Introduction to Machine Learning

Chandresh

The book

Reference book.

1. Bishop, Christopher M. Pattern Recognition and Machine Learning. New York :Springer, 2006
2. Murphy, Kevin P. Machine Learning: A Probabilistic Perspective. Cambridge, MA: MIT Press, 2012. Print.

What is Artifical Intelligence?

Definition

Artificial Intelligence (AI) is the field of computer science focused on building systems that can perform tasks that typically require human intelligence.

- Understanding natural language (e.g., chatbots, translation)
- Recognizing patterns (e.g., image and speech recognition)
- Learning from data (e.g., recommendation systems)
- Making decisions (e.g., autonomous vehicles)

In short: AI enables machines to "think", "learn", and "act" intelligently.

What is Machine Learning?

Definition by Tom M. Mitchell

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.

Informally, machine learning is the is the study of computer algorithms that improve automatically through experience and by the use of data.

Some Basic Terminologies

- **Machine** A tool which can perform some task
- **Learning** is the process of acquiring new understanding, knowledge, values, skills, preferences, and attitude.
- **Knowledge** is familiarity with or awareness of some facts or situation.
- **Intelligence** is the ability to think, learn and act according to a given situation.

A bit of history

- The term machine learning(ML) was coined in 1959 by **Arthur Samuel**.
- ML is a sub-field of broader term called artificial intelligence.
- **Alan Turning** originally proposed the question: Can machines think?
- Later on question was replaced with: Can machines do what we can do?

A bit of history

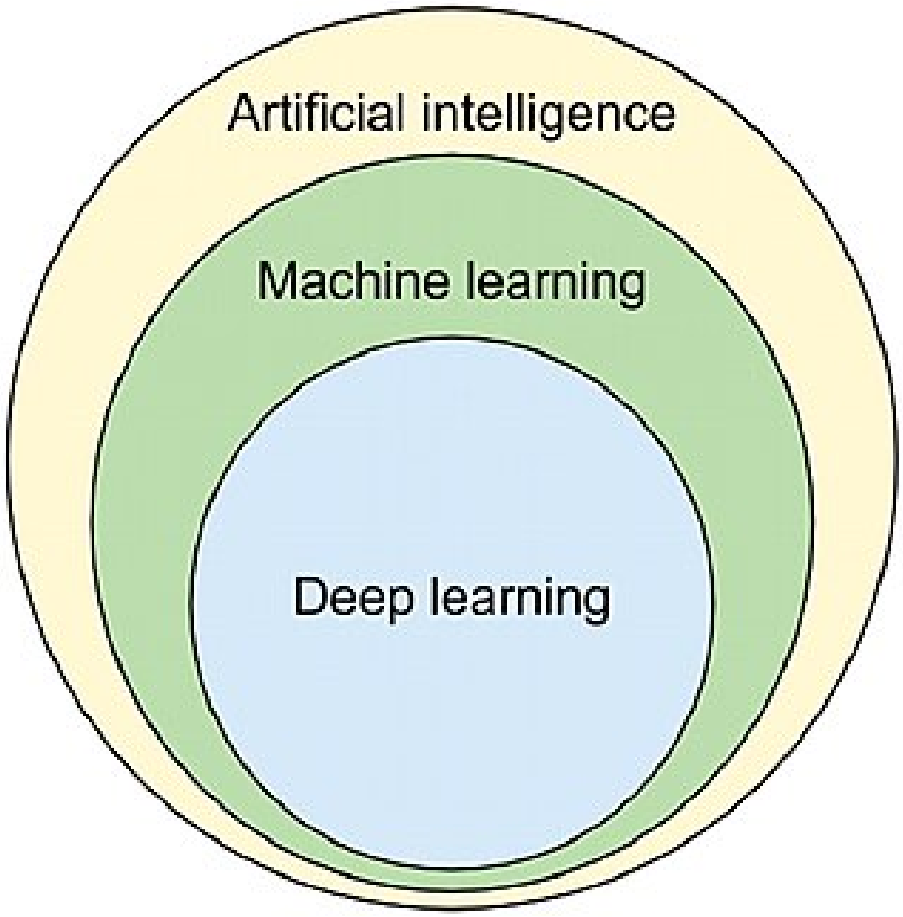


Figure: ML as a sub-field of AI

A bit of history

- 1960s and 70s: **Models of human learning** High-level symbolic descriptions of knowledge, e.g., logical expressions or graphs/networks, e.g., (Karpinski & Michalski, 1966) (Simon Lea, 1974). Winstons (1975) structural learning system learned logic-based structural descriptions from examples.
- 1970s: **Genetic algorithms**, Developed by Holland (1975)
- 1970s - present: **Knowledge-intensive learning** A tabula rasa approach typically fares poorly. To acquire new knowledge a system must already possess a great deal of initial knowledge. Lenats CYC project is a good example.

A bit of history

- 1980 The First Machine Learning Workshop was held at Carnegie-Mellon University in Pittsburgh.
- 1980 Three consecutive issues of the International Journal of Policy Analysis and Information Systems were specially devoted to machine learning.
- 1981 - Hinton, Jordan, Sejnowski, Rumelhart, McLeland at UCSD Back Propagation alg. PDP Book
- 1986 The establishment of the Machine Learning journal.
- 1987 The beginning of annual international conferences on machine learning (ICML). Snowbird ML conference.
- 1988 The beginning of regular workshops on computational learning theory (COLT).
- 1990s Explosive growth in the field of data mining, which involves the application of machine learning techniques.

Difference between ML and AI and Data Mining

- ML learns and predicts based on passive observations.
- AI implies an agent interacting with the environment to learn and take actions that maximize its chance of successfully achieving its goals.
- Data mining focuses on the discovery of (previously) unknown properties in the data.

Types of Machine Learning

- **Supervised learning:** Given a set of features and labels, learn a model that will predict a label to an unseen data point.
- **Unsupervised learning:** Discover patterns in data without using "labels".
- **Reinforcement learning:** How to take action in a given environment where each action results in some pay-off (reward/penalty)
- Others: active learning, online learning, semi-supervised learning, deep learning, meta-learning, and so on.

Supervised Learning

Definition

Formally, a supervised learning on data $\mathcal{D} = \{x_i, y_i\}_{i=1}^{n}$ where $x_i \in \mathcal{R}^d, y_i \in R$ of a model M tried to find a function f such that $f : x_i \mapsto y_i$

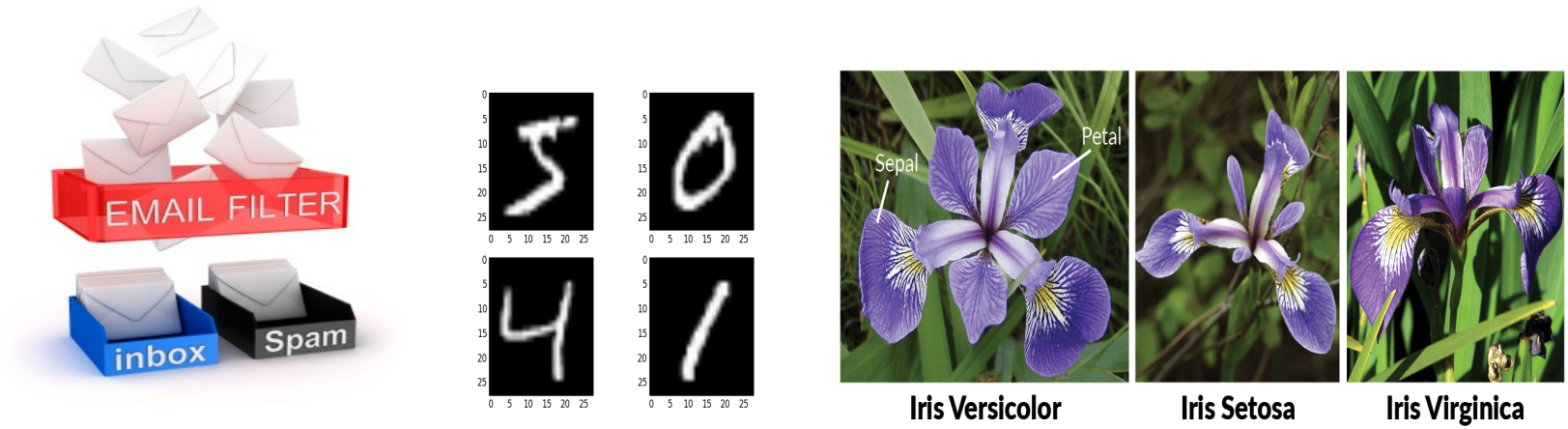
Types:

- Classification
- Regression

Supervised Learning: Classification

In classification problem, the label y is a finite and discrete number also called the number of classes.

- Email classification into spam or ham
- MINIST digit classification
- Iris data classification



(a) (b) (c)

Figure: (a) Email classification (c) Digit classification (c) Iris data

Supervised Learning: Regression

In regression problem, the label y can take any real value.

- Predicting house price based on its attributes
- predict sales based on the past sales
- Predict marks based on hours of study

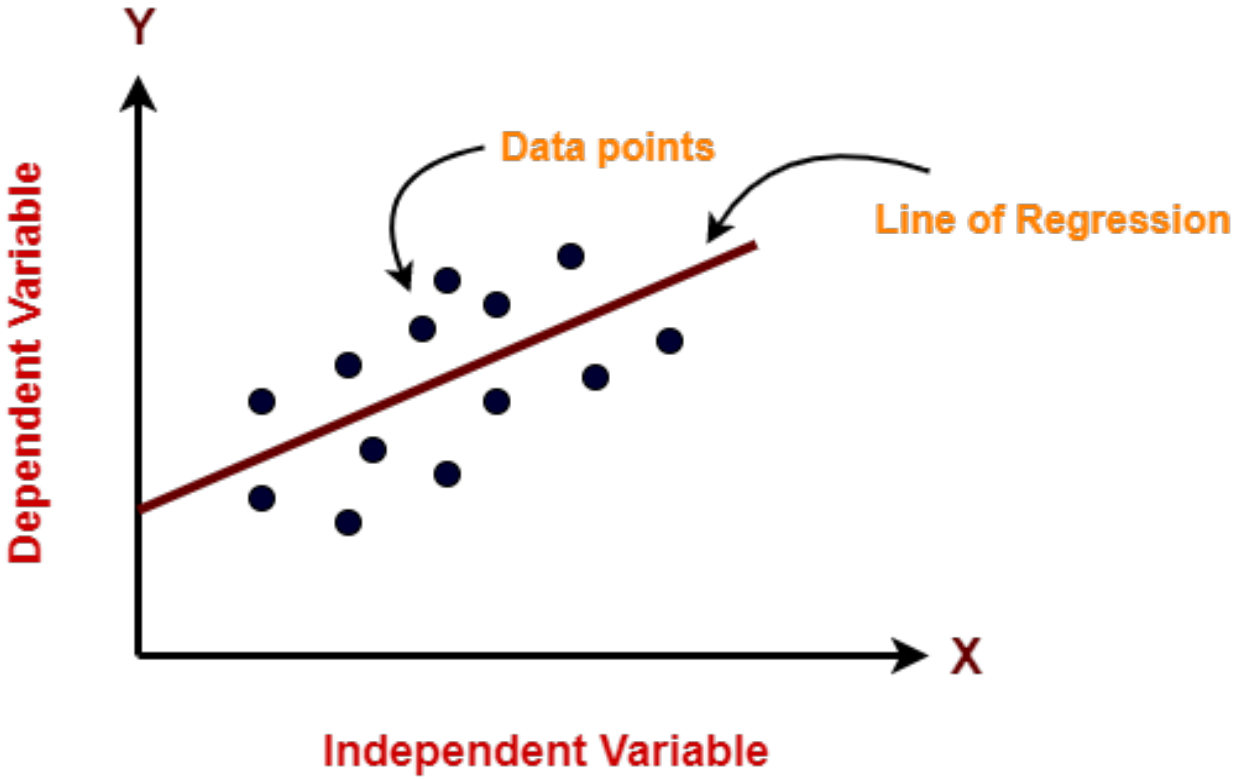


Figure: Linear regression example

Unsupervised Learning

Definition

In unsupervised learning, we are given the output data without any inputs and the goal is to find hidden patterns.

- Unlike supervised learning, we are not told what the desired output is for each input.
- Unsupervised learning is arguably more typical of human and animal learning.

Examples:

- Clustering
- Discovering latent factors
- Collaborative filtering

Unsupervised Learning: Clustering

Clustering is grouping data such that items similar to each other in *some sense* are in the same group

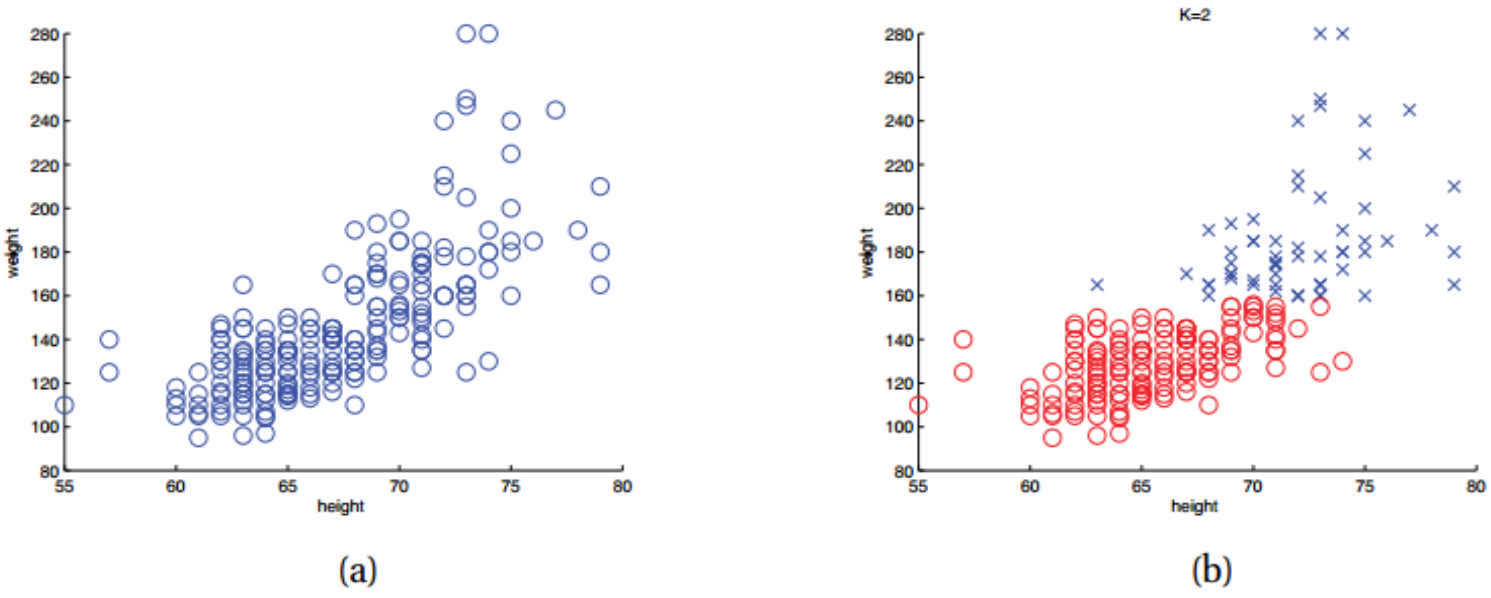


Figure: Clustering

Unsupervised Learning: Discovering latent Factors

- Big data is usually of very high dimension
- Difficult to analyse high dimensional data
- Project high dimension data to a lower dimension called **latent feature**

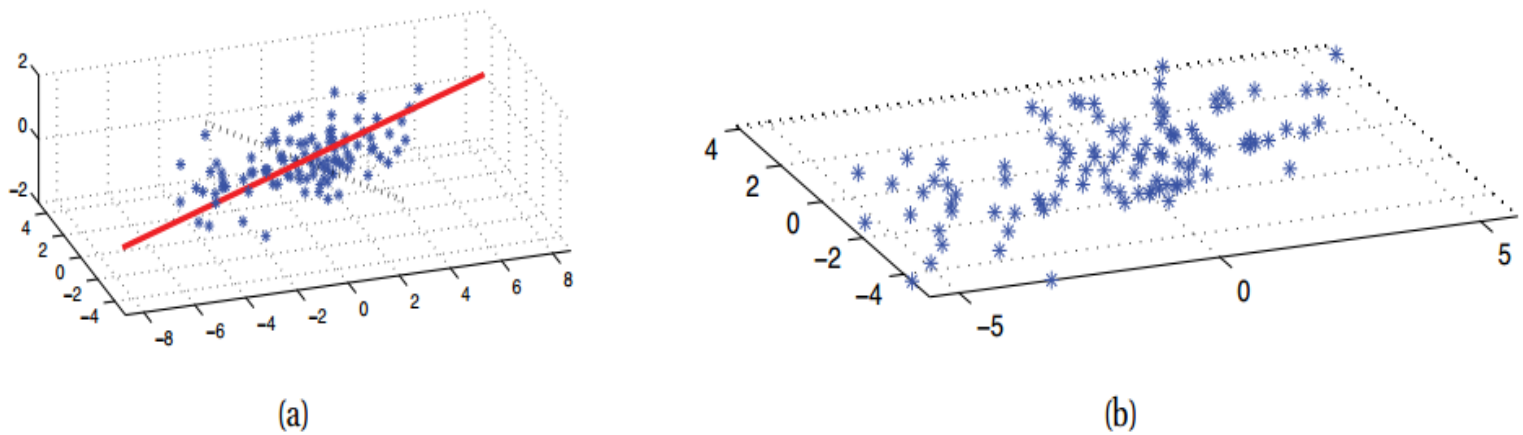


Figure: A set of points that live on a 2d linear subspace embedded in 3d. The solid red line is the first principal component direction. The dotted black line is the second PC direction. (b) 2D representation of the data.

Unsupervised Learning: Collaborative Filtering

Collaborative filtering tackles the problem of predicting unrated movies rating given a matrix of users and their rating. In particular, collaborative filtering uses the idea that users having similar preferences will rate the movie similarly.

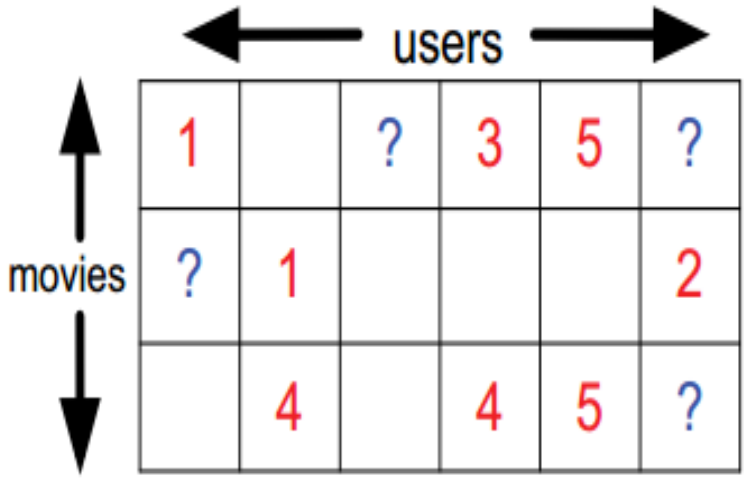


Figure: Example of movie-rating data. Training data is in red, test data is denoted by ?, empty cells are unknown.

Reinforcement Learning

Reinforcement learning aims at the question: how to act or behave when given occasional reward or punishment signals.

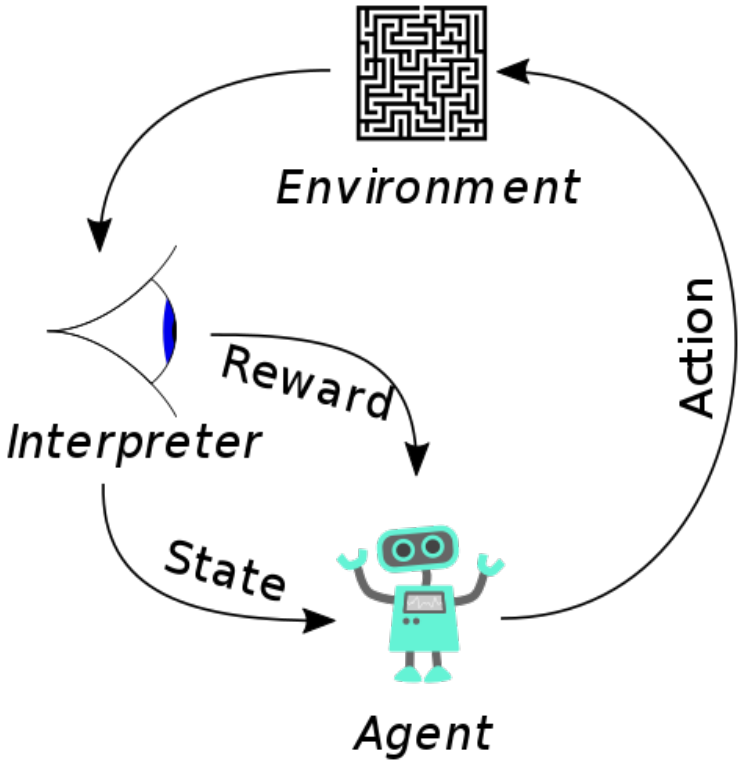


Figure: The typical framing of a Reinforcement Learning (RL) scenario: an agent takes actions in an environment, which is interpreted into a reward and a representation of the state, which are fed back into the agent. Source: wiki

Some Basic Concepts in ML

- Parametric vs non-parametric model
- The curse of dimensionality
- Overfitting
- Model selection
- No-free lunch theorem

Parametric vs Non-parametric Model

- A **parametric** model is one in which the number of parameters is fixed.
  - Parametric models are fast during inference (why?) but the disadvantage is stronger assumption about the nature of the data distribution.
  - An example is linear/logistic regression.
- The parameters grows with the amount of training data in a **non-parametric** model.
  - Non-parametric models are more flexible, but often computationally intractable for large datasets.
  - An example is K-nearest neighbour where the number of parameters is just 1. //It appears that the nearest-neighbor fits have a single parameter k, the number of neighbors. The effective number of parameters of nearest neighbors is N/k and decreases with increasing k. To get an idea of why, note that if the neighborhoods were non-overlapping, there would be neighborhoods and we would fit one parameter (a mean) in each neighborhood.

The Curse of Dimensionality

- Mathematically, the curse of dimensionality is the problem caused by the exponential increase in volume associated with adding extra dimensions to Euclidean space.
- Intuitively, most ML models perform poorly in higher dimensions and this phenomena is called curse of dimensionality.

The Curse of Dimensionality

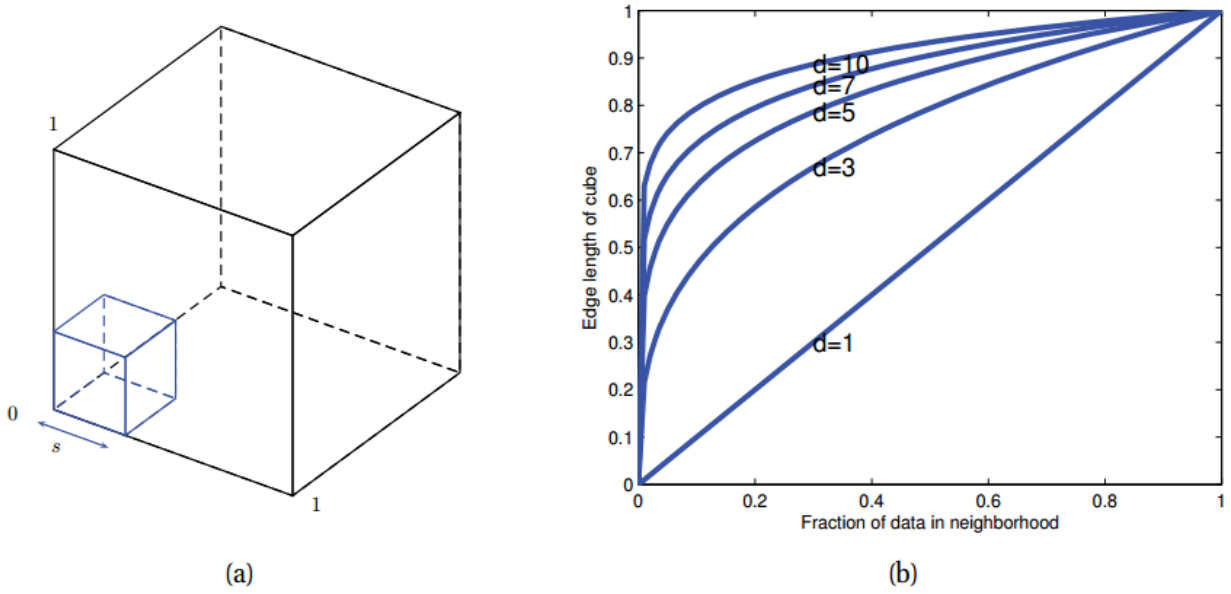


Figure: Illustration of the curse of dimensionality. (a) We embed a small cube of side s inside a larger unit cube. (b) We plot the edge length of a cube needed to cover a given volume of the unit cube as a function of the number of dimensions. Edge length is given by $e_D(f) = f^{1/D}$. If D=10 and we want to take 10% of the data in the neighbourhood, $e_1 0(.1) = 0.80$

Overfitting

Overfitting is the problem of model trying to fix small errors. That is, model tries to memorize small variations in the training data as a result has poor generalization (bad performance on test data).

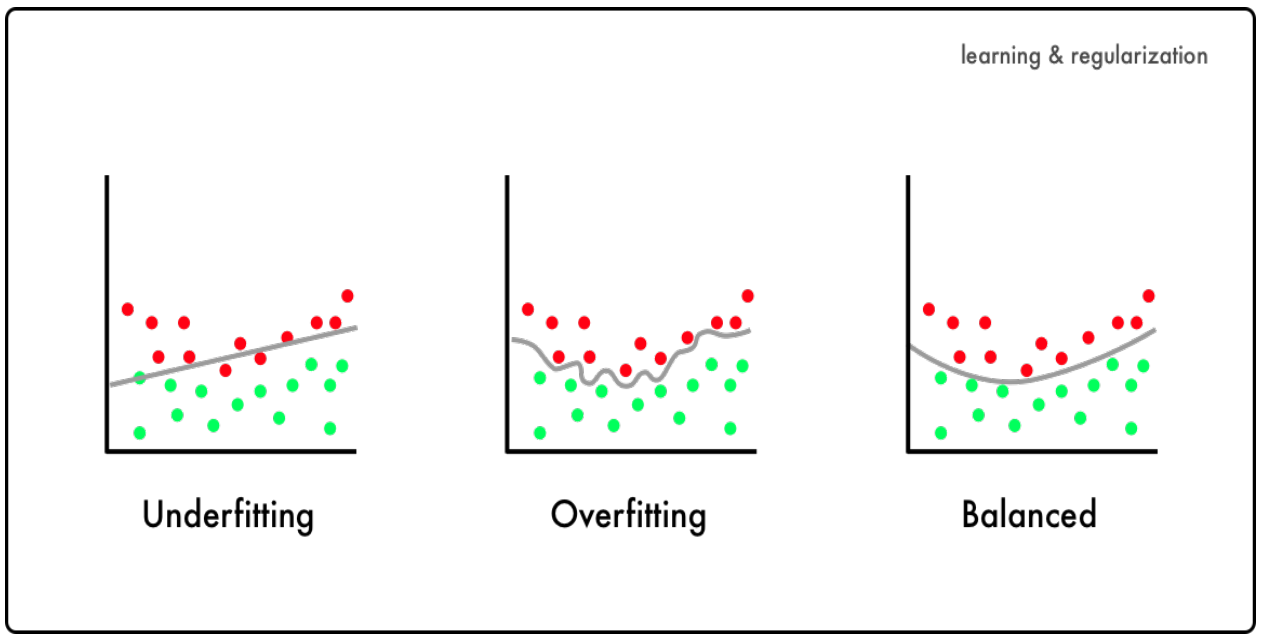


Figure: Polynomial of degrees 2, 20 and 10 fit by least squares to 21 data points.

Overfitting

Ways to prevent overfitting:

1. Hold-out
2. Cross-validation
3. Data augmentation
4. Feature selection
5. L1 / L2 regularization
6. Remove layers / number of units per layer
7. Dropout
8. Early stopping

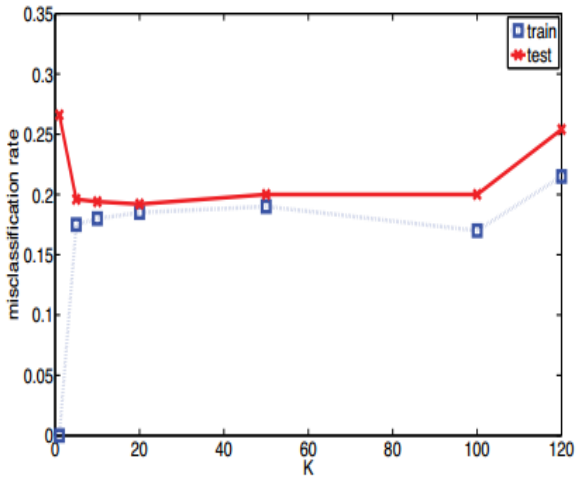
Model Selection

- Model selection deals with the problem of picking the model of right complexity. For example, given polynomials of various degree, which one should we pick?
- Usually we pick model which gives the least mis-classification rate on the training set, i.e.,

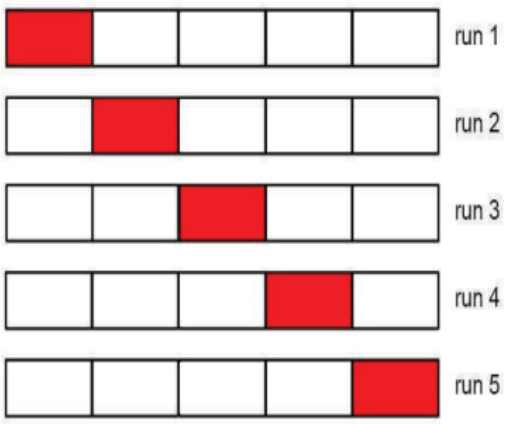
$$err(f, D) = \frac{1}{N} \sum_{n=1}^{N} \mathbb{1}(f(x_i) \neq f(y_i))$$

where $f(x)$ is the model.

Model Selection



(a)



(b)

Figure: (a) Misclassification rate vs K in a K-nearest neighbor classifier. On the left, where K is small, the model is complex and hence we overfit. On the right, where K is large, the model is simple and we underfit. Dotted blue line: training set (size 200). Solid red line: test set (size 500). (b) Schematic of 5-fold cross validation.

No-free Lunch Theorem

Statement

All models are wrong, but some models are useful. George Box [Box et al., 1987]

- We can use methods such as cross validation to empirically choose the best method for our particular problem.
- However, there is no universally best model this is sometimes called the no free lunch theorem [Wolpert, 1996].
- The reason for this is that a set of assumptions that works well in one domain may work poorly in another.

Bibliography I


Box, G. E., Draper, N. R., et al. (1987).
Empirical model-building and response surfaces, volume 424.
Wiley New York.


Wolpert, D. H. (1996).
The lack of a priori distinctions between learning algorithms.
Neural computation, 8(7):1341–1390.