Statistical Machine Learning

Summary
Statistical Machine Learning
May 15, 2020

TECHNISCHE
UNIVERSITÄT
DARMSTADT

Contents

List of Figures

List of Tables

List of Algorithms

1. Introduction

Most of the content in this summary, the ideas, the underlying structure and the image ideas are taken from the lecture "Statistical Machine Learning" by Prof. Jan Peters. It is really just a *summary* of the contents of the lecture.

1.1. Examples

1.2. Classification vs. Regression

1.3. Paradigm

1.4. Key Challenges

1.4.1. Generalization

2. Fundamentals: Linear Algebra

2.1. Vectors

A *vector* is an ordered list of numbers that can be interpreted as an *arrow* in multidimensional spaces:

$$\mathbf{v} \in \mathbb{R}^n \quad \rightarrow \quad v = \begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix} \tag{2.1}$$

Scalar Multiplication Multiplying a vector by a scalar is defined as multiplying each component by that scalar (let $\mathbf{v} \in \mathbb{R}^n$, $\lambda \in \mathbb{R}$):

$$\lambda \mathbf{v} = \lambda \begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix} = \begin{bmatrix} \lambda v_1 \\ \vdots \\ \lambda v_n \end{bmatrix} \tag{2.2}$$

Per component:

$$(\lambda v)_i = \lambda v_i \tag{2.3}$$

> Scalar multiplication is a linear operation.

Addition Adding two vectors is defined by adding the component of both vectors (thus, both vectors must have the same size; let $v, w \in \mathbb{R}^n$):

$$\mathbf{v} + \mathbf{w} = \begin{bmatrix} v_1 + w_1 \\ \vdots \\ v_n + w_n \end{bmatrix} \tag{2.4}$$

Per component:

$$(v + w)_i = v_i + w_i \tag{2.5}$$

> Vector addition is both associative and commutative.

Vector Transpose The *transposed* version $\mathbf{v}^T$ of a vector $\mathbf{v} \in \mathbb{R}^n$ is a vector that was flipped around its main axis (thus, the new vector is a row/column vector if the initial vector as a column/row vector):

$$\begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix}^T = \begin{bmatrix} v_1 & \cdots & v_n \end{bmatrix} \tag{2.6}$$

$$\begin{bmatrix} v_1 & \cdots & v_n \end{bmatrix}^T = \begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix} \tag{2.7}$$

> Transposing a vector twice returns the initial vector
>
> $$\mathbf{v} = \left(\mathbf{v}^T\right)^T$$

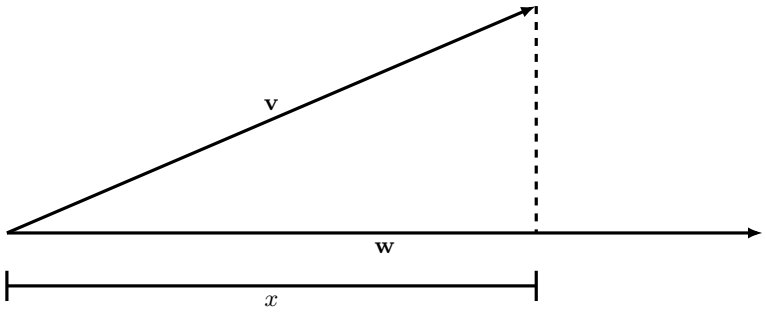

Figure 2.1.: Illustration of Vector Projection

Linear Combination A *linear combination* of multiple vectors $\mathbf{v}_1, \cdots, \mathbf{v}_n \in \mathbb{R}^m$ is the addition of the scaled versions of them (scaled by scalars $\lambda_1, \cdots, \lambda_n \in \mathbb{R}$):

$$u = \lambda_1 v_1 + \cdots + \lambda_n v_n \tag{2.8}$$

If in a group of vectors $\mathbf{v}_1, \cdots, \mathbf{v}_n \in \mathbb{R}^m$, no vector can be represented as a linear combination of the others, they are called *linearly independent*.

Inner and Outer Product and Length The *inner product* of two vectors $\mathbf{v}, \mathbf{w} \in \mathbb{R}^n$ is the sum of the product of the components and is denoted by a single dot $(\mathbf{v} \cdot \mathbf{w})$:

$$\mathbf{v} \cdot \mathbf{w} = \mathbf{v}^T \mathbf{w} = (v_1 w_1) + \cdots + (v_n w_n) \tag{2.9}$$

Therefore, the inner product gives a scalar value. In Cartesian coordinates, this is also called the *scalar product*.

The length $\|\mathbf{v}\|$ of a vector $\mathbf{v} \in \mathbb{R}^n$ is given as the euclidean norm:

$$\|\mathbf{v}\| = (\mathbf{v} \cdot \mathbf{v})^{\frac{1}{2}} = \sqrt{v_1^2 + \cdots + v_n^2}$$

The *outer product* of two vectors $\mathbf{v}, \mathbf{w} \in \mathbb{R}^n$ is defined analogue to the inner product, but with the transpose switched and is denoted by a cross in a circle $(\mathbf{v} \otimes \mathbf{w})$:

$$\mathbf{v} \otimes \mathbf{w} = \mathbf{v} \otimes \mathbf{w}^T \begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix} \otimes \begin{bmatrix} w_1 & \cdots & w_n \end{bmatrix} = \begin{bmatrix} v_1 w_1 & \cdots & v_1 w_n \\ \vdots & \ddots & \vdots \\ v_n w_1 & \cdots & v_n w_n \end{bmatrix} \tag{2.10}$$

This is equivalent to matrix multiplication with two vectors and thus produces a matrix $\mathbf{v} \otimes \mathbf{w} \in \mathbb{R}^{n \times n}$.

Angles between Vectors The *angle* θ between two vectors $\mathbf{v}, \mathbf{w} \in \mathbb{R}$ is given by:

$$\cos\theta = \frac{\mathbf{v} \cdot \mathbf{w}}{\|\mathbf{v}\| \, \|\mathbf{w}\|} \quad \Longleftrightarrow \quad \theta = \arccos \frac{\mathbf{v} \cdot \mathbf{w}}{\|\mathbf{v}\| \, \|\mathbf{w}\|} \tag{2.11}$$

Projections of Vectors A *projection* of a vector $\mathbf{v} \in \mathbb{R}^n$ onto a vector $\mathbf{w} \in \mathbb{R}^n$ results in a scalar value $x \in \mathbb{R}$ which equals the length of the adjacent w.r.t. to the angle between both vectors given that $\mathbf{v}$ is the hypotenuse and the third line is orthogonal to $\mathbf{w}$, see figure 2.1 for an illustration. Then this length is given as:

$$x = \|\mathbf{v}\| \cos\theta = \frac{\mathbf{v} \cdot \mathbf{w}}{\|\mathbf{w}\|} \tag{2.12}$$

2.2. Matrices

A *matrix* is an ordered group of numbers that are ordered in two dimensions:

$$A \in \mathbb{R}^{n \times m} \quad \rightarrow \quad A = \begin{bmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nm} \end{bmatrix} \tag{2.13}$$

Scalar Multiplication Multiplying a matrix by a scalar is defined as multiplying each component by that scalar (let $A \in \mathbb{R}^{n \times m}$, $\lambda \in \mathbb{R}$):

$$\lambda A = \lambda \begin{bmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nm} \end{bmatrix} = \begin{bmatrix} \lambda a_{11} & \cdots & \lambda a_{1m} \\ \vdots & \ddots & \vdots \\ \lambda a_{n1} & \cdots & \lambda a_{nm} \end{bmatrix} \tag{2.14}$$

Per component:

$$(\lambda A)_{ij} = \lambda a_{ij} \tag{2.15}$$

Scalar multiplication is a linear operation.

Addition Adding two matrices is defined by adding the component of both matrices (thus, both matrices must have the same size; let $A, B \in \mathbb{R}^{n \times m}$):

$$A + B = \begin{bmatrix} a_{11} + b_{11} & \cdots & a_{1m} + b_{1m} \\ \vdots & \ddots & \vdots \\ a_{n1} + b_{n1} & \cdots & a_{nm} + b_{nm} \end{bmatrix} \tag{2.16}$$

Per component:

$$(A + B)_{ij} = a_{ij} + b_{ij} \tag{2.17}$$

Matrix addition is both associative and commutative.

Transpose The *transposed* version A^T of a matrix $A \in \mathbb{R}^{n \times m}$ is a matrix $A^T \in \mathbb{R}^{m \times n}$ that was flipped around its main axis:

$$\begin{bmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nm} \end{bmatrix}^T = \begin{bmatrix} a_{11} & \cdots & a_{n1} \\ \vdots & \ddots & \vdots \\ a_{1m} & \cdots & a_{nm} \end{bmatrix} \tag{2.18}$$

Transposing a matrix twice returns the initial matrix

$$A = \left(A^T\right)^T$$

Matrix Multiplication The *matrix multiplication* of two matrices is only possible of the number of columns of the first matrix equals the number of rows of the second matrix (i.e. $A \in \mathbb{R}^{n \times m}$, $B \in \mathbb{R}^{m \times o}$). The resulting matrix then has the dimensions $AB \in \mathbb{R}^{n \times o}$. Matrix multiplication is defined as follows:

$$AB = \begin{bmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nm} \end{bmatrix} \cdot \begin{bmatrix} b_{11} & \cdots & b_{1o} \\ \vdots & \ddots & \vdots \\ b_{m1} & \cdots & b_{mo} \end{bmatrix} = \begin{bmatrix} a_{11}b_{11} + \cdots + a_{1m}b_{m1} & \cdots & a_{11}b_{1o} + \cdots + a_{1m}b_{mo} \\ \vdots & \ddots & \vdots \\ a_{n1}b_{11} + \cdots + a_{nm}b_{m1} & \cdots & a_{n1}b_{1o} + \cdots + a_{nm}b_{mo} \end{bmatrix} \tag{2.19}$$

Per component:

$$(AB)_{ij} = \sum_{k=1}^{n} A_{ik} B_{kj} \tag{2.20}$$

Matrix multiplication is only associative and distributive w.r.t. matrix addition.

Inverse Let $I_n \in \mathbb{R}^{n \times n}$ be the identity matrix with all ones on the main diagonal and the rest zeros. If the dimension is clear, the n can be left out.

Using this definition, the *inverse* of a matrix $A \in \mathbb{R}^{n \times n}$ is defined as matrix $A^{-1} \in \mathbb{R}^{n \times n}$ that holds the following equation:

$$AA^{-1} = A^{-1}A = I_n \tag{2.21}$$

If such a matrix exists, the matrix A is called *regular* or *nonsingular*.

Pseudoinverse If a matrix $A \in \mathbb{R}^{n \times m}$ is not squared (i.e. $n \neq m$), there exists no inverse matrix. Instead, *Pseudoinverse Matrices* can be used to "invert" such a matrix. Left and right pseudoinverse are mutually exclusive, meaning that the one can only exist if the other does not (whilst neither have to exist).

The *left pseudoinverse* does only exist of the matrix has full column rank and is defined as:

$$A^{\#} = \left(A^T A\right)^{-1} A^T \quad \Longrightarrow \quad A^{\#} A = I_m \tag{2.22}$$

The *right pseudoinverse* does only exist if the matrix has full row rank and is defined as:

$$A^{\#} = J^T \left(J J^T\right)^{-1} \quad \Longrightarrow \quad A A^{\#} = I_n \tag{2.23}$$

Properties

Symmetricity A squared matrix $A \in \mathbb{R}^{n \times n}$ is *symmetric* iff $A^T = A$. This implies that:

- The inverse matrix A^{-1} is also symmetric.
- A can be decomposed into $A = QDQ^T$, where D is a diagonal matrix with all eigenvalues of A and Q is a matrix with all columns as the eigenvectors of A.

Definite Quadratic Form Let $A \in \mathbb{R}^{n \times n}$ be a squared matrix and let $\sigma(A)$ be its spectral. Then the matrix A is

$$\begin{cases} \text{positive definite} & \forall \lambda \in \sigma : \lambda > 0 & \Longleftrightarrow & \mathbf{x}^T A \mathbf{x} > 0 \\ \text{negative definite} & \forall \lambda \in \sigma : \lambda < 0 & \Longleftrightarrow & \mathbf{x}^T A \mathbf{x} < 0 \\ \text{positive semi-definite} & \forall \lambda \in \sigma : \lambda \geq 0 & \Longleftrightarrow & \mathbf{x}^T A \mathbf{x} \geq 0 \\ \text{positive semi-definite} & \forall \lambda \in \sigma : \lambda \leq 0 & \Longleftrightarrow & \mathbf{x}^T A \mathbf{x} \leq 0 \\ \text{indefinite} & \text{else} \end{cases} \tag{2.24}$$

for all vectors $\mathbf{x} \in \mathbb{R}^n$.

Regularity/Nonsingularity All of the following are equivalent w.r.t. a matrix $A \in \mathbb{R}^{n \times n}$:

- The matrix is regular.
- The matrix is nonsingular (or not singular).
- There exists a matrix $A^{-1} \in \mathbb{R}^{n \times n}$ with $AA^{-1} = A^{-1}A = I_n$.
- The determinant of the matrix is nonzero: $\det A \neq 0$.
- The matrix has full row rank.
- The matrix has full column rank.

2.3. Operations and Linear Transformations

Change of Basis

Linear Transformations

2.4. Eigenvalues and -vectors

Basis

Linear Transformations

2.5. Wrap-Up

- Vectors and matrices
- Operations on vectors and matrices
- Eigenvectors and -values
- Linear transformations

3. Fundamentals: Statistics

For a much more detailed introduction into the basic concepts (e.g. random variables), please lookup the summary of "Math 3: Stochastic and Statistics" (see https://www.dmken.com/cs), but notice that it is in German.

3.1. Random Variables

A *random variable* is a number that is determined by chance, draw according to a probability distribution.

3.2. Distributions

A probability distribution describes the probability that a random variable will equal a certain value (or lie in a certain range).

3.2.1. Uniform Distribution

All data/all values are equally likely within a bounded region R with size R.

$$p(x) = \frac{1}{R} \tag{3.1}$$

The distribution is plotted in figure 3.1.

3.2.2. Discrete Distributions

The random variables take *discrete values* (can be infinite, but countably infinity) and their probabilities sum up to 1:

$$\sum_i p(x_i) = 1 \tag{3.2}$$

A discrete distribution is described by a *probability mass function* which is a normalized histogram.

Bernoulli Distribution

A *Bernoulli random variable* only takes on two values, e.g. 0 or 1.

Parameters

μ The probability that the variable equals 1.

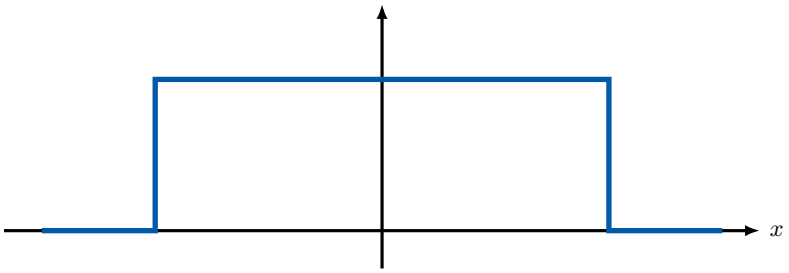

Figure 3.1.: Uniform Distribution

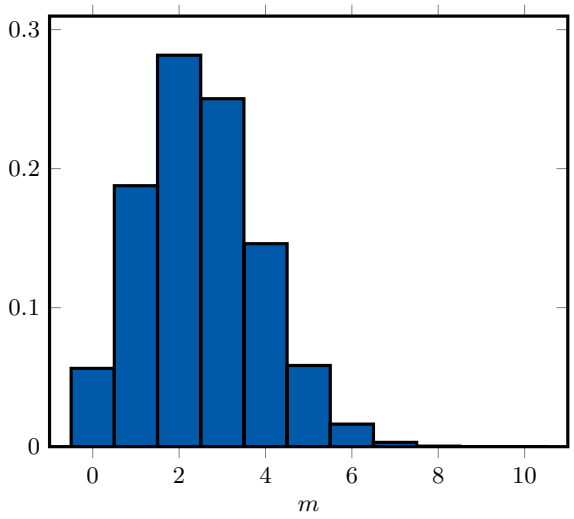

Figure 3.2.: Binomial Distribution $\text{Bin}(m \mid 10, 0.25)$

Properties

$$x \in \{0, 1\} \tag{3.3}$$
$$p(x = 1 \mid \mu) = \mu \tag{3.4}$$
$$\text{Bern}(x \mid \mu) = y^x (1 - \mu)^{1-x} \tag{3.5}$$
$$\text{E}(x) = \mu \tag{3.6}$$
$$\text{Var}(x) = \mu(1 - \mu) \tag{3.7}$$

Binomial Distribution

Binomial variables are a sequence of N Bernoulli variables.

Parameters

μ The probability that one variable equals 1.

N The number of trials/samples.

Properties

$$\text{Bin}(m \mid N, \mu) = \binom{N}{m} \mu^m (1 - \mu)^{N-m} \tag{3.8}$$
$$\text{E}(m) = N\mu \tag{3.9}$$
$$\text{Var}(m) = N\mu(1 - \mu) \tag{3.10}$$

See figure 3.2 for a visualization of $\text{Bin}(m \mid 10, 0.25)$.

Multinoulli Distribution

Multinoulli variables (also called *categorical variables*) are a generalization of Bernoulli variables where each variable can have multiple (namely K) outputs. The random variables is a vector with one-hot-encoding.

Parameters

$\boldsymbol{\mu}$ The entry μ_i defines the probability that the entry x_i equals 1.
All entries must be $\mu_i \geq 0$ and $\sum_{k=1}^{K} \mu_k = 1$.

K The number of classes/outcomes.

Properties

$$\mathbf{x} = [0, 0, 1, 0, 0, 0]^T \tag{3.11}$$

$$p(x_i \mid \boldsymbol{\mu}) = \mu_i \tag{3.12}$$

$$p(\mathbf{x} \mid \boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_k^{x_k} \tag{3.13}$$

$$\mathrm{E}(\mathbf{x} \mid \mu) = \sum_{\mathbf{x}} p(\mathbf{x} \mid \boldsymbol{\mu})\mathbf{x} = \boldsymbol{\mu}^T \tag{3.14}$$

Multinomial Distribution

Multinomial variables are a sequence of N Multinoulli variables.

Parameters

$\boldsymbol{\mu}$ The entry μ_i defines the probability that, for one variable, the entry x_i equals 1. All entries must be $\mu_i \geq 0$ and $\sum_{k=1}^{K} \mu_k = 1$.

K The number of classes/outcomes.

N The number of trials/samples.

Properties

$$\mathrm{Mult}(m_1, m_2, \cdots, m_K \mid \boldsymbol{\mu}, N) = \binom{N}{m_1, m_2, \cdots, m_K} \prod_{k=1}^{K} \mu_k^{m_k} \tag{3.15}$$

$$\mathrm{E}(m_k) = N\mu_k \tag{3.16}$$

$$\mathrm{Var}(m_k) = N\mu_k(1 - \mu_k) \tag{3.17}$$

$$\mathrm{Cov}(m_j, m_k) = -N\mu_j\mu_k \tag{3.18}$$

Poisson Distribution

A *Poisson distribution* is a binomial distribution where the number of trials goes to infinity $N \to \infty$ and the success of each trial goes to zero $\mu \to 0$, s.t. $N\mu = \lambda$ is constant.

Parameters

λ Defines the expectation value and the variance at once.

Properties

$$p(m \mid \lambda) = \frac{\lambda^m}{m!} e^{-\lambda} \tag{3.19}$$

$$\mathrm{E}(m) = \lambda \tag{3.20}$$

$$\mathrm{Var}(m) = \lambda \tag{3.21}$$

See figure 3.3 for a visualization of $p(m \mid 5)$.

3.2.3. Continuous Distributions

The random variables take *discrete values* (infinite, can be uncountable) and their probability density function integrates to 1:

$$\int_{-\infty}^{+\infty} p(x)\,\mathrm{d}x = 1 \tag{3.22}$$

A continuous distribution is described by a *probability density function* $p(x)$.

The probability that a random variable x falls into the interval (a, b) is

$$P(a < x < b) = \int_{a}^{b} p(x)\,\mathrm{d}x \tag{3.23}$$

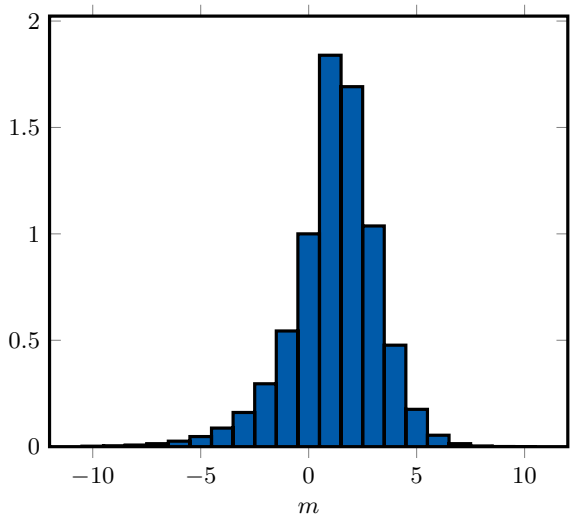

Figure 3.3.: Poisson Distribution $p(m \mid 5)$

Gaussian Distribution

Parameters

μ The expectation value.

σ^2 The variance.

Properties

$$p(x) = \mathcal{N}(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2}(x-\mu)^2 \right\} \tag{3.24}$$

$$\mathrm{E}(x) = \mu \tag{3.25}$$

$$\mathrm{Var}(x) = \sigma^2 \tag{3.26}$$

A Gaussian distribution has more really useful properties:

- A Gaussian has soft tails, i.e. they fade away smoothly.
- Gaussians are often good models for data and provide analytical solutions.

See figure 3.4 for a visualization of $\mathcal{N}(x \mid 0, 1)$ (the standard Gaussian distribution).

3.2.4. Multivariate Gaussian Distribution

Gaussians can be applied to D-dimensional data $x_1, x_2, \cdots$ using multivariate Gaussian distributions.

Parameters

$\boldsymbol{\mu}$ A vector $\mu \in \mathbb{R}^D$ of the expectation values for each dimension.

Σ The *covariance matrix* containing the variance for each dimension on its main axis and the covariances on the other spaces. It is symmetric and defined as:

$$\Sigma = \begin{bmatrix} \mathrm{Var}(x_1) & \mathrm{Cov}(x_1, x_2) & \cdots & \mathrm{Cov}(x_1, x_D) \\ \mathrm{Cov}(x_2, x_1) & \mathrm{Var}(x_2) & \cdots & \mathrm{Cov}(x_2, x_D) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(x_D, x_1) & \mathrm{Cov}(x_D, x_2) & \cdots & \mathrm{Var}(x_D) \end{bmatrix} \tag{3.27}$$

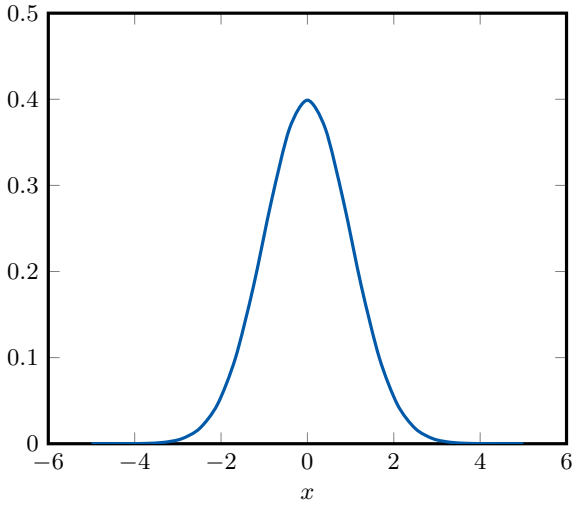

Figure 3.4.: Standard Gaussian Distribution $\mathcal{N}(x \mid 0, 1)$

Properties

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}, \Sigma) = \frac{1}{\sqrt{2\pi}^D \sqrt{\det \Sigma}} \exp\left\{ -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right\} \tag{3.28}$$

$$\mathrm{E}(\mathbf{x}) = \boldsymbol{\mu} \tag{3.29}$$

$$\mathrm{Var}(x_i) = \Sigma_{ii} \tag{3.30}$$

Geometry

Moments

3.2.5. Partitioned Gaussian Distributions

3.3. Central Limit Theorem

The distribution of the sum of N i.i.d. random variables becomes increasingly Gaussian as N increases. That is, with $N \to \infty$, it converges towards a Gaussian.

3.4. Probability Rules

Joint Distribution

$$p(x, y) \tag{3.31}$$

Marginal Distribution

$$p(y) = \int p(x, y)\, \mathrm{d}x \tag{3.32}$$

Conditional Distribution

$$p(y \mid x) = \frac{p(x, y)}{p(x)} \tag{3.33}$$

Probabilistic/Stochastic Independence

$$p(x, y) = p(x)p(y) \tag{3.34}$$

Chain Rule of Probabilities

$$p(x_1, \cdots, x_n) = p(x_1 \mid x_2, \cdots, x_n)p(x_2, \cdots, x_n) \tag{3.35}$$
$$= p(x_1 \mid x_2, \cdots, x_n)p(x_2 \mid x_3, \cdots, x_n) \cdots p(x_{n-1} \mid x_n)p(x_n) \tag{3.36}$$

Bayes Rule

$$p(y \mid x) = \frac{p(x \mid y)p(y)}{p(x)} \tag{3.37}$$

- Posterior: $p(y \mid x)$
- Likelihood: $p(x \mid y)$
- Prior: $p(y)$
- Normalization Factor: $p(x) = \int p(x, y)\,\mathrm{d}y = \int p(x \mid y)p(y)\,\mathrm{d}y$

3.5. Expectation, Variance and Moments

3.5.1. Expectation

The *expectation* value of a random variable x with a distribution $p(x)$ is defined as:

$$\mathrm{E}_{x\sim p(x)}\left(f(x)\right) = \mathrm{E}_x\left(f(x)\right) = \mathrm{E}\left(f(x)\right) = \begin{cases} \sum_x f(x)\,p(x) & \text{for discrete distributions} \\ \int_x f(x)\,p(x)\,\mathrm{d}x & \text{for continious distributions} \end{cases} \tag{3.38}$$

This gives a similar formula for the *conditional expectation*:

$$\mathrm{E}_{x\sim p(x\mid y)}\left(f(x)\right) = \mathrm{E}_x\left(f(x)\right) = \mathrm{E}\left(f(x)\right) = \begin{cases} \sum_x f(x)\,p(x \mid y) & \text{for discrete distributions} \\ \int_x f(x)\,p(x \mid y)\,\mathrm{d}x & \text{for continious distributions} \end{cases} \tag{3.39}$$

With enough samples, the expectation value can be approximated using the arithmetic mean:

$$\mathrm{E}\left(f(x)\right) \approx \frac{1}{N}\sum_{i=1}^{N} f(x_i) \tag{3.40}$$

Calculation Rules Let x, y be random variables and $\alpha \in \mathbb{R}$.

$$\mathrm{E}(\alpha x) = \alpha\,\mathrm{E}(x) \tag{3.41}$$
$$\mathrm{E}(x + y) = \mathrm{E}(x) + \mathrm{E}(y) \tag{3.42}$$
$$\mathrm{E}(xy) = \mathrm{E}(x)\,\mathrm{E}(y) \tag{3.43}$$

Equation 3.43 only holds if x and y are statistically independent.

3.5.2. Variance and Covariance

The *variance* measures the spread of the variable in relation to its mean:

$$\mathrm{Var}(x) = \mathrm{E}\left(\left(x - \mathrm{E}(x)\right)^2\right) = \mathrm{E}\left(x^2\right) - \left(\mathrm{E}(x)\right)^2 \tag{3.44}$$

The *covariance* measures the correlation between two variables (how much the variables change together):

$$\mathrm{Cov}(x, y) = \mathrm{E}_{x,y}(xy) - \mathrm{E}_x(x)\,\mathrm{E}_y(y) \tag{3.45}$$
$$\mathrm{Cov}(\mathbf{x}, \mathbf{y}) = \mathrm{E}_{\mathbf{x},\mathbf{y}}(\mathbf{x}\mathbf{y}^T) - \mathrm{E}_{\mathbf{x}}(\mathbf{x})\,\mathrm{E}_{\mathbf{y}}(\mathbf{y}^T) \tag{3.46}$$

This gives the following very important rule (with $\boldsymbol{\mu}$ and Σ from the Gaussian):

$$\mathrm{E}(\mathbf{x}\mathbf{x}^T) = \boldsymbol{\mu}\boldsymbol{\mu}^T + \Sigma \tag{3.47}$$

3.5.3. Moments

A *moment* is defined as

$$m_n = E(x^n) \tag{3.48}$$

The *central moment* is defined as

$$cm_n = \mathrm{E}\left((x-\mu)^n\right) \tag{3.49}$$

which leads to another definition of the variance, skewness and kurtosis:

$\boldsymbol{cm}_2$ Variance (measure of spreading)

$\boldsymbol{cm}_3$ Skewness (measure of asymmetry)

$\boldsymbol{cm}_4$ Kurtosis (measure of heavy/light tailed-ness)

3.6. Exponential Family

The *exponential family* is a large class of distributions that are analytically interesting, because taking the log of them simplifies them a lot. All distributions of this family are unimodal with the following general form:

$$p(\mathbf{x} \,|\, \boldsymbol{\eta}) = h(\mathbf{x}) g(\boldsymbol{\eta}) \exp\left\{\boldsymbol{\eta}^T \mathbf{u}(\mathbf{x})\right\} \tag{3.50}$$

where $\boldsymbol{\eta}$ is the natural parameter and

$$g(\boldsymbol{\eta}) \int_{-\infty}^{+\infty} h(\mathbf{x}) \exp\left\{\boldsymbol{\eta}^T \mathbf{u}(\mathbf{x})\right\} = 1 \tag{3.51}$$

holds. g can be interpreted as a normalization to make this property hold true.

3.6.1. Example: Bernoulli Distribution

The Bernoulli distribution is part of the exponential family and decomposes as

$$\begin{aligned} \mathrm{Bern}(x \,|\, \mu) &= \mu^x (1-\mu)^{1-x} && (3.52)\\ &= \exp\left\{x \ln(\mu) + (1-x)\ln(1-\mu)\right\} && (3.53)\\ &= (1-y)\exp\left\{\ln\left(\frac{\mu}{1-\mu}\right) x\right\} && (3.54) \end{aligned}$$

with the logistic sigmoid

$$\sigma(\eta) = \frac{1}{1+\exp(-\eta)} \tag{3.55}$$

and

$$\eta = \ln\left(\frac{\mu}{1-\mu}\right) \tag{3.56}$$

we can write the Bernoulli distribution as

$$p(x \,|\, \mu) = \sigma(-\eta)\exp(\eta x) \tag{3.57}$$

which, in the exponential family form, gives:

$$h(x) = 1 \qquad g(\eta) = \sigma(-\eta) \qquad u(x) = x \tag{3.58}$$

3.6.2. Example: Gaussian Distribution

The Gaussian distribution is part of the exponential family and decomposes as

$$\begin{aligned} \mathcal{N}\left(x \,|\, \mu, \sigma^2\right) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} && (3.59)\\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x - \frac{\mu^2}{2\sigma^2}\right\} && (3.60)\\ &= h(x) g(\boldsymbol{\eta}) \exp\left\{\boldsymbol{\eta}^T \mathbf{u}(x)\right\} && (3.61) \end{aligned}$$

with

$$\eta = \begin{bmatrix} -\frac{1}{2\sigma^2} & \frac{\mu}{\sigma^2} \end{bmatrix}^T \qquad h(x) = 1 \qquad g(\boldsymbol{\eta}) = \sqrt{-\frac{\eta_1}{\pi}} \exp\left\{\frac{\eta_2^2}{4\eta_1}\right\} \qquad \mathbf{u}(x) = \begin{bmatrix} x^2 \\ x \end{bmatrix} \tag{3.62}$$

3.7. Information Theory and Entropy

Information theory is about how to represent information compactly (as few bits as possible) and therefore about compression.
This raises three questions:

- How to measure complexity?
- How to measure the "distance" between probability distributions?
- How to reconstruct data?

3.7.1. Information and Entropy

- Information is hiding in data.
- E.g. in the English alphabet, every letter has a different probability p_i of occurring.
- A lower probability indicates that the data point contains more information.
- The *average information*, called *entropy* can be calculated as

$$H(p) = -\sum_i p_i \log_2(p_i) \tag{3.63}$$

3.7.2. Kullback-Leibler Divergence

The *Kullback-Leibler Divergence* is a similarity measurement between probability distributions, defined by

$$\mathrm{KL}(p \,\|\, q) = -\int p(x) \ln\big(q(x)\big)\,\mathrm{d}x - \left(-\int p(x) \ln\big(p(x)\big)\,\mathrm{d}x\right) \tag{3.64}$$

$$= -\int p(x) \ln\left(\frac{q(x)}{p(x)}\right)\mathrm{d}x \tag{3.65}$$

The KL divergence represents the average additional number of bits required to specify a symbol x, if the underlying probability distribution is the estimated $q(x)$ and not the true one $p(x)$.
Some properties:

- $\mathrm{KL}(p \,\|\, q) \neq \mathrm{KL}(q \,\|\, p)$ not a distance
- $\mathrm{KL}(p \,\|\, q) \geq 0$ non-negative distance
- $\big(\forall x : p(x) = q(x)\big) \implies \mathrm{KL}(p \,\|\, q) = 0$

There exist other metrics than KL, but KL is deeply connected to maximum likelihood estimation.

3.8. Wrap-Up

- Random variables (both continuous and discrete)
- Probability distributions
- Basic rules of probability theory
- Expectation and variance
- Gaussian distribution and its importance
- Information and entropy

4. Fundamentals: Optimization

> "All learning problems are essentially optimization problems on data" (Christopher G. Atkeson, Professor at CMU)

All machine learning problems are optimization problems in the form

$$\min_{\theta} J(\theta, \mathcal{D}) \tag{4.1}$$

$$\text{s.t.} \quad f(\theta, \mathcal{D}) = 0 \tag{4.2}$$

$$g(\theta, \mathcal{D}) \geq 0 \tag{4.3}$$

with parameters θ to enable learning, a data set $\mathcal{D}$ to learn from, a cost function $J(\theta, \mathcal{D})$ to measure the performance and equality and inequality constraints $f(\theta, \mathcal{D}) = 0$, $g(\theta, \mathcal{D}) \geq 0$.

4.1. Convexity

A set $C \subseteq \mathbb{R}^n$ is *convex* iff for all $\mathbf{x}, \mathbf{y} \in C$ and for all $\alpha \in [0, 1]$ the following holds:

$$\alpha\mathbf{x} + (a - \alpha)\mathbf{y} \in C \tag{4.4}$$

Intuition: Every point on a line drawn between two arbitrary points in space lie in the set itself. The set has no "bays".

A function $f : \mathbb{R}^n \to \mathbb{R}$ is *convex* iff for all $\mathbf{x}, \mathbf{y} \in \text{Domain}(f)$ and for all $\alpha \in [0, 1]$ the following holds:

$$f\big(\alpha\mathbf{x} + (1 - \alpha)\mathbf{y}\big) \leq \alpha f(\mathbf{x}) + (1 - \alpha) f(\mathbf{y}) \tag{4.5}$$

Intuition: The drawn line between two arbitrary points on the function do not cross the function (they may touch, so linear functions are also convex).

If f is differentiable, it is convex iff for all $\mathbf{x}, \mathbf{y} \in \text{Domain}(f)$ the following holds:

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla_{\mathbf{x}}(\mathbf{y} - \mathbf{x}) \tag{4.6}$$

If f is twice differentiable, it is convex iff for all $\mathbf{x} \in \text{Domain}(f)$ the following holds:

$$\nabla^2_{\mathbf{x}} f(\mathbf{x}) \preceq 0 \tag{4.7}$$

> **Warning:** Differentiability is not a condition for convexity!

4.2. Cost Functions

An ideal cost function is convex. But most of the time, they are not...

4.2.1. Common Cost Functions

Table 4.1 lists common cost functions for classification, regression, density estimation and clustering.

4.3. Constrained/Unconstrained Optimization

The general form of a constrained optimization problem is

$$\max_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) \tag{4.8}$$

$$\text{s.t.} \quad \mathbf{f}(\boldsymbol{\theta}) = 0 \tag{4.9}$$

$$\mathbf{g}(\boldsymbol{\theta}) \geq 0 \tag{4.10}$$

with a cost function $J(\boldsymbol{\theta})$, some equality constraints $\mathbf{f}(\boldsymbol{\theta})$ and inequality constraints $\mathbf{g}(\boldsymbol{\theta})$.

Problem	Example Cost Functions	Resulting Method
Classification	$\min_\theta \sum_{i=1}^n \ln\left(1+\exp\left(-y_i x_i^T\theta\right)\right)$	Logistic Regression
	$\min_{\theta_1,\theta_2} \sum_{i=1}^n \left(y_i - g\left(\theta_2^T g\left(\theta_1^T x_i\right)\right)\right)^2$	Neural Network Classification
	$\min_\theta \lVert\theta\rVert^2 + C\sum_{i=1}^n \xi_i$ s.t. $\xi_i - \left(1 - y_i x_i^T\theta\right) \geq 0,\ \xi_i \geq 0$	Support Vector Machines
Regression	$\min_\theta \sum_{i=1}^n \left(y_i - \phi(x_i)^T\theta\right)^2$	Linear Regression
	$\min_{\theta_1,\theta_2,\theta_3} \sum_{i=1}^n \left(y_i - \theta_3^T g\left(\theta_2^T g\left(\theta_1^T x_i\right)\right)\right)^2$	Neural Network Regression
Density Estimation	$\min_\theta \sum_{i=1}^n \ln\left(p(x_i \mid \theta)\right)$	General Formulation
Clustering	$\min_{\mu_1,\cdots,\mu_k} \sum_{j=1}^k \sum_{i\in C_j} \lVert x_i - \mu_i\rVert^2$	

Table 4.1.: Common Cost Functions

4.4. Lagrange Multipliers

With a constrained optimization problem in the general form, the *Lagrangian* is defined as

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\epsilon}) = J(\theta) + \boldsymbol{\lambda}^T \mathbf{f}(\boldsymbol{\theta}) + \boldsymbol{\mu}^T\left(\mathbf{g}(\boldsymbol{\theta} + \boldsymbol{\epsilon}^2)\right) \tag{4.11}$$

The coefficients $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ are called *Lagrangian Multipliers*, the variables $\boldsymbol{\epsilon}$ are called *slack variables* and are used to convert the inequality constraints into equality constraints.

To solve the optimization problem, take the derivatives w.r.t. θ, λ and μ and set them to zero:

$$\nabla_{\boldsymbol{\theta}}\mathcal{L} = 0 \qquad \nabla_{\boldsymbol{\lambda}}\mathcal{L} = 0 \qquad \nabla_{\boldsymbol{\mu}}\mathcal{L} = 0 \tag{4.12}$$

If this results in any $\epsilon_i = 0$, the inequality constraint is called *active* and the solution lies on the edge of that constraint. To check whether the result really is a minima/maxima, take the second derivative of the cost function w.r.t. θ, $\nabla^2_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$, and check whether the resulting Hessian is positive or negative definite, yielding that the found solution is a minima or maxima, respectively.

4.4.1. Dual Formulation

Given the so-called *primal problem*

$$\min_{\boldsymbol{\theta}}\ J(\boldsymbol{\theta}) \tag{4.13}$$

$$\text{s.t.} \qquad \mathbf{f}(\boldsymbol{\theta}) = 0 \tag{4.14}$$

$$\mathbf{g}(\boldsymbol{\theta}) \geq 0 \tag{4.15}$$

with the Lagrangian

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\epsilon}) = J(\theta) + \boldsymbol{\lambda}^T \mathbf{f}(\boldsymbol{\theta}) + \boldsymbol{\mu}^T\left(\mathbf{g}(\boldsymbol{\theta} + \boldsymbol{\epsilon}^2)\right) \tag{4.16}$$

the *dual problem* is

$$\max_{\boldsymbol{\lambda},\boldsymbol{\mu}}\ \hat{\mathcal{L}}(\boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\epsilon}) = \min_{\boldsymbol{\theta}}\ \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\epsilon}^2) \tag{4.17}$$

$$\text{s.t.} \quad \boldsymbol{\lambda} \geq \mathbf{0} \tag{4.18}$$

$$\boldsymbol{\mu} \geq \mathbf{0} \tag{4.19}$$

- If $\boldsymbol{\lambda}^*$ is the solution for the dual problem, then $\hat{\mathcal{L}}(\boldsymbol{\lambda}^*)$ is a *lower bound* for the primal problem due to two concepts:
  - *Minimax inequality:* For any function with two arguments $\phi(x, y)$, the maximin is less or equal to the minimax:

$$\max_y \min_x\ \phi(x, y) \leq \min_x \max_y\ \phi(x, y) \tag{4.20}$$

  - *Weak duality:* The primal values are always greater or equal to the dual values:

$$\min_{\boldsymbol{\theta}} \max_{\substack{\boldsymbol{\lambda}\geq\mathbf{0}\\ \boldsymbol{\mu}\geq\mathbf{0}}}\ \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \geq \max_{\substack{\boldsymbol{\lambda}\geq\mathbf{0}\\ \boldsymbol{\mu}\geq\mathbf{0}}} \min_{\boldsymbol{\theta}}\ \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \tag{4.21}$$

- In machine learning, the dual is often far more useful than the primal.
- That is because $\hat{\mathcal{L}}$ is a concave function and easy to optimize, even if J and the constraints may be nonconvex.
- Given some $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$, the dual is an unconstrained problem.

4.4.2. Example

Given the following optimization problem (in the real numbers):

$$\arg\max_{x,y} J(x,y) = x + y \tag{4.22}$$

$$\text{s.t.} \quad x^2 + y^2 - 1 = 0 \tag{4.23}$$

$$2 - x \geq 0 \tag{4.24}$$

the Lagrangian is written as

$$\mathcal{L}(x, y, \lambda, \mu, \epsilon) = x + y + \lambda(x^2 + y^2 - 1) + \mu(2 - x + \epsilon^2) \tag{4.25}$$

Take the derivatives:

$$\nabla_x \mathcal{L} = 1 + 2\lambda x - \mu \tag{4.26}$$

$$\nabla_y \mathcal{L} = 1 + 2\lambda y \tag{4.27}$$

$$\nabla_\lambda \mathcal{L} = x^2 + y^2 - 1 \tag{4.28}$$

$$\nabla_\mu \mathcal{L} = 2 - x + \epsilon^2 \tag{4.29}$$

$$\nabla_\epsilon \mathcal{L} = 2\mu\epsilon \tag{4.30}$$

Settings them to zero gives the insight that either $\mu = 0$ or $\epsilon = 0$ must be true. This must be done per case.

Case 1: $\mu = 0$ This yields the following equation system:

$$0 = 1 + 2\lambda x \tag{4.31}$$

$$0 = 1 + 2\lambda y \tag{4.32}$$

$$0 = x^2 + y^2 - 1 \tag{4.33}$$

$$0 = 2 - x + \epsilon^2 \tag{4.34}$$

with the solution $x = y = \pm\frac{1}{\sqrt{2}}$ and $\epsilon^2 = \frac{1}{\sqrt{2}} - 2$, so the inequality constraint is not active as the solution fulfills the equation $x < 2$.

Case 2: $\epsilon = 0$ This yields the following equation system:

$$0 = 1 + 2\lambda x - \mu \tag{4.35}$$

$$0 = 1 + 2\lambda y \tag{4.36}$$

$$0 = x^2 + y^2 - 1 \tag{4.37}$$

$$0 = 2 - x \tag{4.38}$$

Which yields the following solutions:

$$x_1 = 2 \qquad x_2 = 2 \tag{4.39}$$

$$y_1 = -i\sqrt{3} \qquad y_2 = i\sqrt{3} \tag{4.40}$$

As $\epsilon = 0$, the solution must lie on the edge of the inequality constraint, thus $x = 2$. As only real solutions where wanted, this solution can be discarded.

4.5. Numerical Optimization

For a lot of optimization problems, the solution cannot be computed analytically, so these have to be approximated using numerical optimization.

The performance of numerical methods can be measured with the following questions:

- Does the algorithm converge to the optimal solution?
- How many steps does it take to converge?
- Is the convergence smooth or bumpy?
- Does it work for all types of functions or just a special type (e.g. convex)?