[AIMLBD] MACHINE LEARNING, BIG DATA, ARTIFICIAL INTELLIGENCE per medicina e chirurgia high tech

L01: Machine Learning Overview

Dott. Giorgio De Magistris

demagistris@diag.uniroma1.it

CORSO DI LAUREA IN MEDICINA E CHIRURGIA HIGH TECH

SAPIENZA
UNIVERSITÀ DI ROMA

FACOLTÀ DI INGEGNERIA DELL'INFORMAZIONE, INFORMATICA E STATISTICA

DIPARTIMENTO DI INGEGNERIA INFORMATICA, AUTOMATICA E GESTIONALE

What is Machine Learning?

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.

Mitchell (1997)

Experience

We can partition the ML algorithms into 3 macro categories according to which kind of experience is allowed during training.

Supervised Learning

The dataset is a set of tuple (x,y) where x is the input and y is the target.

Unsupervised Learning

The dataset is just a set of samples

Reinforcement Learning

The dataset is acquired sensing the environment

Task

Supervised Learning Tasks:

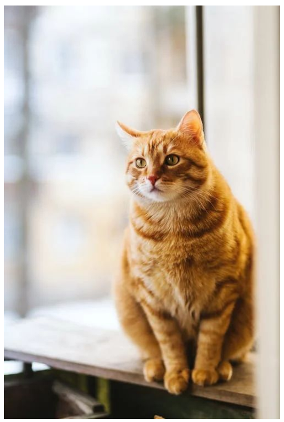
- Classification
- Regression
- …

Unsupervised Learning Tasks:

- Dimensionality Reduction
- Clustering
- Density Estimation
- …

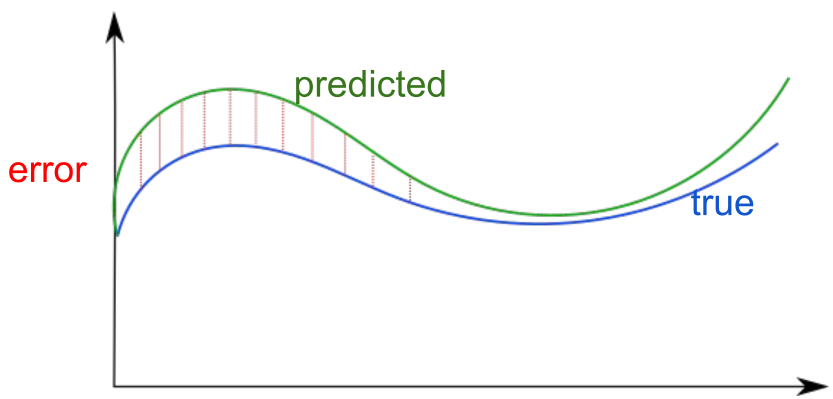
Performance

The performance measure depends on the task, and it is necessary to quantify the ability of the ML algorithm to solve the specific task. It could be a measure of the error to minimize or a score to maximize.



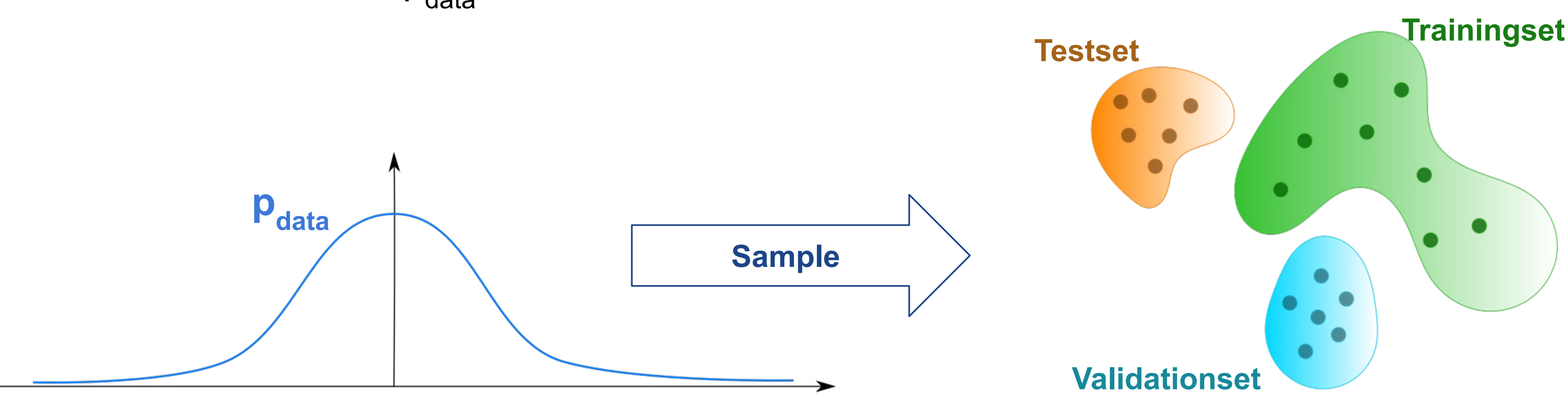
Is this a mammal?

$(-1)^{ytrue} \log(ypred)$



Data generation process

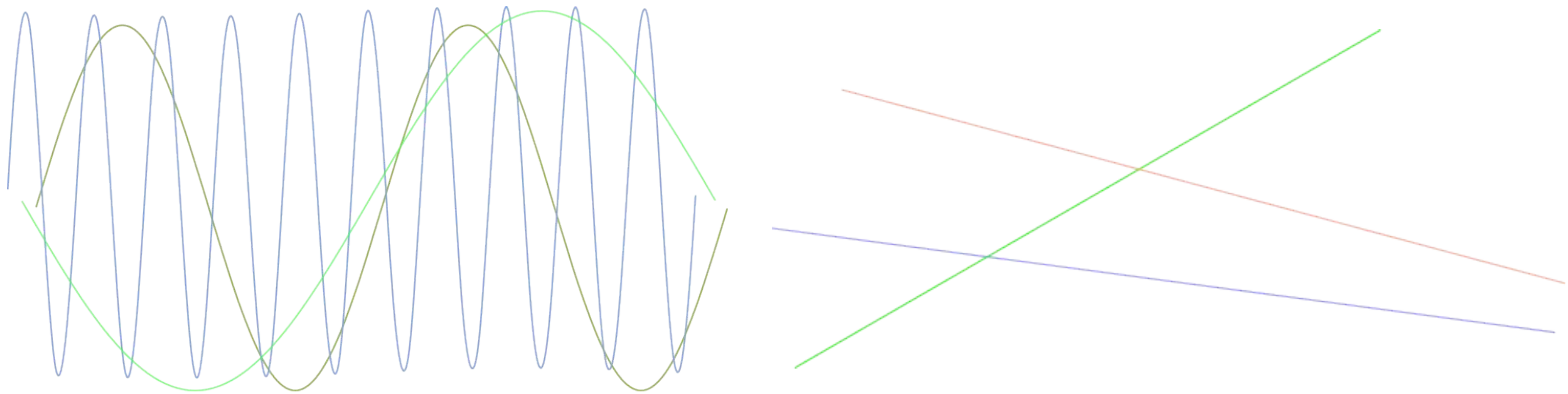
- Assume data is sampled from an (unknown) probability distribution p_{data}
- We usually make the *"independent and identically distributed"* (**I.I.D.**) assumption, meaning that each data sample comes from the same distribution p_{data} and that samples are independent
- Assume we have a dataset $D = \{x_1,...,x_n\}$. The independence assumption allows us to factorize the probability of observing the dataset as:
  - $P(D) = P(X_1=x_1, \dots, X_n=x_n) = \prod^n_{i=1} P(X_i=x_i)$
- We also assume that **trainingset**, **validationset** and **testset** are sampled from the same distribution p_{data}



Hypotheses Space and Inductive Bias

- The Hypotheses space is the set of all the functions that the ML algorithm is allowed to learn

- The prioritization of some hypotheses (restriction of hypothesis space) is an inductive bias.



Generalization

- We optimize the ML model on the trainingset in order to minimize the **training error**
- However we are interested in the error on new data, that is not encountered during training
- The error on the testset is called **generalization error** and it quantifies the ability of the model to generalize on unseen data

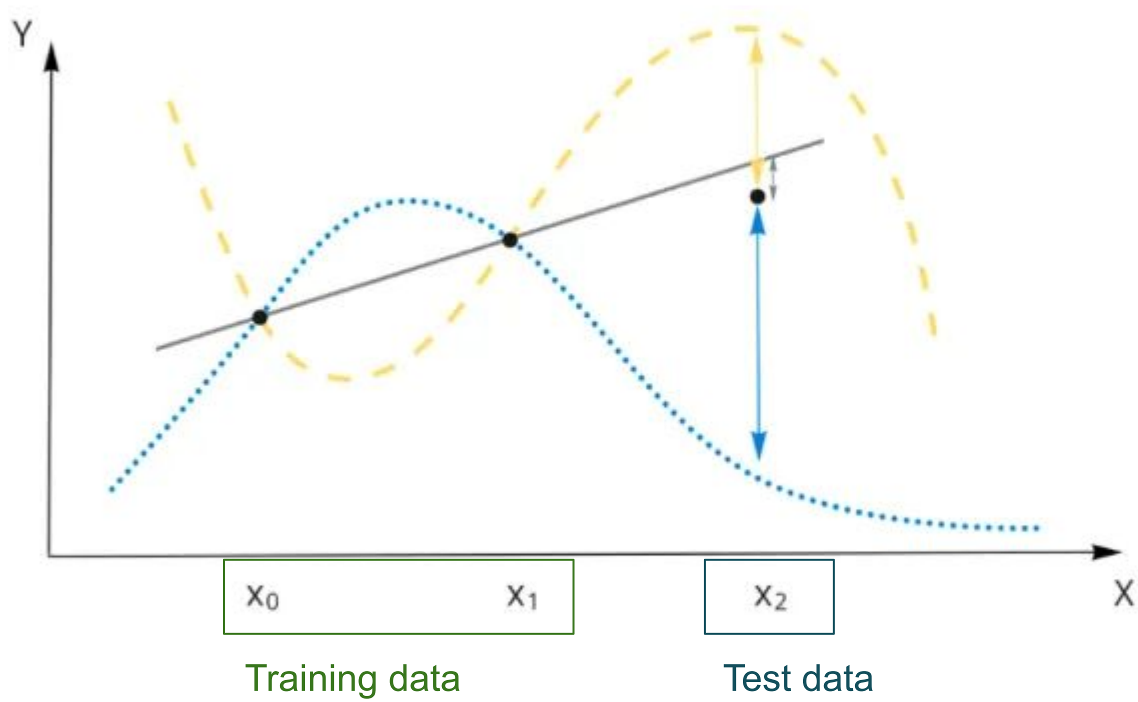


Image credit: [The Inductive Bias of ML Models, and Why You Should Care About It]

Overfitting and Underfitting

- When training a ML model two things can happen:
  - The model struggles to minimize the training error, this phenomenon is called **Underfitting**
  - The model reaches a small training error, but the gap between training error and generalization error is large, this phenomenon is called **Overfitting**

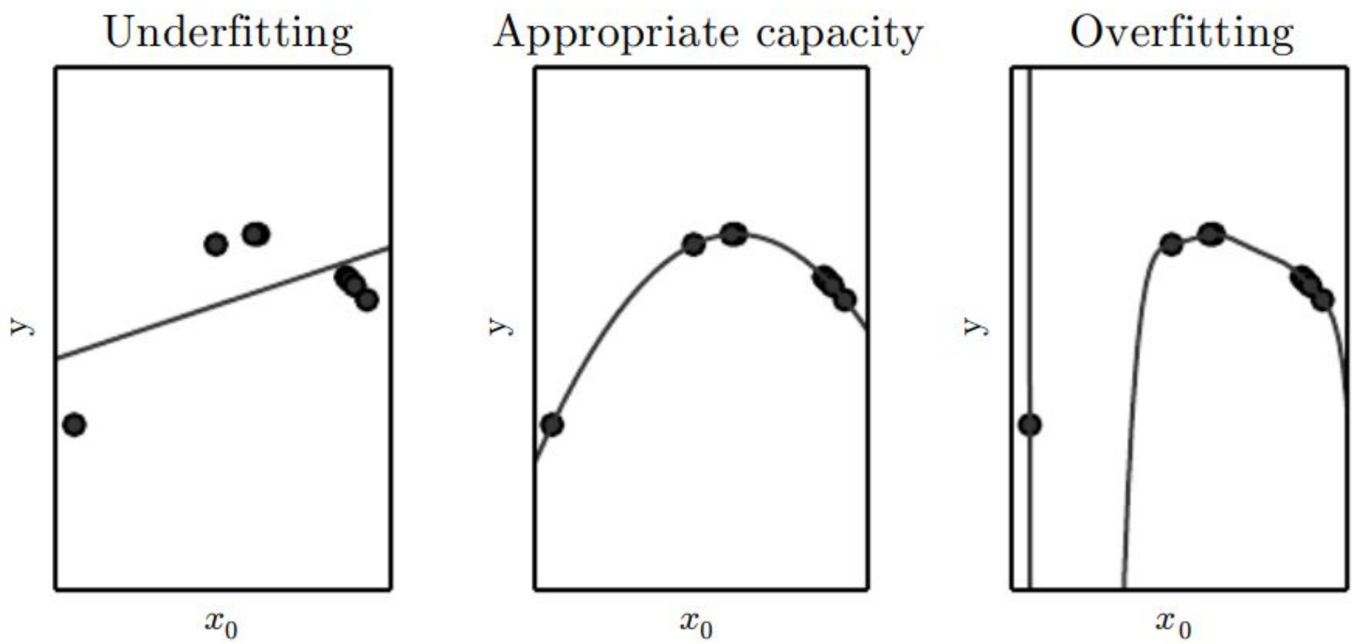


Image credit: [Deep Learning Book]

Capacity

- I can balance the overfitting and underfitting acting on the model capacity
- The model capacity is the complexity of the model:
  - A model with high capacity can fit a wide range of functions
  - A model with limited capacity can fit a small range of functions
- Usually I have the best generalization error when:
  - the capacity of the model is well proportionate to the task
  - I have a lot of training examples

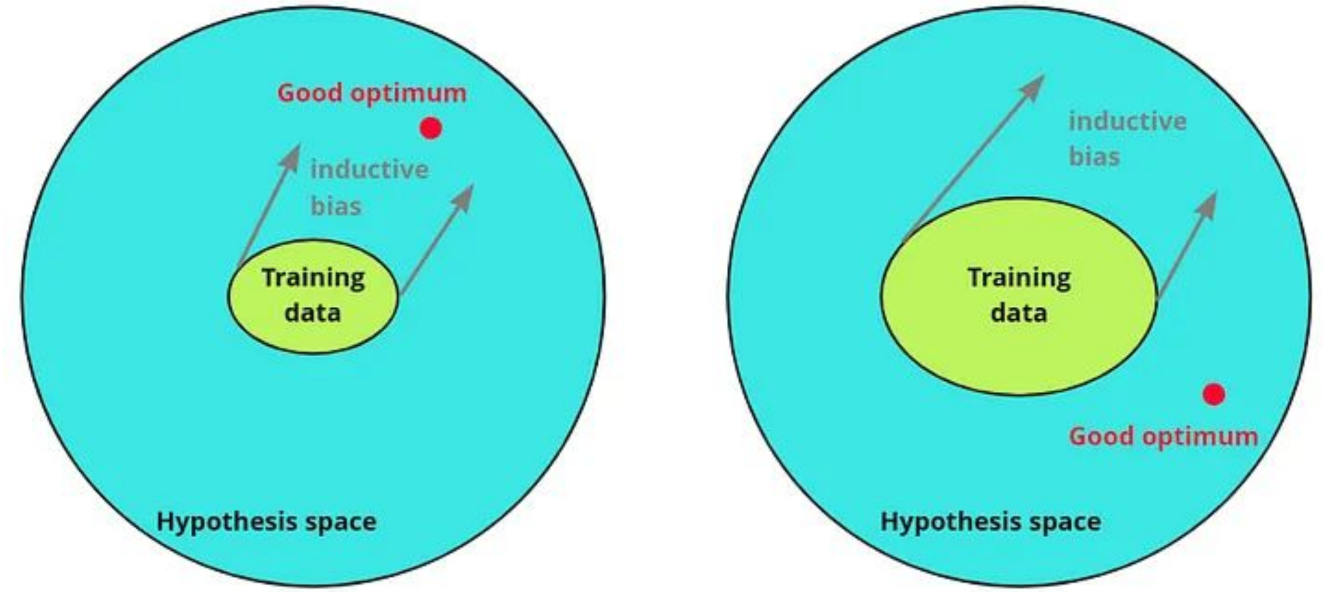
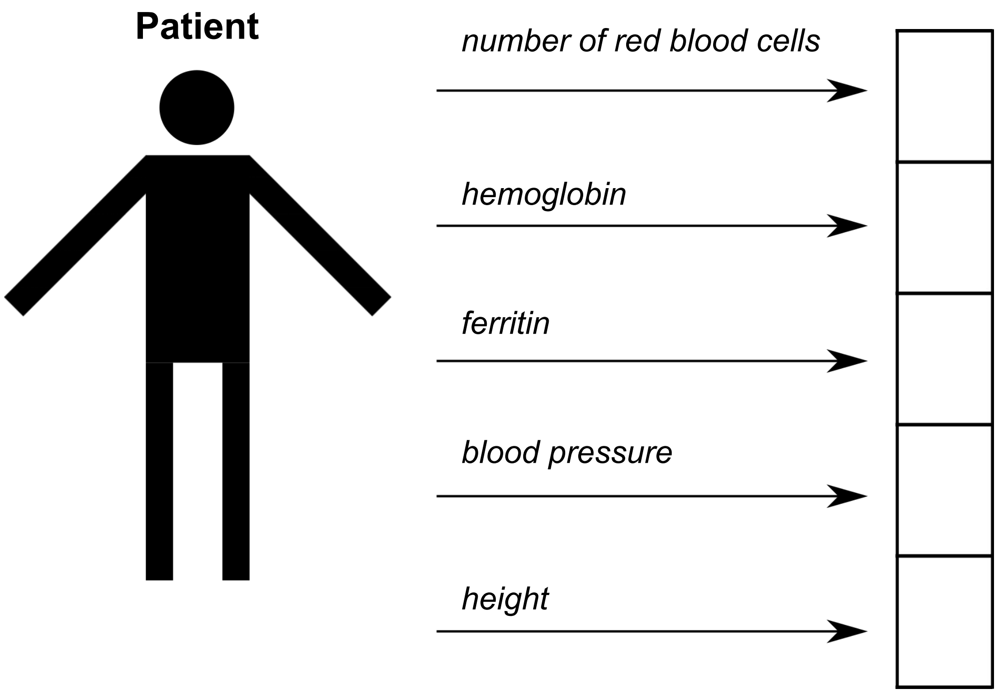


Image credit: [The Inductive Bias of ML Models, and Why You Should Care About It]

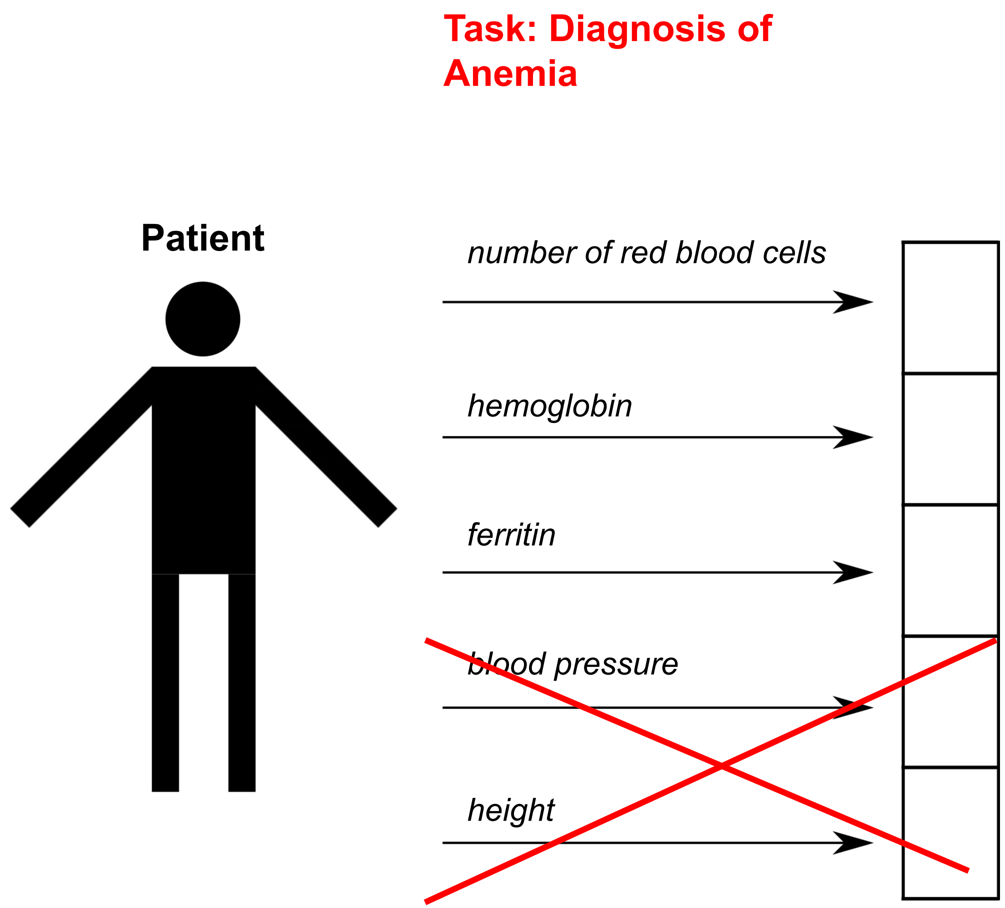
Feature and Feature Vector

- Each data sample is described as a set of features
- A feature is an individual characteristic of the observed object
- Usually it is represented as a real value (even if different types of feature exist, like categorical features or graph features)
- In the case of real valued features, a sample is represented as a vector which components are the values of the features called feature vector



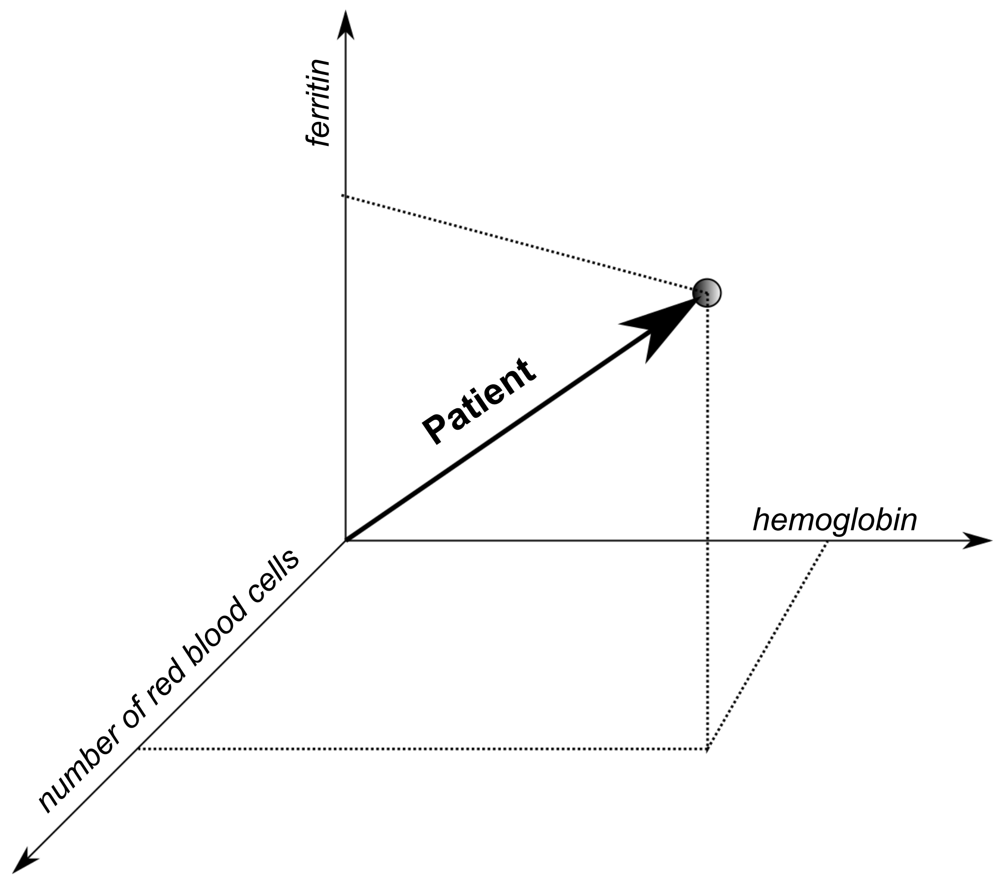
Features and Task

- I can use many different sets of feature to describe the same object
- Which is the best representation?
- It depends on the task !



Feature Space

- Once I've chosen a representation, the set of all feature vectors form a Feature Space
- I can represent an object as a vector in the feature space, which coordinates are the values of the features
- Usually a feature space with n features is represented as the vector space R^n



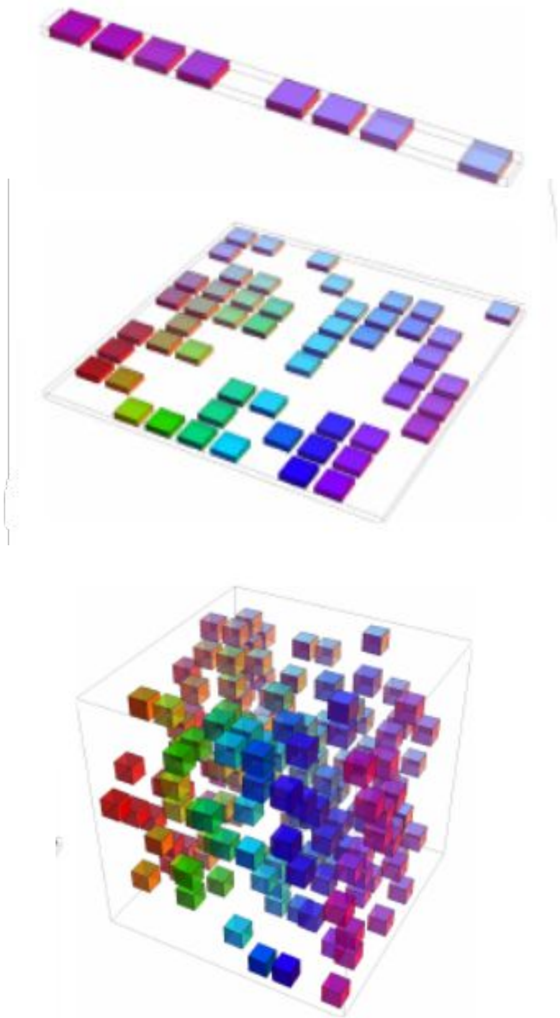
Learning

- In this context, learning means to find a function that maps the input feature space to the desired output

- The feature space must exhibit a sort of structure, otherwise no learning is possible

$$f(\nearrow) = Anemic$$

The Curse of Dimensionality



- Assume we represent our objects using n features, and each feature has m admissible values
- The feature space has m^n elements, which is exponential in n
- In general to learn a function over a feature space we need at least a number of samples that is proportional to the size of the feature space
- Hence the number of samples grows exponentially with respect to the dimensionality of the feature space (n)

Image credit: [Deep Learning Book]

Intrinsic Dimensionality

- How can we learn when the dimensionality of the input is higher than the number of samples?

- Often the input data that comes in an high dimensional space (images, audio, text etc.) but can be represented in a lower dimensional feature space in which learning is feasible

- The intrinsic dimensionality is the minimum number of meaningful dimensions needed to capture the essential characteristics or structure of the data without introducing unnecessary noise or redundancy

- The intrinsic dimensionality of a dataset can be influenced by various factors, such as the type of data (e.g., images, text, time series) and the nature of the problem (e.g., classification, regression)