Why Data Analysis

- Machine Learning does not work like any data in, good prediction out.
- Although we wish...
- It works with very clean datasets.
- However, real datasets require data analysis to get them clean.
- For example all the datasets used for large language models. A lot of processing and selection is required to arrive at these datasets.
- Truth in machine learning is, "garbage in, garbage out".

Why Data Analysis

- When we do Machine Learning, we typically encounter data points where our model does not work to our satisfaction.
- We need to be able to identify these cases and analyse what characterises them.
- Data analysis skills, i.e. statistical skills are mandtory throughout the whole process of Machine Learning.

Content and Goals

Ostbayerische Technische Hochschule Amberg-Weiden

- Descriptive statistics of data
- Introduction to inferential statistics
- Probability
- Common probability distributions
- Conditional probabilities, Bayes' theorem
- Outliers

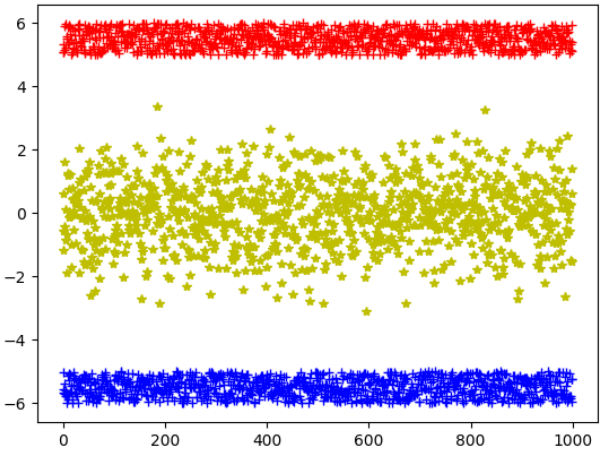
Probability

Ostbayerische Technische Hochschule Amberg-Weiden

- Machine Learning is the science of finding relationships in data in the presence of uncertainty.

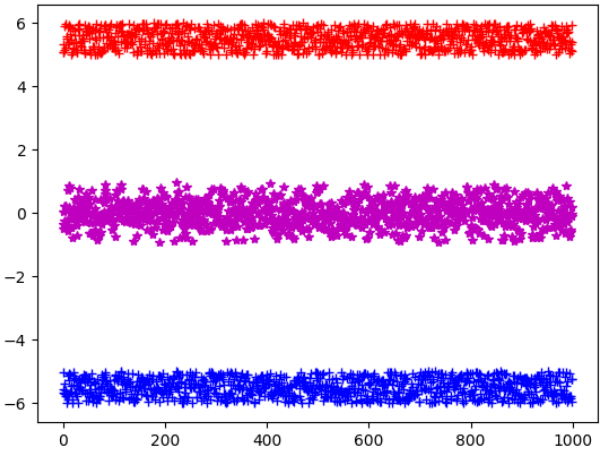
Probability

- Can you derive a relationship between the differently colored data points?



Probability

- Now?



Probability

- Machine Learning is the science of finding relationships in data in the presence of uncertainty.
- Uncertainties can have different sources:
  - Inherent randomness in the system: E.g. random fluctuations due to temperature, quantum systems, random choice of people in a poll, ...
  - Incomplete observation: A system can behave determinsitically but you do not recognize it because you cannot measure all properties of a system. Example: shell game, as the presenter you know under which cup you place the ball, while from the contestant's point of view it is random.
  - Incomplete model: We neglect certain properties of the real system for the sake of simplicity of our model. E.g. you measure the pressure of a gas in a machine while neglecting temperature effects. However, if your measurement values are collected at different temperatures, the corresponding effect seems random to you.
  - Measurement uncertainty: Every measurement has an error adding a further source of uncertainty.

Uncertainty

- In summary, various effects add uncertainty to our data.
- In planned experiments, you try to avoid uncertainty as far as possible, e.g. using laboratory systems instead of real systems.
- In Machine Learning, they are usually unavoidable,
- We work with **probabilities**.
- A probability for a specific outcome of an experiment is the fraction of this specific outcome among all outcomes during an infinite number of repetitions of the experiment. $\longrightarrow$ frequentist view
- A probability of an outcome is a measure of how certain we are that we see this outcome when we conduct the experiment. $\longrightarrow$ Bayesian view

Random Variables

- A random variable X is a variable which takes different values or states.
- Two cases: either X is dicrete (countable, also countably infinite), or continuous (not countably infinite).
- Example for discrete random variable: Throwing a die yields a number in $\{1, 2, 3, 4, 5, 6\}$.
- Example for an continuous random variable: Throw a dart arrow at a wall, measure the exact location (arbitrarily exact).
- Notice in reality you can map everything to a discrete case, due to measurement limitations. However, for practical reasons we consider everything continuous which is naturally a continuous variable.

Random Variables
Some more details

- Especially discrete variables can be subdivided further:
  - **categorical variable:** distinguishing different kinds of things, like e.g. species, color, gender, ...
  - categorical variables cannot be put into an order
  - **binary variable:** special case of categorical variable with exactly two categories
  - **ordinal variable:** Discrete variable with an order, e.g. a rank, (discrete) positions, price categories (cheap, normal, expensive), grades, ...

Random Variables
Data Scientist's view

Ostbayerische Technische Hochschule Amberg-Weiden

- Data science is all about automatically extract relationships between random variables.
- Frankly speaking, this is statistics.
- Machine Learning just adds a bunch of automation.
- And usually removes (much) rigorosity.
- "Every model is wrong, but some are useful."

Random Variables
Modelling

Ostbayerische Technische Hochschule Amberg-Weiden

- In modelling you typically differentiate between **independent** and **dependent** variables.
- Dependent variables are affected by changes in the independenten variables.
- Usually, this is just a question of definition:
  - The independent variables are those you measure or manipulate (in case you do experiments)
  - The dependent variables are those you observe.
  - Attention! Independent variables are not necessarily independent (of each other).
  - Which variable is the dependent one depends on the causality you assume.
  - Example: Time spent in front of a computer, and binary category "IT professional".
  - Are you an IT professional because you spent a lot of time in front of a computer or vice versa?

Discrete Random Variables

Ostbayerische Technische Hochschule Amberg-Weiden

- A discrete random variable X is a variable which takes different values or states x_i.
- Each value x_i gets a probability $P(x_i)$ assgined for X taking the value x_i.
- $P(x)$ is the **probability mass funtion** assigning a probability to each value x.
- These probabilitites must obey certain rules:
  - The domain of P is the whole range of possible values of X.
  - $0 \leq P(x_i) \leq 1$, with a probability of zero indicating that the value is impossible, while a probability of one indicates that the value is absolutely certain.
  - $\sum_i P(x_i) = 1$: The sum over all probabilities is one, some value will be taken.

Continuous Random Variables

Ostbayerische Technische Hochschule Amberg-Weiden

- The case of a continuous random variable X is similar, but not equal.
- We deal with a probability density function $p(x)$.
- It does not represent the probability that X assumes the value x.
- Rather, the probability to find the value of X within an infinitesimal region Δx around x is given by $p(x)\Delta x$.
- Probability densities also obey certain rules:
  - The domain of p is the whole range of possible values of X.
  - $0 \leq p(x)$
  - $\int p(x)dx = 1$: The integral over all probabilities is one, some value will be taken.
- The probability of X assuming a value within the a region R (e.g. in one dimension an interval $[a; b]$) is given by

$$P(X \in R) = \int_R p(x)dx = \int_a^b p(x)dx \tag{1}$$

- with the last equality applying for the one-dimensional example.

Joint Probability

- Probability mass functions and densities can be defined for several random variables, too.
- E.g. $P(X_1 = x, X_2 = y)$ is the **joint probability** that X_1 has the value x and at the same time X_2 has the value y.
- Notation: Random variables are separated by commata.
- If the probability (density) depends on a set of parameters those are noted right of the random variables separated by a semicolon, e.g. $P(X; \theta)$ in case the probability depends on parameters θ.

Marginal Probability

- If a probability (density) depends on several random variables but we are only interested in a subset of them we sum (integrate) over those we are not interested in.
- Example: Given $P(X_1, X_2)$ we are interested in $P(X_1)$ we get

$$P(X_1 = x) = \sum_y P(X_1 = x, X_2 = y) \tag{2}$$

- or in the continuous case

$$p(x) = \int p(x, y) dy \tag{3}$$

Conditional Probability

The probability that a random variable X_1 has a value $X_1 = x$ under the condition that another random variable X_2 has a value $X_2 = y$ is given by

$$P(X_1 = x|X_2 = y) = \frac{P(X_1 = x, X_2 = y)}{P(X_2 = y)} \tag{4}$$

Interpretation:

- Consider a group of 30 people from three different countries, with equal distribution, i.e. 10 people per country. Assume that every person only speakes the mother tongue of his or her country but no other (foreign) language.
- The random variables are the country C a person comes from and the language L the person speaks.
- The joint probability to find a person from the first country ($C = 1$) speaking the language of the first country ($L = 1$) is $P(C = 1, L = 1) = 10/30 \approx 0.33$.

Conditional Probability

$$P(X_1 = x|X_2 = y) = \frac{P(X_1 = x, X_2 = y)}{P(X_2 = y)} \tag{5}$$

Interpretation:

- The joint probability to find a person from the first country ($C = 1$) speaking the language of the first country ($L = 1$) is $P(C = 1, L = 1) = 10/30 \approx 0.33$.
- The conditional probability to find a person from country 1 given that he or she speaks the language of this country however is given by $P(C = 1|L = 1) = \frac{P(C=1,L=1)}{P(L=1)} = \frac{10/30}{10/30} = 1$.
- That means if we knew the persons speaks language 1 we know for sure he or she is from country 1.
- Notice that you cannot condition on an impossible event.

Conditional Probability

- From the definition of the conditional probability we can derive the product rule of probability theory.
- For two random variables we directly see it from the previous definition:

$$P(X_1 = x, X_2 = y) = P(X_2 = y|X_1 = x)P(X_1 = x) \tag{6}$$

- For an arbitrary number of n random variables X_i, $i = 1, \ldots n$, we can easily extend it.
- $n = 3$:

$$\begin{aligned} P(X_1, X_2, X_3) &= P(X_3|X_1, X_2) \cdot P(X_1, X_2) = P(X_3|X_1, X_2) \cdot P(X_2|X_1) \cdot P(X_1) \\ &= P(X_1) \cdot P(X_2|X_1) \cdot P(X_3|X_1, X_2) \end{aligned}$$

- or generally,

$$P(X_1, X_2, \ldots X_n) = P(X_1) \prod_{i=2}^{n} P(X_i|X_1, \ldots X_{i-1}). \tag{7}$$

Independence of Random Variables

- Two random variables are independent if their joint probability distribution factorizes accoring to

$$\forall x, y \quad p(X_1 = x, X_2 = y) = p(X_1 = x)p(X_2 = y) \tag{8}$$

- They are conditionally independent if this factorization applies dependent on a third variable:

$$\forall x, y, z \quad p(X_1 = x, X_2 = y|X_3 = z) = p(X_1 = x|X_3 = z)p(X_2 = y|X_3 = z) \tag{9}$$

Distribution of Random Variables

Ostbayerische Technische Hochschule Amberg-Weiden

- When working with data, your whole dataset is a collection of values for random variables.
- Example: Each column of the iris dataset you worked with in the introduction represents a random variable.
- What can you infer about these random variables from the values you got in the dataset?
- Let's talk about the term "distribution" first.

Distribution of Random Variables

Ostbayerische Technische Hochschule Amberg-Weiden

- A probability distribution how likely the possible values of a random variable occur.
- A probability distribution can be described
  - Graphically (however, this is not an exact way)
  - By all its moments (expectation (mean) values for the variables X, X^2, X^3, X^4, ...)
  - Some by a closed formula.

Expectation Value and Variance

- Expectation value and variance are the first two moments of a probability distribution.
- We assume in the following we know the probability mass or density function, respectively, for our random variable.
- We will see afterwards how we compute these quantities from data.
  - The expected value, or mean, of a random variable X is defined as

$$E_{X\sim P}[X] = \sum_i x_i P(X = x_i) \quad \text{or} \quad E_{X\sim P}[X] = \int_{-\infty}^{\infty} xp(x)dx. \tag{10}$$

  - The variance of a random variable X is defined as

$$\text{Var}(X) = E[(X - E[X])^2]. \tag{11}$$

  - The standard deviation of a random variable X is given by $\text{Std}(X) = \sqrt{\text{Var}(X)}$.

Higher Moments

- The third moment is called the **skewness**.
- It measures the lack of symmetry in the distribution.
- It is the normalized third moment:

$$\text{skew}[X] = E\left[\left(\frac{(X - \mu)}{\sigma}\right)^3\right] \tag{12}$$

- Normally distributed variables have a skewness of zero.

Higher Moments

- The fourth moment is called the **kurtosis**.
- Kurtosis measures how much scores cluster at the end of the distribution (tails).
- It therefore is a measure for the pointyness
- It is the normalized third moment:

$$\text{kurt}[X] = E\left[\left(\frac{(X-\mu)}{\sigma}\right)^4\right] \tag{13}$$

- Normally distributed variables have a kurtosis of zero.

Bernoulli Distribution

- For a single, binary random variable

$$\begin{aligned} P(X=1) &= p \\ P(X=0) &= 1-p \\ P(x) &= p^x(1-p)^{1-x} \quad (x \in \{0;1\}) \\ E_x[x] &= p \\ \text{Var}_x[x] &= p(1-p) \end{aligned}$$

- Example: Tossing a coin.
- Can you deal with a non-binary random variable using the Bernoulli distribution? For example for throwing a (balanced) die?

Gaussian or Normal Distribution

- The Gaussian distribution or normal distribution is (due to its significance in statistics) probably the most widely used continuous probability distribution.
- It plays a central role in limit theorems in statistics.

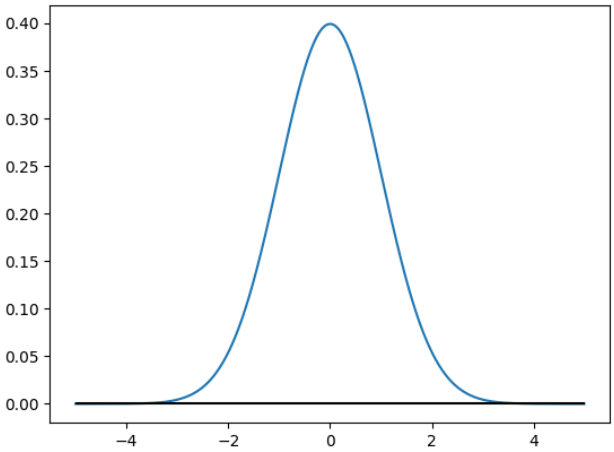
$$N(x;\mu,\sigma) = \sqrt{\frac{1}{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \tag{14}$$

- Mean value and variance are given by

$$E_{\mathcal{N}}[X] = \mu \qquad \mathrm{Var}_{\mathcal{N}}(X) = \sigma^2 \tag{15}$$

  - Verify the mean value using the defintiion $E_{\mathcal{N}}[X] = \int_{-\infty}^{\infty} x\mathcal{N}(x;\mu;\sigma^2)dx$.

Gaussian or Normal Distribution



Empirical Mean and Standard Deviation

- To compute the empirical mean $\overline{x}$ from data $\{x_i\}_{i=1}^N$, you calculate

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{16}$$

- To compute the empirical variance s^2 from data $\{x_i\}_{i=1}^N$, you calculate

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2 \tag{17}$$

Further Important Empirical Quantities

- The center of a distribution is often well described by the **median**.
- The median value of a set of a data sample $\{x_i\}_{i=1}^N$ is chosen such that 50% of the datapoints are larger and 50% are smaller (or equal).
- To compute it, first sort the values $\{x_i\}_{i=1}^N$ in ascending order.
  - In case you have an odd number of values there is a clear center value. It is found at position $\lceil \frac{N}{2} \rceil$ (when counting starts at 1)
  - In case of an even number take the average of the two center values $\frac{x_{N/2}+x_{N/2+1}}{2}$.
- The advantage of the median compared to the mean: It is less susceptible to outliers.
- *Example:* Consider the values $5.1, 5.2, 5.5, 5.3, 5.2, 10.3$ with an obvious outlier. Compute mean and median.

Further Important Empirical Quantities

Ostbayerische Technische Hochschule Amberg-Weiden

- **Mode** of the data
- Simple: The data point (or interval) that occurs the most in the data.
- Determine the mode by sortiung the scores and look for the largest one.
- Example: For the Gaussian distrbution the mean value is the mode.
- Attention: It may happen that several values have the same score.
- Multimodal distribution (bimodal in case of two)

Exploring Data Graphically

Ostbayerische Technische Hochschule Amberg-Weiden

- Histogram plots
- Discrete random variable: Plot a histogram with a bar per possible value which shows the number of occurrences of that particular value or, better, the rate (number / total number of data points)
- Continuous random variable: Subdivide the range of values into disjunct intervals (bins) item A variant, especially in case of continuous variables, is a density plot.
- In this case the histogram counts are divided by the total counts and the bin width.
- The resulting graph has an integral of 1 when integrating over all values.