October 7, 2025

Faculty of Computer Science, Heinrich-Heine-University Düsseldorf

Deep Learning

Winter term 25/26 – Exercise Sheet **Math Recap**

This is a short recap of the mathematical concepts used within the Deep Learning lecture. The concepts are given in a simplified form, for more rigorous definitions we refer to the relevant literature.
Perspective of probability theory: Distributions are given, infer properties of realizations.
Perspective of statistics: Realizations (data) are given, infer underlying distributions.

1 Probability Spaces & Random Variables

Let's assume we study some probabilistic experiment like the outcome of throwing a dice. For this, we need the notion of probability spaces.

Definition 1.1 (Probability space)**.** A probability space $(\Omega, \mathbb{P})$ consists out of two components:

1. A set Ω. (Sample space)
   This set contains all possible outcomes of our probabilistic experiment. A subset $A \subset \Omega$ is called an **event**.

2. A map $\mathbb{P} : \mathcal{P}(\Omega) \to [0, 1]$. (Probability distribution)
   For every event $A \subset \Omega$ this assigns a probability $\mathbb{P}(A) \in [0, 1]$ to A. It holds that $\mathbb{P}(\Omega) = 1$ and $\mathbb{P}(\emptyset) = 0$. For countably many disjoint subsets of Ω $A_1, A_2, A_3, \ldots$ it holds that $\mathbb{P}\Big(\bigcup_{i=1}^{\infty} A_i\Big) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$.

Definition 1.2 (Discrete distributions)**.** Let $(\Omega, \mathbb{P})$ be a probability space. If Ω is countable (e.g., if Ω has a one-to-one correspondence to $\mathbb{N}$) we call $\mathbb{P}$ a discrete probability distribution.

For a discrete distribution, we can assign a probability $\mathbb{P}(\{w\})$ to singular elements $w \in \Omega$ and for $A \subset \Omega$ we have $\mathbb{P}(A) = \sum_{w \in A} \mathbb{P}(\{w\})$.

Example 1.1 (Discrete distributions)**.**
(Fair) Coin throw:

1. $\Omega = \{\text{heads}, \text{tails}\}$
2. $\mathbb{P}(\text{heads}) = \mathbb{P}(\text{heads}) = \frac{1}{2}$

(Unfair) Coin throw:

1. $\Omega = \{\text{heads}, \text{tails}\}$
2. $\mathbb{P}(\text{heads}) = \frac{2}{3}$, $\mathbb{P}(\text{heads}) = \frac{1}{3}$

Fair dice throw:

1. $\Omega = \{1, 2, 3, 4, 5, 6\}$
2. $\mathbb{P}(\{i\}) = \frac{1}{6}$ for all $i \in \{1, \dots, 6\}$.
   $\mathbb{P}(A) = \sum_{i \in A} \frac{1}{6} = \frac{\#\text{ Elements in } A}{6}$ for $A \in \mathcal{A}$

Definition 1.3 (Continuous distributions)**.** Let $(\Omega, \mathbb{P})$ be an uncountable probability space (most often $\Omega = \mathbb{R}$ or $\Omega = \mathbb{R}^n$). Then we call $\mathbb{P}$ a continuous probability distribution.

For a continuous distribution, probabilities of singular elements (e.g., numbers in $\mathbb{R}$) are zero. Positive probabilities are assigned to intervals $[a, b]$. This is done via a so called **probability density function**. A probability density function (pdf) $f : \Omega \to \mathbb{R}$ is a map such that $f(x) \geq 0$ for all $x \in \Omega$, f is integrable and $\int_\Omega f(x)\mathrm{d}x = 1$.
Continuous distribution $\mathbb{P}$ is defined with the help of pdf as

$$\mathbb{P}([a, b]) = \int_a^b f(x)\mathrm{d}x, \qquad a < b \in \Omega.$$

Often density functions have parameters θ to specify the exact distribution. In that case, we write $f(x|\theta)$. (In the literature sometimes also $f_\theta(x)$).

Example 1.2 (Continuous distributions)**.** Uniform distribution on the interval $[a, b]$:

1. $\Omega = [a, b]$
2. $f(x) = \dfrac{1}{b-a} \qquad \Rightarrow \qquad \mathbb{P}([x, y]) = \displaystyle\int_x^y \frac{1}{b-a}\mathrm{d}x = \frac{x-y}{b-a}$

Gaussian distribution on $\mathbb{R}$ with parameters μ, σ^2

1. $\Omega = \mathbb{R}$
2. $f(x) = \dfrac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad \Rightarrow \qquad \mathbb{P}([x, y]) = \dfrac{1}{\sqrt{2\pi\sigma^2}} \displaystyle\int_x^y e^{-\frac{(x-\mu)^2}{2\sigma^2}} \mathrm{d}x$

Definition 1.4 (Random variable)**.** A random variable is a function $X : \Omega \to \mathbb{R}$. Note, that it can be a mapping to any measurable set, not necessarily $\mathbb{R}$. For $A \subset \mathbb{R}$ we then call

$$\mathbb{P}(X = A) = \mathbb{P}(\{w \in \Omega : X(w) \in A\})$$

the probability that X is A. We often write $\mathbb{P}_X(A)$ for $\mathbb{P}(X = A)$ and call $\mathbb{P}_X$ the probability distribution of X.
It is important to notice that very often the value that random variable gets when some probabilistic event happened is called itself a random variable.

Example 1.3 (Random variables)**.** Discrete random variables:

- Outcome of fair dice throw:

1. $\Omega = \{1, 2, 3, 4, 5, 6\}$
2. $X(i) = i$ for all $i \in \Omega$
   $\Rightarrow \mathbb{P}(X = 1) = \mathbb{P}(\{w \in \Omega : X(w) = w = 1\}) = \mathbb{P}(\{1\}) = \frac{1}{6}$
   $\Rightarrow \mathbb{P}(X \text{ is even}) = \mathbb{P}(\{w \in \Omega : X(w) = w \in \{2, 4, 6\}\}) = \frac{3}{6} = \frac{1}{2}$

- Sum of numbers in double dice throw:

1. $\Omega = \{(1, 1), (1, 2), \ldots, (6, 5), (6, 6)\}$
2. $X((i, j)) = i + j$ for all $(i, j) \in \Omega$
   $\Rightarrow \mathbb{P}(X = 4) = \mathbb{P}(\{(1, 3), (2, 2), (3, 1)\}) = \frac{3}{36} = \frac{1}{12}$

Continuous random variables:

- Gaussian random variable with parameters μ, σ^2

1. $\Omega = \mathbb{R}$
2. $X(i) = i$ for all $i \in \Omega$
   $\mathbb{P}(X \in [x, y]) = \dfrac{1}{\sqrt{2\pi\sigma^2}} \displaystyle\int_x^y e^{-\frac{(x-\mu)^2}{2\sigma^2}} \mathrm{d}x$

Definition 1.5 (Joint distributions)**.** Let X, Y be two random variables on some probability space $(\Omega, \mathbb{P})$. For $A, B \in \Omega$ we denote by $\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(\{w \in \Omega : X(w) \in A \cap Y(w) \in B\})$ the **joint probability distribution** of X and Y.

We call X and Y **independent**, if for all $A, B \in \Omega$

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B).$$

If X, Y are continuous, that means if X, Y are independent, their joint density function $f_{X,Y}$ is the product of their individual densities f_X, f_Y.
If the two variables are not independent, then probability of one given another can be calculated as following

$$P(X|Y) = \frac{P(X, Y)}{P(Y)}$$

2 Expectation & Variance

Definition 2.1 (Expected value)**.**
Let X be a discrete real-valued random variable that takes values in $\{x_1, x_2, \ldots\}$. If

$$\mathbb{E}[X] = \sum_{i=1}^{\infty} x_i \mathbb{P}(X = x_i)$$

is finite, we call $\mathbb{E}[X]$ the expected value or expectation of X. The expected value can be seen as the weighted average, with probabilities being the weights.

Let X be a continuous real-valued random variable with probability density function f_X. If

$$\mathbb{E}[X] = \int_{\mathbb{R}} x f_X(x) \mathrm{d}x$$

is finite, we call $\mathbb{E}[X]$ the expected value of X.

The expectation is linear. That means, for two random variables X, Y and some constants $a, b \in \mathbb{R}$ we have that

$$\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y].$$

Example 2.1 (Expected Value)**.**
(Fair) Dice Throw:

$$\mathbb{E}[X] = \sum_{i=1}^{6} x_i \cdot P(X = x_i) = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = 3.5$$

Definition 2.2 (Variance)**.** The variance of a random variable X is the expected value of the squared deviation from the mean of X:

$$\mathbb{V}ar(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right].$$

2.1 Central Limit Theorem

Limit theorems describe how empirical averages of independent random variables $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ are distributed when n approaches infinity. Informally, the central limit theorem (CLT) establishes that, in some situations, when independent random variables are added, their properly normalized sum tends toward a normal distribution (a "bell curve") even if the original variables themselves are not normally distributed.

Theorem 2.2 (Central Limit Theorem)**.** Suppose $\{X_1, \ldots, X_n\}$ is a sequence of independent random variables defined on the same probability space with $\mathbb{E}[X_i] = \mu$ and $\mathrm{Var}[X_i] = \sigma^2 < \infty$. Then as n approaches infinity, the random variables $\sqrt{n}(\bar{X}_n - \mu)$ converge in distribution to a normal distribution $\mathcal{N}(0, \sigma^2)$

$$\sqrt{n}\left(\bar{X}_n - \mu\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2\right).$$

3 Maximum likelihood method

Let $\{x_1, \ldots, x_n\}$ be values sampled independently from a random variable X on some probability space $(\Omega, \mathbb{P}_\theta)$ where the distribution has unknown parameters $\theta \in \mathbb{R}^d$. If X is a continuous random variable, we define the **likelihood** of a $\theta \in \mathbb{R}^d$ given $\{x_1, \ldots, x_n\}$ as

$$\mathcal{L}(\theta) := f(x_1, \ldots, x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta).$$

The likelihood $\mathcal{L}(\theta)$ says how likely are the observations $x_1, \ldots x_n$ under the assumption about the underlying density $f(x|\theta)$ to have parameters θ. Since we do not know the actual value of $\theta \in \mathbb{R}^d$ we have to estimate it. We do that by the so called **maximum likelihood estimator**

$$\hat{\theta} = \text{argmax}_\theta \mathcal{L}(\theta).$$

So $\hat{\theta}$ is the parameter under which our observations $x_1, \ldots x_n$ are most likely.

The maximum likelihood method is simply the estimation of θ by calculating the maximum of $\mathcal{L}(\theta)$. This is often done by maximizing the so called **log-likelihood function**

$$\ell(\theta) := \ln \mathcal{L}(\theta)$$

(turns products into sums and is therefore easier to derive and optimize for). As the logarithm is strictly monotonically decreasing, the position of the maximum of $L(\theta)$ and $l(\theta)$ are equivalent.
Often based on observations $x_1, \ldots, x_n$ from an unknown generative process $X \sim \mathbb{P}_{\text{data}}$, we want to estimate the parameters of our model, a parametric density function $f(x|\theta)$, which describes how a random variable X is distributed in $\mathcal{X}$. Intuitively, we look for a model density that well resembles the target distribution $\mathbb{P}_{\text{data}}$. Now, the idea is to find θ by maximizing the likelihood-function.

4 (Multivariate) Taylor approximation

A Taylor series (which constitutes Taylor approximation) is a series expansion of a function about a point. A one-dimensional Taylor series is an expansion of a real function $f(x)$ about a point $x = a$ is given by

$$f(x) = f(a) + \frac{f'(a)}{1!}(x-a)^1 + \frac{f''(a)}{2!}(x-a)^2 + \ldots$$

This is a nice polynomial approximation of a function which is widely used when the function is complicated. It can be also applied to multivariate function. Note, that in this case first derivative becoming Jacobian $[\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_n}]$ and second derivative - Hessian $(\text{Hessian} f)_{ij} \equiv \frac{\partial^2 f}{\partial x_i \partial x_j}$.

4.1 Eigenvalues and eigenvectors

An eigenvector or characteristic vector of a linear transformation is a nonzero vector that changes at most by a constant factor when that linear transformation is applied to it. The corresponding eigenvalue, often represented by λ, is the multiplying factor. So, for matrix A if $Au = \lambda u$, then u is eigenvector with eigenvalue λ. Eigenvalues of the Hessian of $f(x)$ are characteristic of the curvature of the surface described by the $f(x)$.