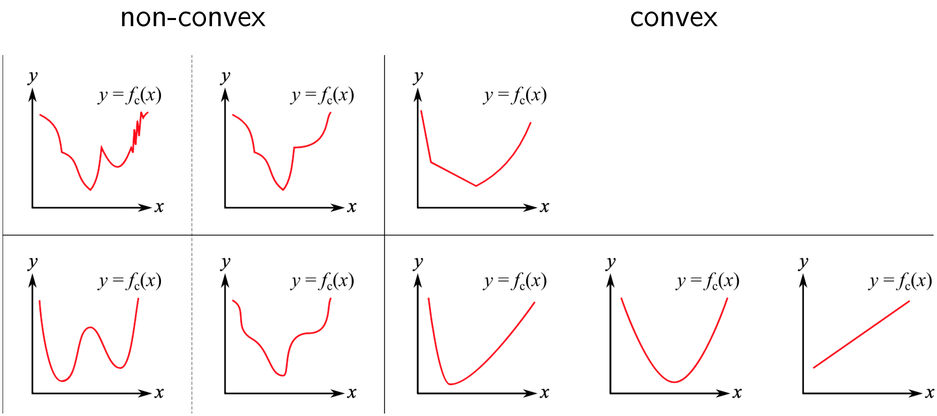
Constrained Optimization (Ctd.)

$$\min_{\boldsymbol{\theta}} \mathcal{J}(\theta) = \ldots \qquad \longleftarrow \text{convex function}$$

$$\text{s.t. } f(\theta) = 0 \qquad \longleftarrow \text{linear function}$$

$$g(\theta) \geqslant 0 \qquad \longleftarrow \text{convex set}$$

Cost Functions

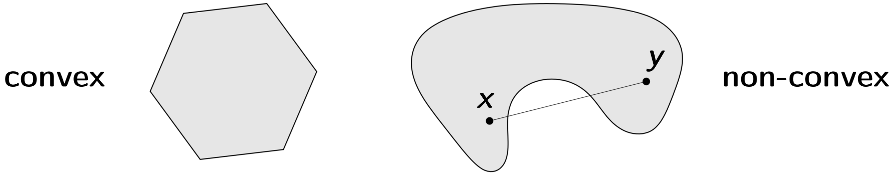


Convexity – Convex Sets

- A set $C \subseteq \mathbb{R}^n$ is convex, if $\forall \boldsymbol{x}, \boldsymbol{y} \in C$ and $\forall \alpha \in [0, 1]$

$$\alpha \boldsymbol{x} + (1 - \alpha)\boldsymbol{y} \in C \tag{45}$$

- This is the equation line segment between $\boldsymbol{x}$ and $\boldsymbol{y}$

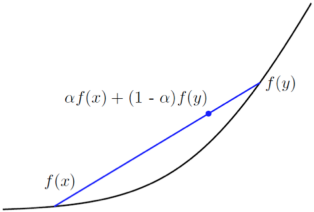


Convexity – Convex Functions

- A function $f : \mathbb{R}^n \mapsto \mathbb{R}$ is convex, if $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathrm{dom}(f)$ and $\forall \alpha \in [0, 1]$

$$f(\alpha \boldsymbol{x} + (1 - \alpha)\boldsymbol{y}) \leqslant \alpha f(\boldsymbol{x}) + (1 - \alpha) f(\boldsymbol{y}) \tag{46}$$

- Examples are linear functions $f(\boldsymbol{x}) = \boldsymbol{a}^\intercal \boldsymbol{x} + b$ and quadratic functions $f(\boldsymbol{x}) = \boldsymbol{x}^\intercal \boldsymbol{A} \boldsymbol{x} + \boldsymbol{b}^\intercal \boldsymbol{x} + c$



Convexity (Ctd.)

- Why are convex cost functions so appealing?
- Local solutions are global optima
- Efficient implementations of optimizers are available

Important

Constrained Optimization

- **How to solve this optimization problem?**

$$\min_{x,y} \mathcal{J}(x, y) = 2y + x$$

subject to (s. t.):

$$f(x, y) = y^2 + xy - 1 = 0$$

- Convert the problem to an unconstrained one
- This is done using **Lagrange multipliers** α

The Concept of Lagrange Multipliers

General Lagrange function: $\mathcal{L}(x, y, \lambda) = \mathcal{J}(x, y) + \lambda f(x, y)$

$$\min_{x,y} \mathcal{J}(x, y) = 2y + x$$

s. t.:

$$f(x, y) = y^2 + xy - 1 = 0$$

Step ❶: Differentiate w. r. t. x, y and λ:

I. $\nabla_x \mathcal{L} = 1 + \lambda y$

II. $\nabla_y \mathcal{L} = 2 + 2\lambda y + \lambda x$

III. $\nabla_\lambda \mathcal{L} = y^2 + xy - 1$

The Concept of Lagrange Multipliers (Ctd.)

Step ❷: Set equations to zero:

$$\begin{aligned} \text{I.} \quad & 1 + \lambda y && \stackrel{!}{=} 0 \\ \text{II.} \quad & 2 + 2\lambda y + \lambda x && \stackrel{!}{=} 0 \\ \text{III.} \quad & y^2 + xy - 1 && \stackrel{!}{=} 0 \end{aligned}$$

Step ❸: Substitute:

$$\begin{aligned} \text{I.} \quad & \lambda = -\frac{1}{y} \\ \text{I.} \rightarrow \text{II.} \quad & x = 0 \\ \text{II.} \rightarrow \text{III.} \quad & y = \pm 1 \end{aligned}$$

Numerical Optimization

- Different numerical optimization algorithms exist for optimizing a function numerically on a computer if we can't solve it analytically
- **Gradient descent:** Incrementally update an estimate of the parameters:

$$\boldsymbol{\theta}_{new} \longleftarrow \boldsymbol{\theta}_{old} + \alpha \delta \boldsymbol{\theta} \tag{47}$$

- After each update: $\mathcal{J}(\boldsymbol{\theta}_{new}) < \mathcal{J}(\boldsymbol{\theta}_{old})$
- The algorithms differ in the number of iterations required, the computational cost, the convergence guarantees, the robustness with noisy cost functions and their memory usage

Numerical Optimization Algorithms

- **Gradient-based methods:**
  - Gradient descent (with constant, variable step size α)
  - (L-)BFGS (Broyden-Fletcher-Goldfarb-Shanno)
  - Conjugate gradient descent
- **Non-gradient based methods:**
  - Genetic algorithms
  - Non-Linear simplex
  - Nelder-Mead

Numerical techniques may not find the global optimum!

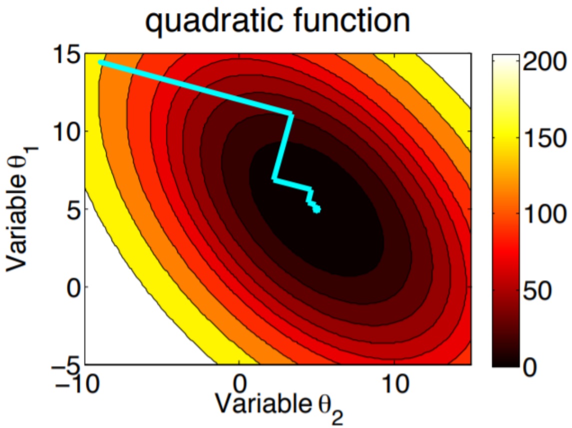
Gradient Descent

- Most basic algorithm (and most commonly used)
- Go into the direction of the **steepest descent**
- The gradient points in the direction of the maximum ($\rightarrow$ subtract gradient)

$$\boldsymbol{\theta}^{(new)} \longleftarrow \boldsymbol{\theta}^{(old)} - \alpha \nabla_{\boldsymbol{\theta}} \mathcal{J}(\boldsymbol{\theta}^{(old)}) \quad (48)$$

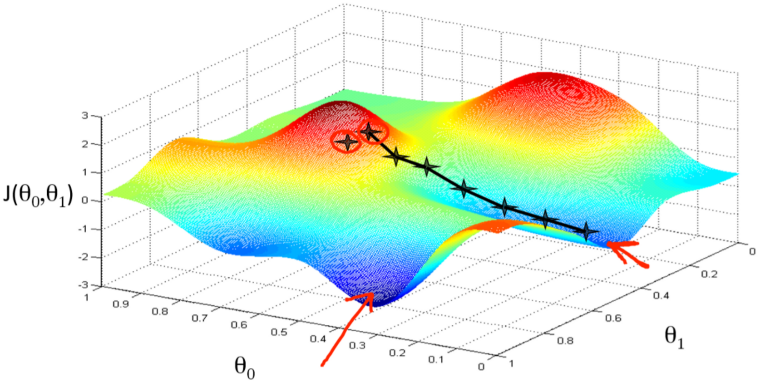
- The parameter updates tend to 'zig-zag' down the valley (see next slide)
- Gradient descent is a 1st-order method

Gradient Descent (Ctd.)



Initialization

Initialization also matters...



Newton's Method

- We want to solve: ($\boldsymbol{H}$ is the **Hessian**, $\boldsymbol{g}$ the **Jacobian**)

$$\delta\boldsymbol{\theta} = \arg\min_{\delta\boldsymbol{\theta}} \Big[\underbrace{c + \boldsymbol{g}^\intercal \delta\boldsymbol{\theta} + \frac{1}{2}\delta\boldsymbol{\theta}^\intercal \boldsymbol{H} \delta\boldsymbol{\theta}}_{\text{Taylor series expansion}} \Big] \tag{49}$$

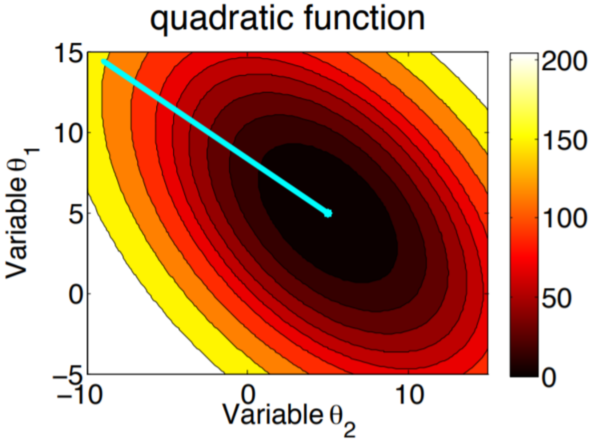
- We have to differentiate and set to zero:

$$\nabla_{\delta\boldsymbol{\theta}} \left[c + \boldsymbol{g}^\intercal \delta\boldsymbol{\theta} + \frac{1}{2}\delta\boldsymbol{\theta}^\intercal \boldsymbol{H} \delta\boldsymbol{\theta} \right] = \boldsymbol{g} + \boldsymbol{H}\delta\boldsymbol{\theta} \overset{!}{=} \boldsymbol{0} \tag{50}$$

- Which yields the solution:

$$\delta\boldsymbol{\theta} = -\boldsymbol{H}^{-1}\boldsymbol{g} \tag{51}$$

Newton's Method (Ctd.)



Want to learn more about Optimization?

- Deep Learning book chapters 4.3, 4.4 and 8
  (Link chapters 4.3, 4.4, Link chapter 8) are highly recommended
- Boyd & Vandenberghe, Convex Optimization (Link)
- Stanford convex optimization course (Link)
- MOOC on constrained optimization (Link)

Section:
Wrap-Up

DHBW
Duale Hochschule
Baden-Württemberg

Summary

-

Important

Self-Test Questions

1

What's next...?

Unit I	Machine Learning Introduction
Unit II	Mathematical Foundations
Unit III	**Bayesian Decision Theory**
Unit IV	Probability Density Estimation
Unit V	Regression
Unit VI	Classification I
Unit VII	Evaluation
Unit VIII	Classification II
Unit IX	Clustering
Unit X	Dimensionality Reduction

Recommended Literature and further Reading

Thank you very much for the attention!

Topic: *** Applied Machine Learning Fundamentals *** Mathematical Foundations
Date: November 18, 2019

Contact:
Daniel Wehner
SAP SE
daniel.wehner@sap.com

Do you have any questions?