Machine Learning, Basics for Chemists

Juhani Teeriniemi

Aalto University
School of Chemical Technology

Pitäisi ehkä esitellä geneettisen algoritmin sijaan/lisäksi yksinkertaisempi polkuparametrisointimenetelmä.
Pitäisikö kertoa PCAsta.

Machine Learning, Introduction

Why to use machine learning?

- Almost all science is fitting models to datasets.
- Experiments are designed to collect data from which knowledge is extracted by finding simple models that explain the observations
  This is called induction.
- Manual analysis of data is slow and costly.
- Manual analysis is often too difficult and would require a person with specific knowledge about the problem.

⟶ There is growing interest for computer models that can analyze data and extract knowledge automatically.
This is called *machine learning*.

Machine Learning, Introduction

Some applications of machine learning:

- Chemistry (chemoinformatics / toxicology).
- Bioinformatics (gene expressions).
- Mechanics (robotics / adaptive control).
- Signal processing (climatology / financing).
- Image processing (computer vision / face recognition).
- Natural language processing (machine translation).
- Marketing (customer classification).
- …

Machine Learning, Introduction

- Machine learning emphasizes combination of statistical methods and algorithms.
- Machine learning cares the problem structure (for example, are we working with correlated or uncorrelated variables), but not much about some rigorous aspects like local minimum vs. stationary points or optimization bounds.
- Machine learning is strongly related to data mining and big data. Data mining emphasizes algorithms that are sufficiently fast for large data volumes.

Machine Learning, Methods

Machine Learning Methods

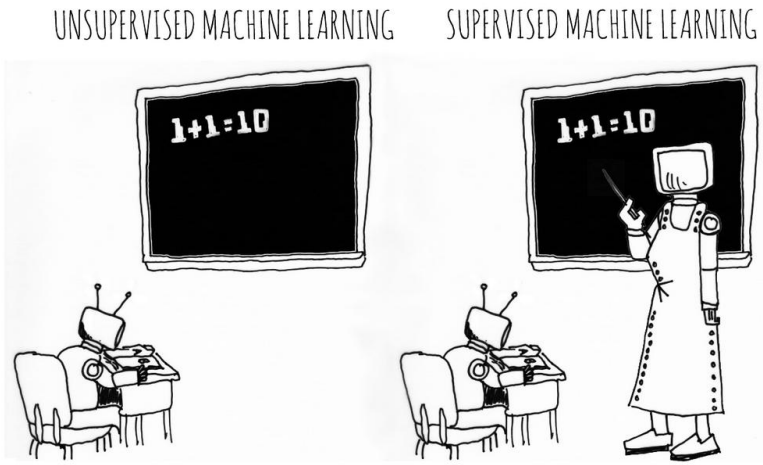
Supervised Learning
- The aim is to learn mapping from input to output.
- Output is known by a supervisor.

Unsupervised Learning
- The aim is to find ”something interesting”.
- Output is not known (known also as density estimation in statistics).

Reinforcement Learning
- Seemingly good solutions are encouraged to develop further.

Classification
- Outputs are discrete.

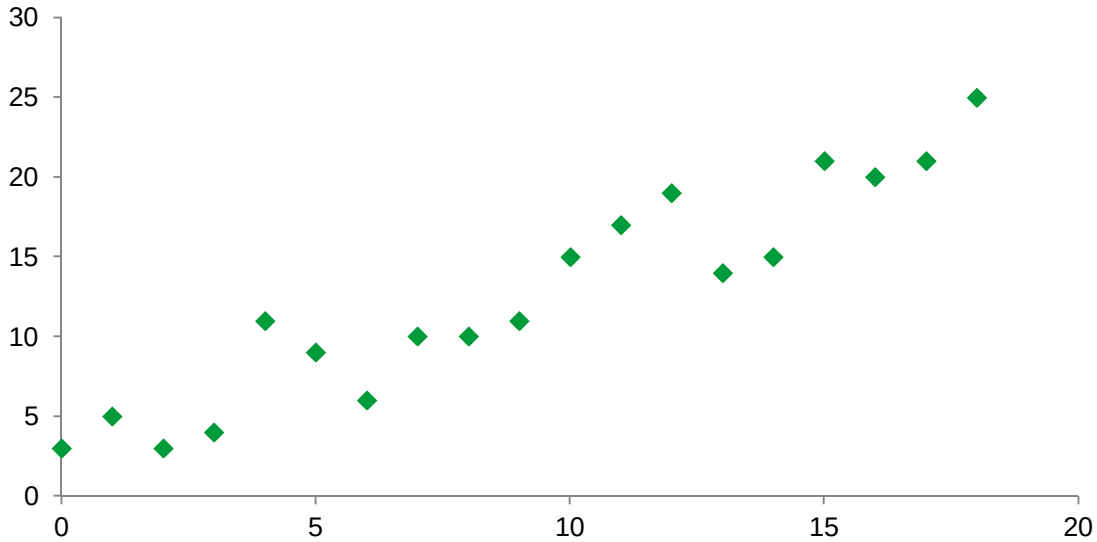
Regression
- Outputs are continuous.
- Typically noisy input.



Machine Learning, Regression and Validation

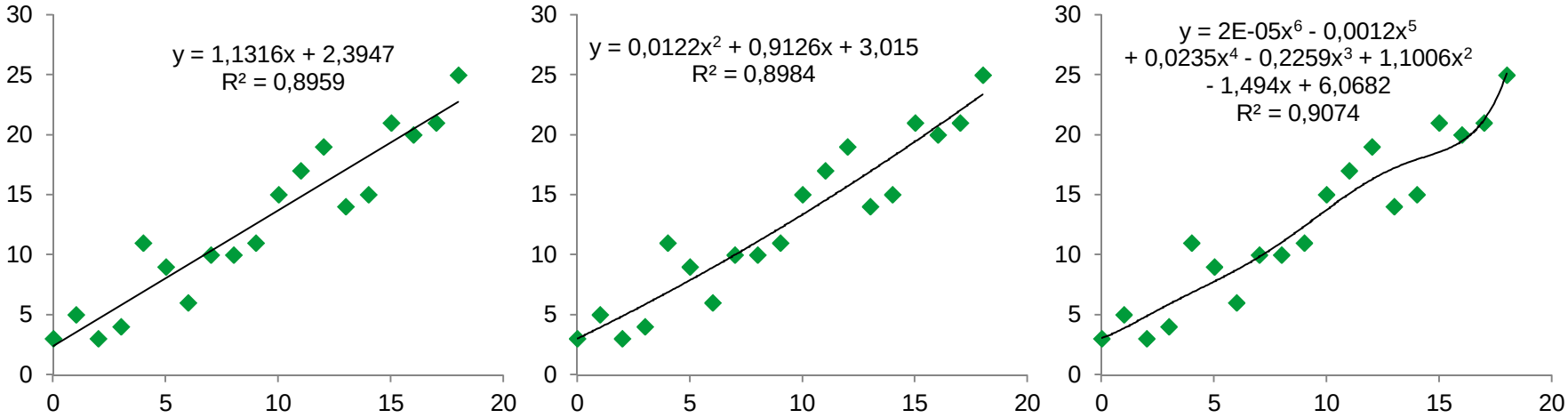
Example of the regression problem:

”What is best polynomial model to explain given input set”



- It is useful to have concept of hypothesis class. Here we assume that input can be explained by some polynomial function. Therefore, our class of hypothesis is all polynomial functions. Hypothesis class is a form of *inductive bias*.

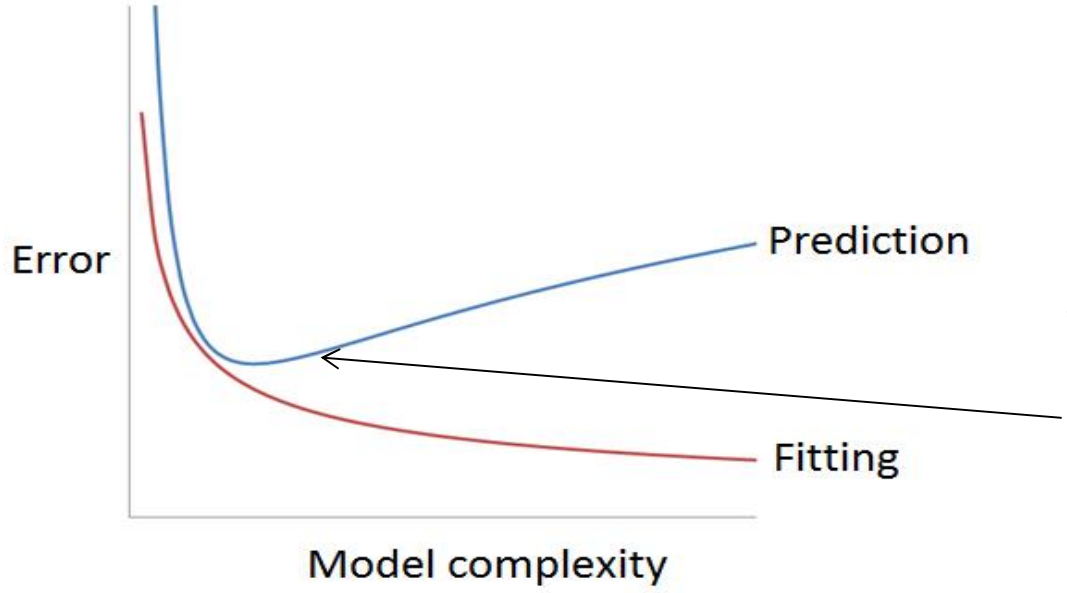
Machine Learning, Regression and Validation



- Different polynomials are called different hypothesis that are all within our class of hypothesis (polynomial functions)

Machine Learning, Regression and Validation

- The problem of generalization: how well our hypothesis will predict future examples that are not part of the training set?



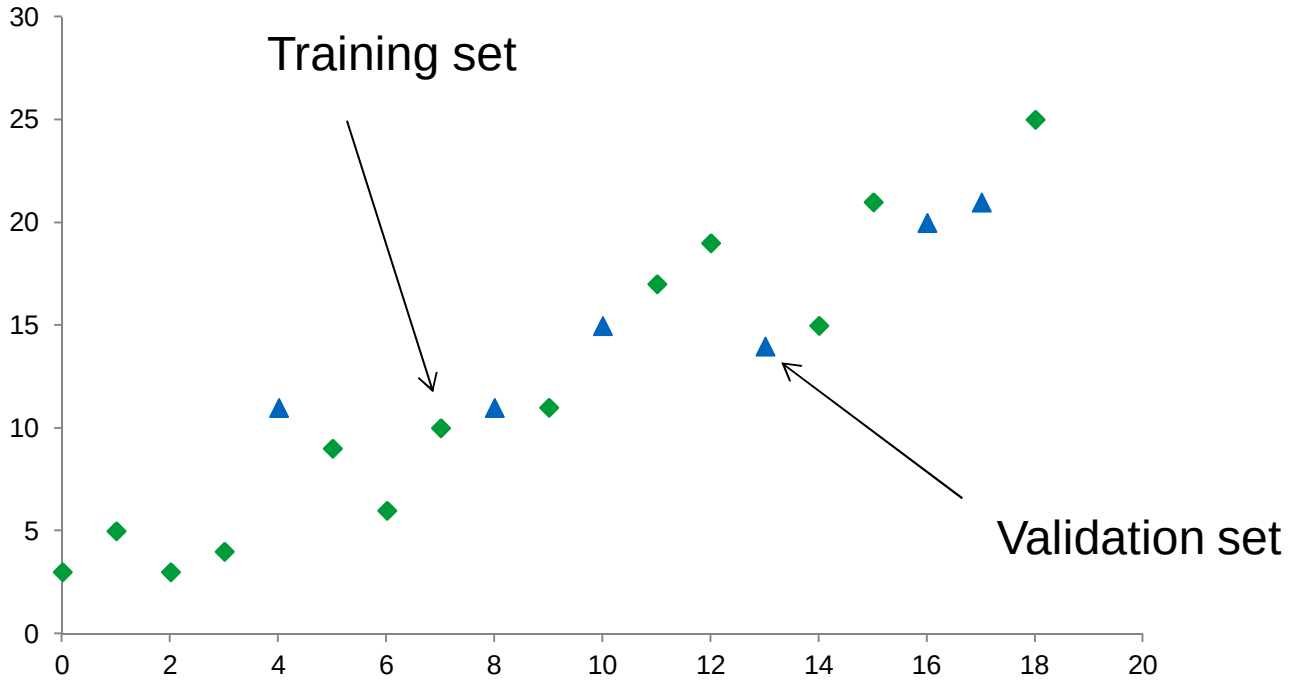
- The best model is characterized by minimal prediction error

- Prediction or generalization error can be evaluated by cross-validation or by repeated sub-sampling

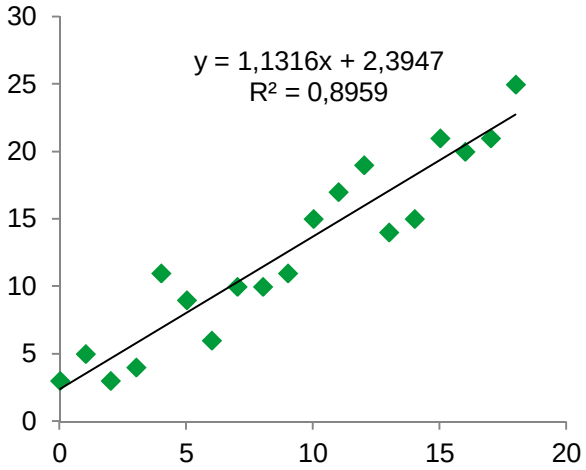
Machine Learning, Regression and Validation

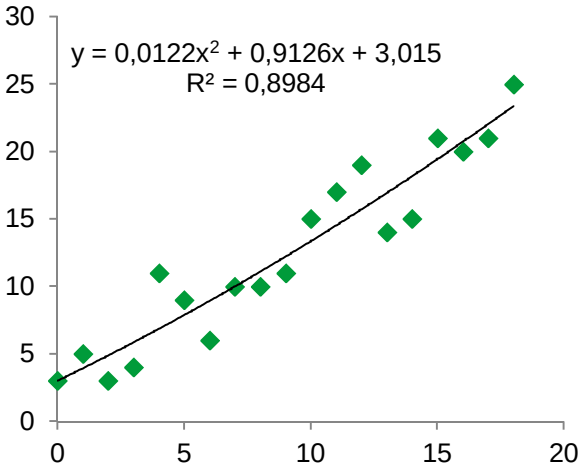
Example of regression problem:

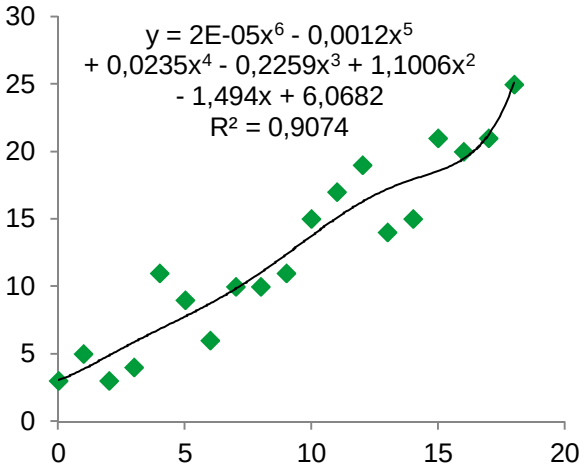
”What is best polynomial model to explain given input set”



Machine Learning, Regression and Validation

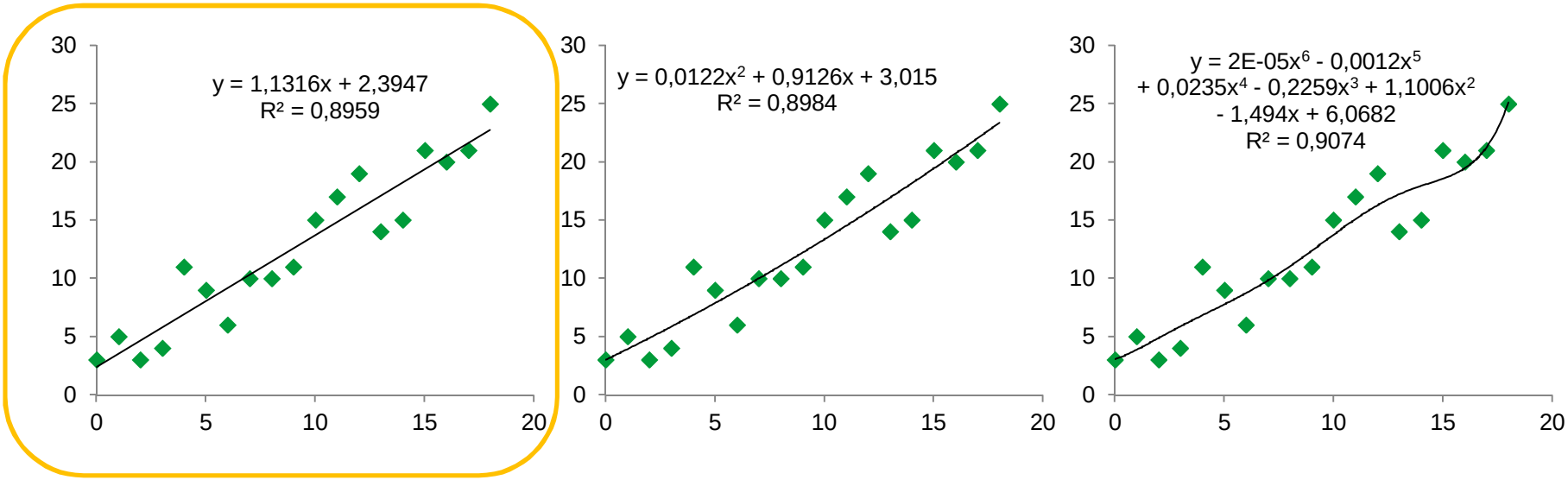






Split No.	1st order	2nd order	6th order
1	2.138	2.033	1.758
2	2.555	2.565	7.503
3	2.674	2.617	2.756
4	1.721	1.868	140.031
5	2.863	3.031	3.180
Mean validation error	2.39	2.42	31.05

Machine Learning, Regression and Validation



Split No.	1st order	2nd order	6th order
1	2.138	2.033	1.758
2	2.555	2.565	7.503
3	2.674	2.617	2.756
4	1.721	1.868	140.031
5	2.863	3.031	3.180
Mean validation error	2.39	2.42	31.05

However, expected error of our best model is not 2.39.

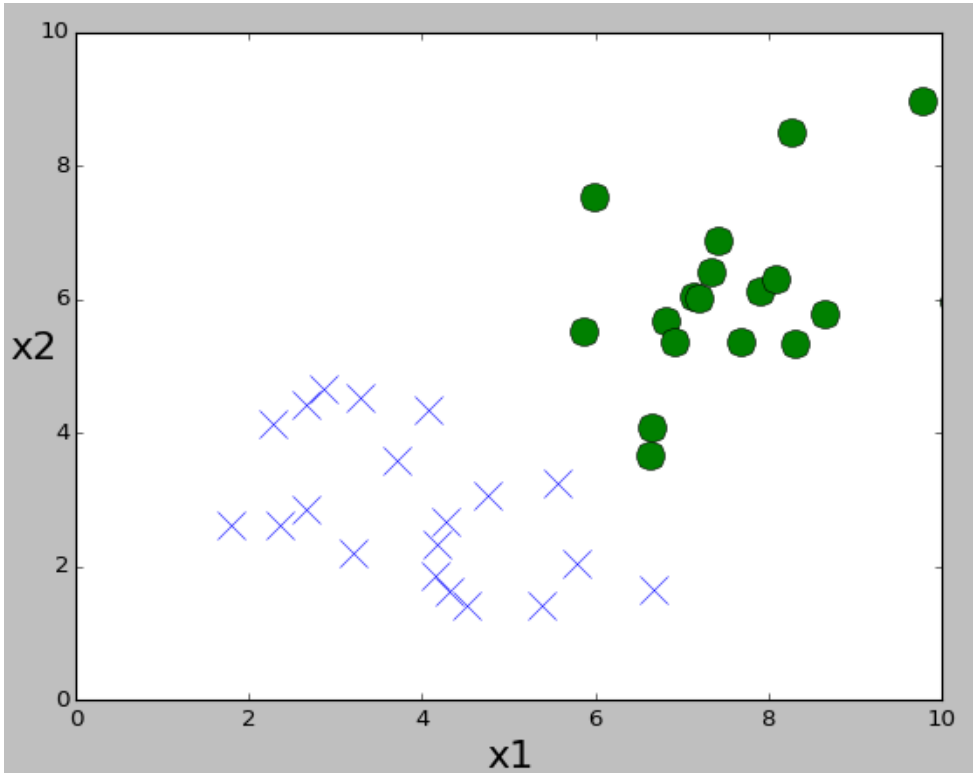
We would need a third set that contains examples that were not in original input set.

The third set is called the *test set* or sometimes *the publication set.*

Machine Learning, Classification

Example of classification problem:

”How to determine the class of an instance?”

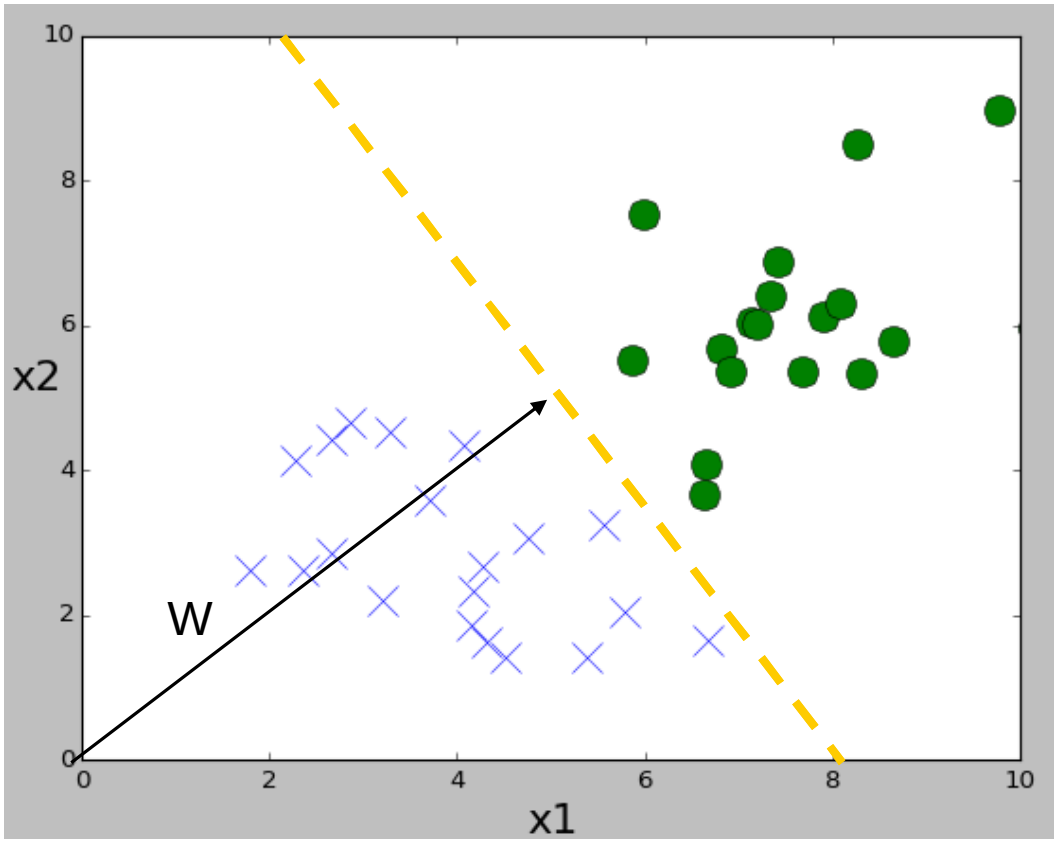


Machine Learning, Classification

Two dimensional problem (x_1 and x_2).

$$z = w^T x = \sum_i w_i x_i$$

gives projection of x onto w.



Machine Learning, Classification

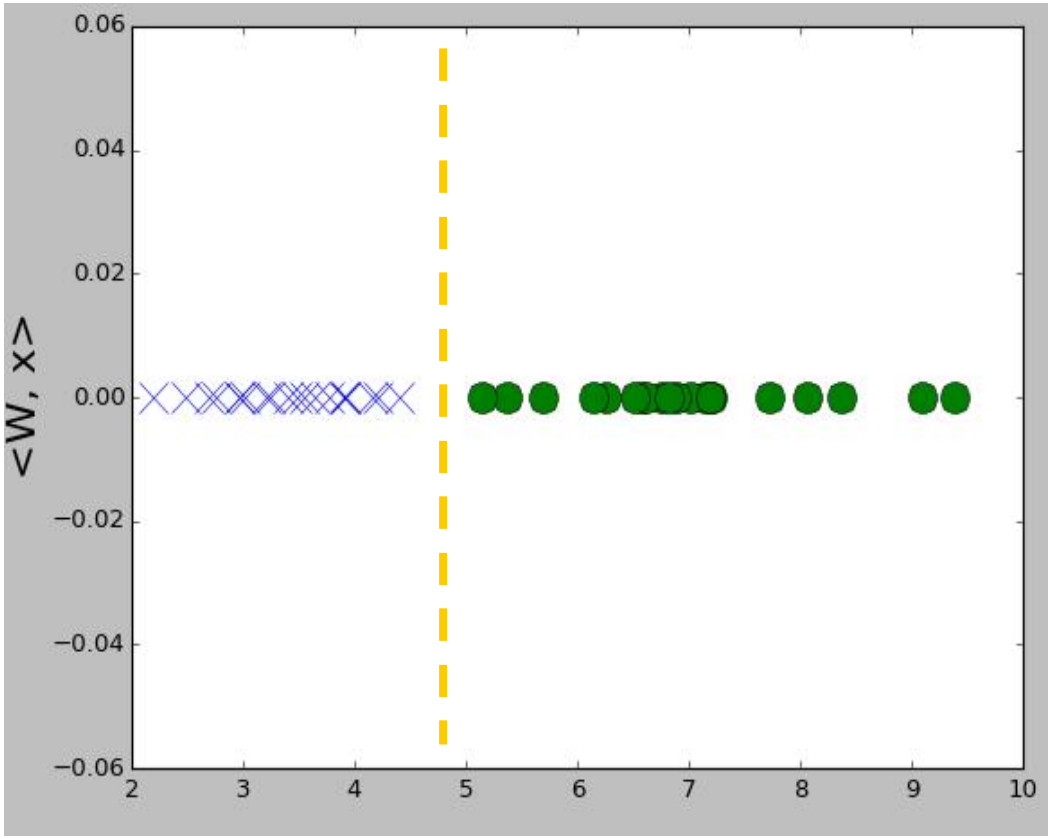
Two dimensional problem (x_1 and x_2).

$$z = w^T x = \sum_i w_i x_i$$

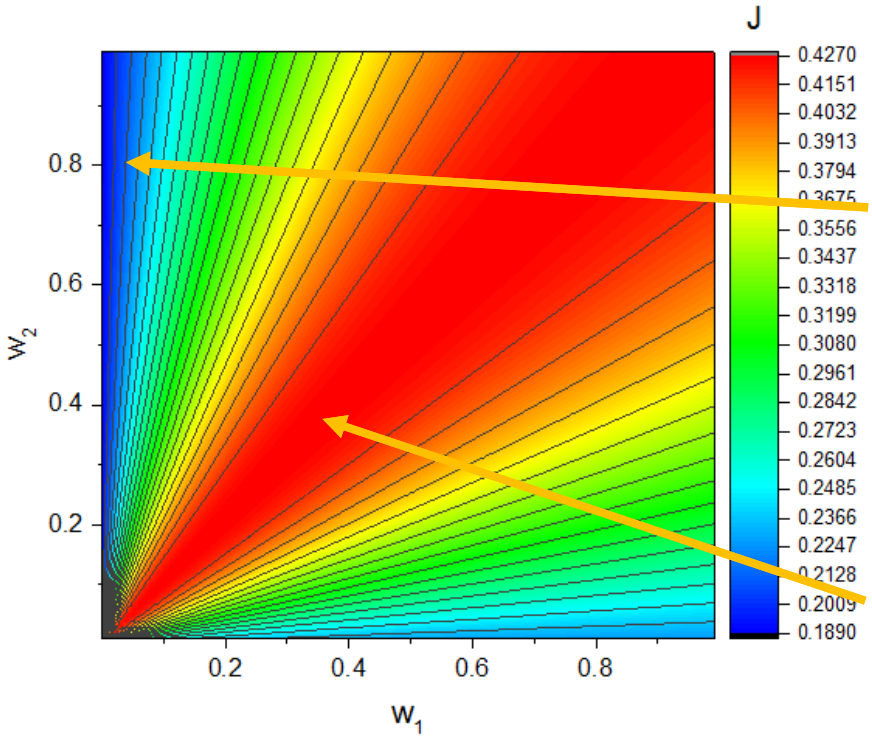
gives projection of x onto w.

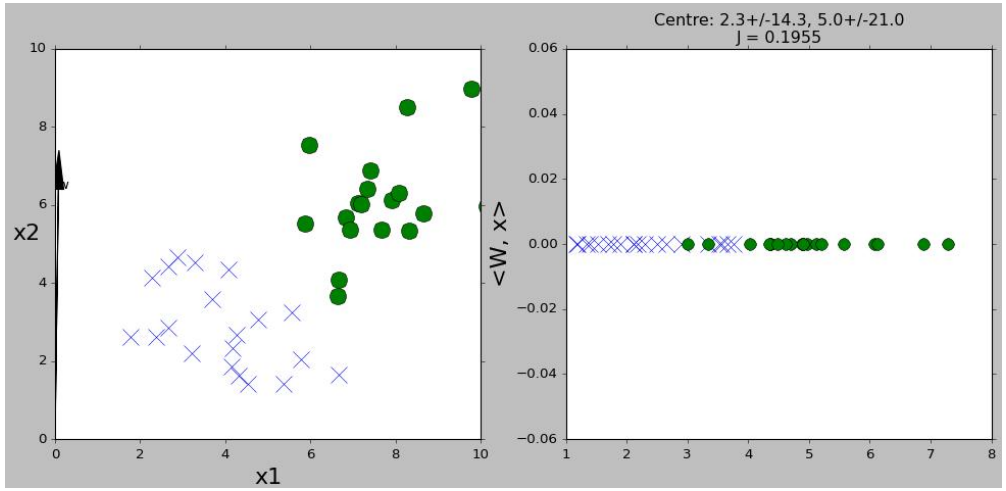
Now decision boundary (discriminant) is simply:

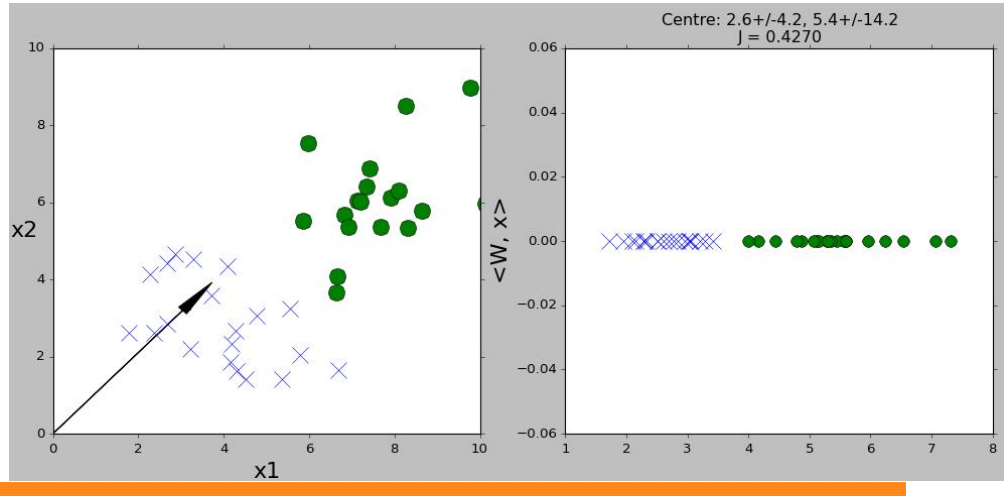
Class { +1 if z > 4.8
else: -1



Machine Learning, Classification





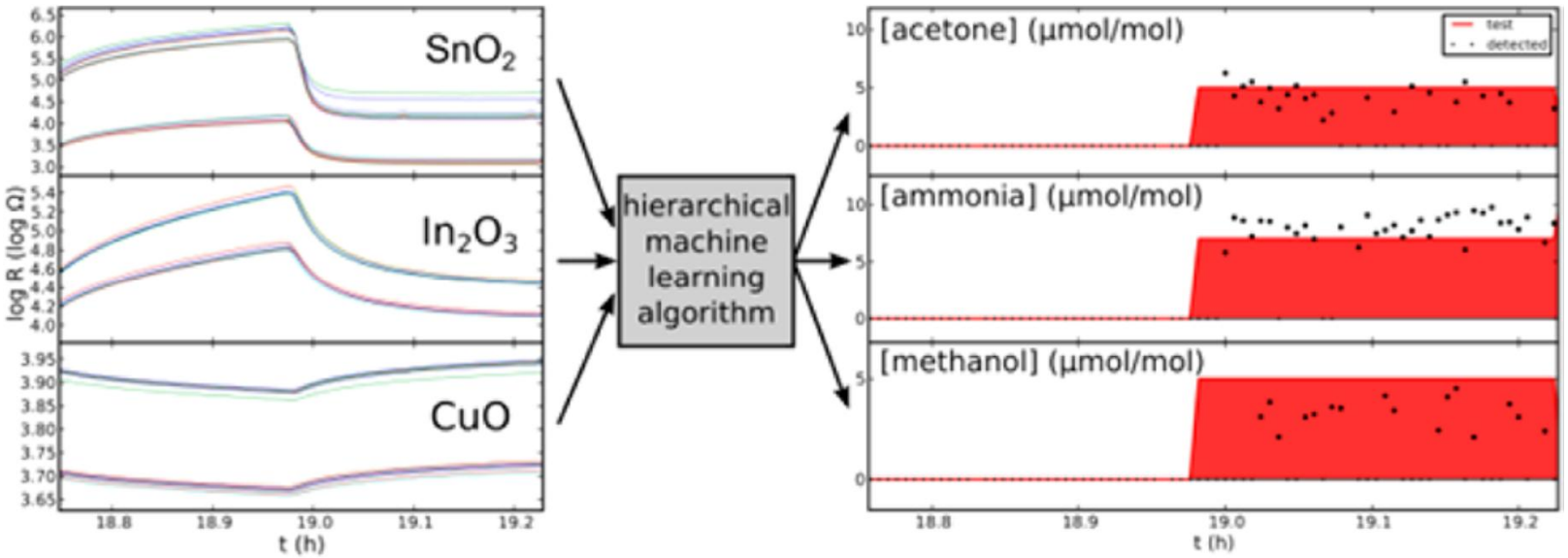


Use Fischer's linear discriminant as W:

$$\max_{W} J(W) = \frac{(m_1 - m_2)^2}{s_1^2 - s_2^2}$$

Machine Learning, Classification

The linear discriminant analysis used in detecting biomarkers from sensor data.[1]



[1] P. Rogers *et al.,* Machine Learning Applied to Chemical Analysis: Sensing Multiple Biomarkers in Simulated Breath Using a Temperature-Pulsed Electronic-Nose, *Analytical Chemistry*, **2012**.

Machine Learning, Clustering

Machine Learning Methods

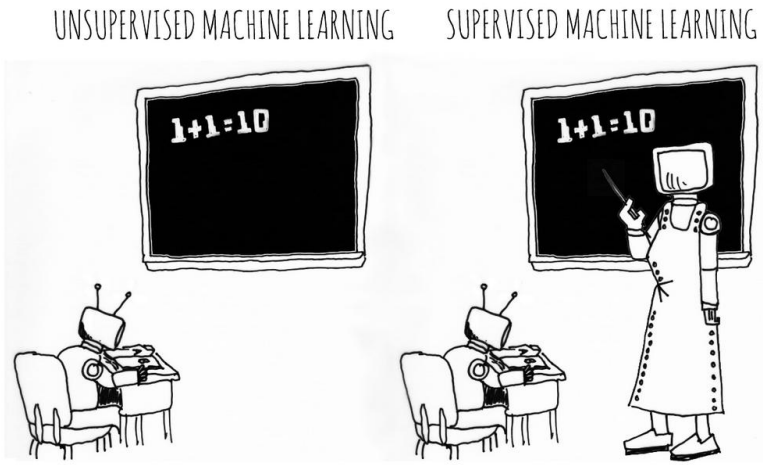
Supervised Learning
- The aim is to learn mapping from input to output.
- Output is known by a supervisor.

Unsupervised Learning
- The aim is to find ”something interesting”.
- Output is not known (known also as density estimation in statistics).

Reinforcement Learning
- Seemingly good solutions are encouraged to develop further.

Classification
- Outputs are discrete.

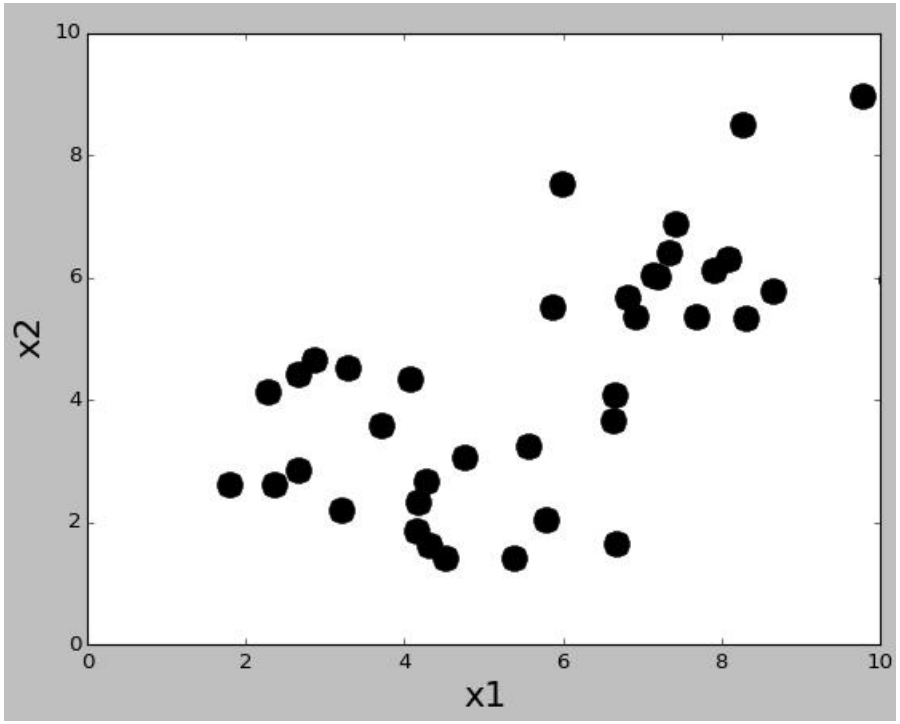
Regression
- Outputs are continuous.
- Typically noisy input.



Machine Learning, Clustering

An alternative approach: k-means clustering.

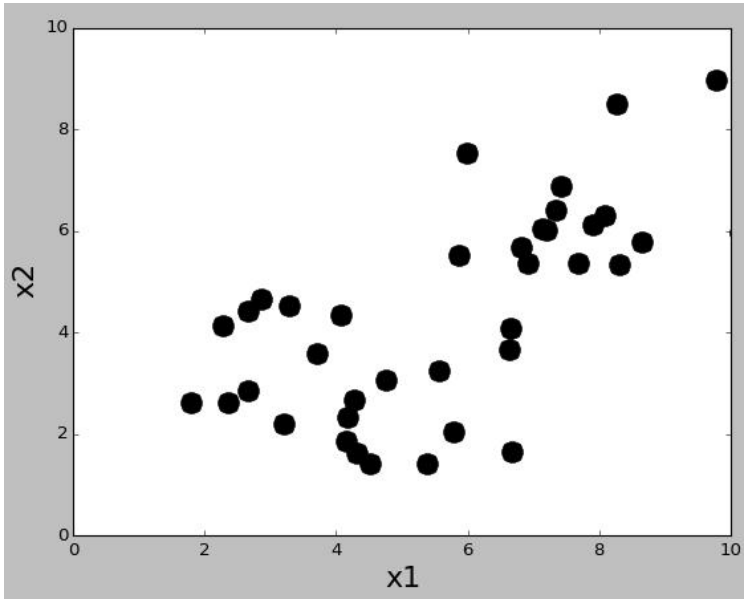
What is a cluster? How many clusters do you see?



Machine Learning, Clustering

- Informally:
  Cluster analysis or clustering is the task of assigning set of objects into groups so that the objects in the same group are more similar to each other than to those in other clusters.[1]
- In contrast to previous example, this is an unsupervised learning problem:

Previous classification problem was:
Training data: *{x, c}*
Task: predict *c* for new data vector *x*.
Clustering classification problem is:
Data *{x}*
Task: find some *c* for existing data.



[1] Wikipedia.

Machine Learning, Clustering

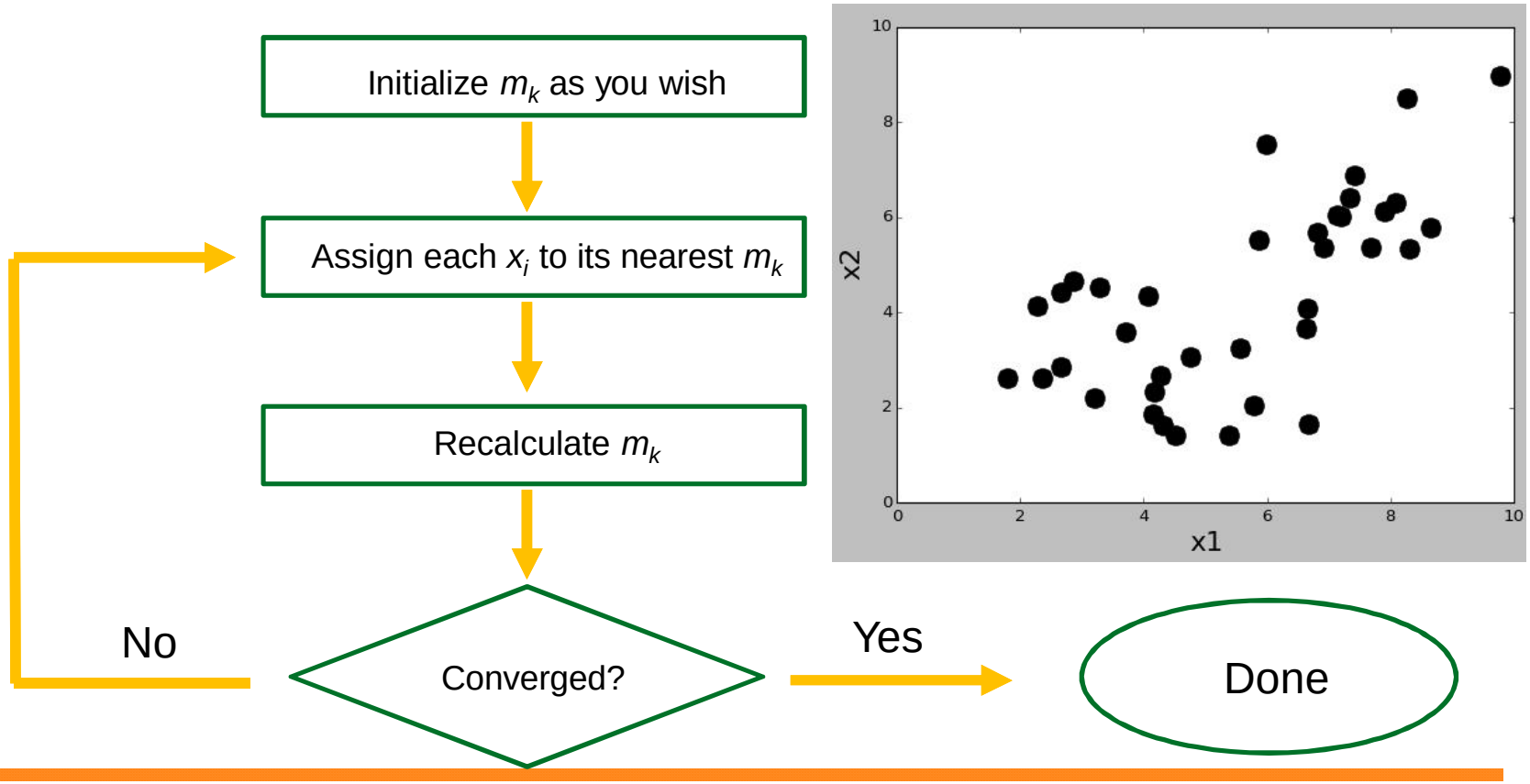
- How can we represent a “cluster”?
- A simple way is to use the mean of the data points.
- Error function to be minimized:

$$\varepsilon\left(\{m_k\}_{k=1}^{K}\middle|X\right) = \sum_{i=1}^{N} \min_{k} \left|\left|x_i - m_k\right|\right|^2$$

- That is: find mean vectors m_i such that distances of given datapoints from the mean vectors is minimized.
- This might sound easy, but finding the global minimum for K-means is NP-hard (≈ too expensive).

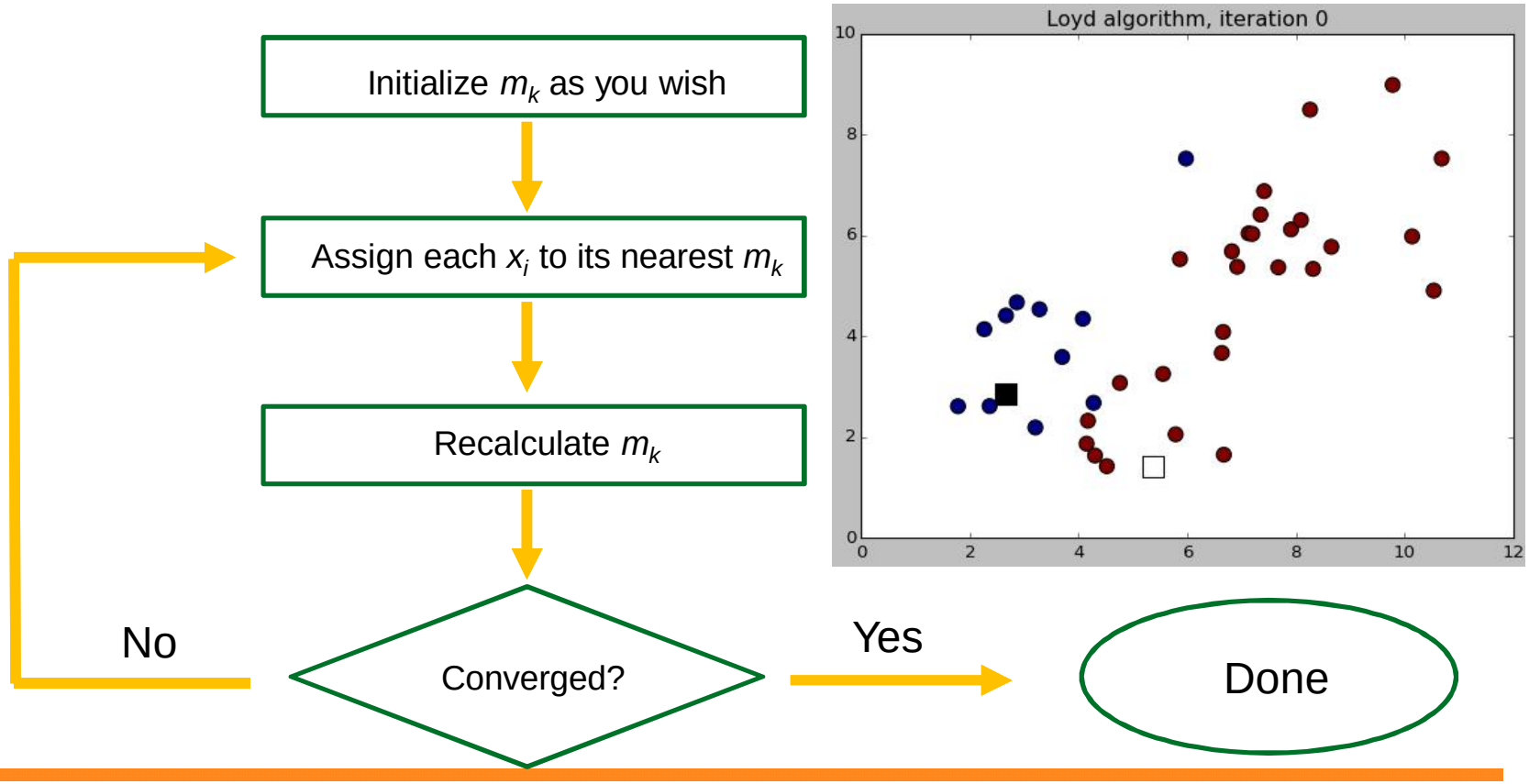
Machine Learning, Clustering

- Loyd's algorithm is the most famous method to find local minimum for the K-means cost function.



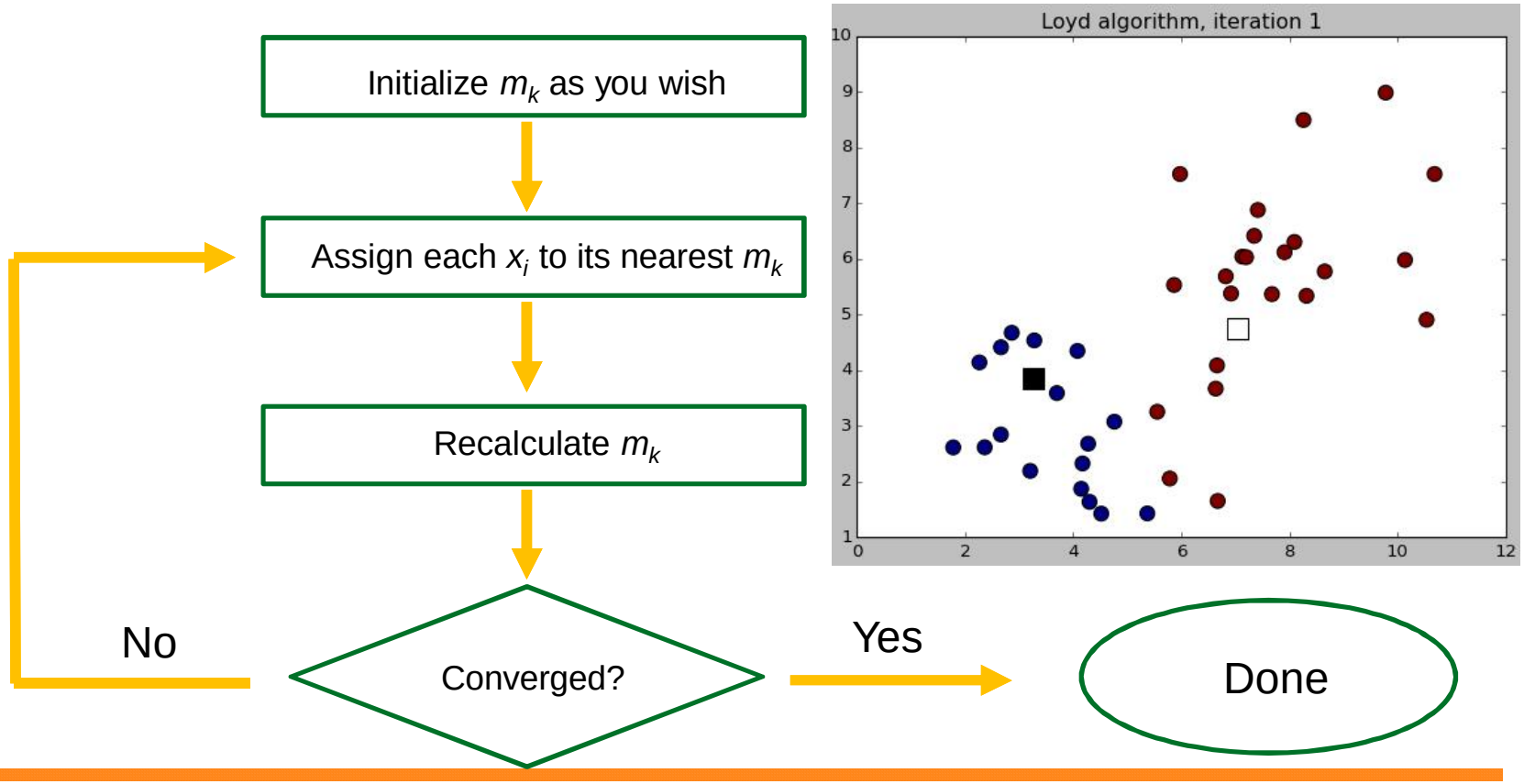
Machine Learning, Clustering

- Loyd's algorithm is the most famous method to find local minimum for the K-means cost function.



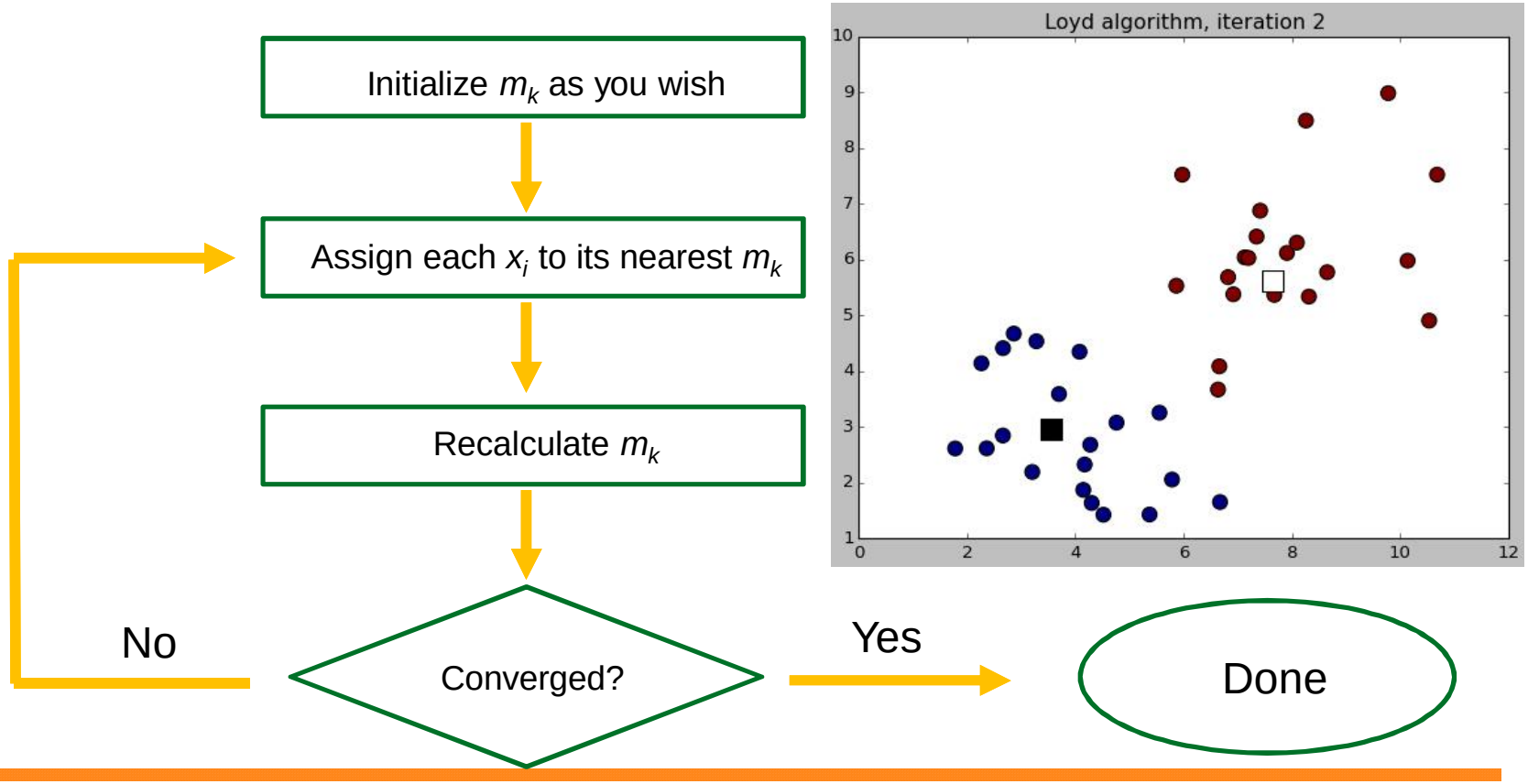
Machine Learning, Clustering

- Loyd's algorithm is the most famous method to find local minimum for the K-means cost function.



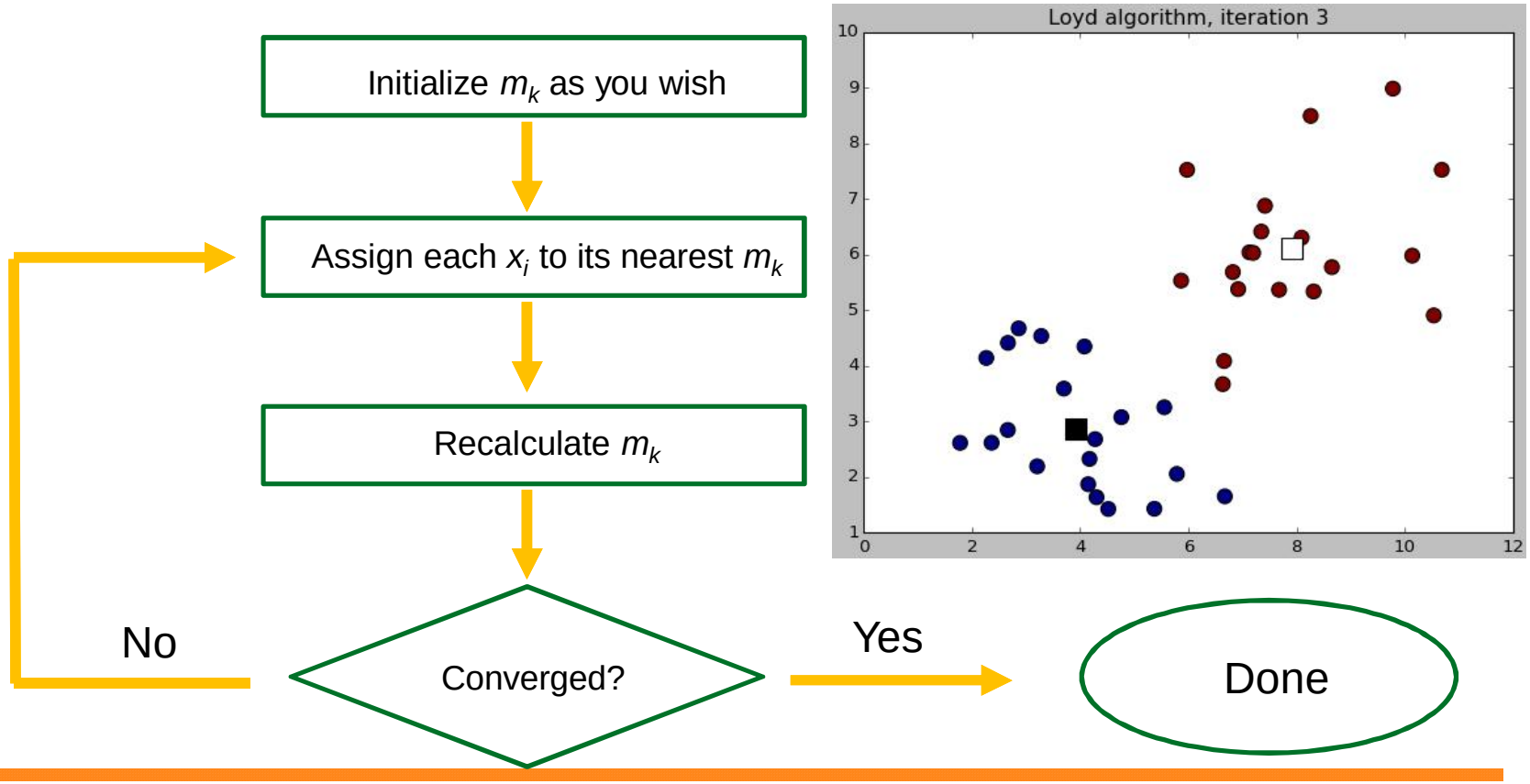
Machine Learning, Clustering

- Loyd's algorithm is the most famous method to find local minimum for the K-means cost function.



Machine Learning, Clustering

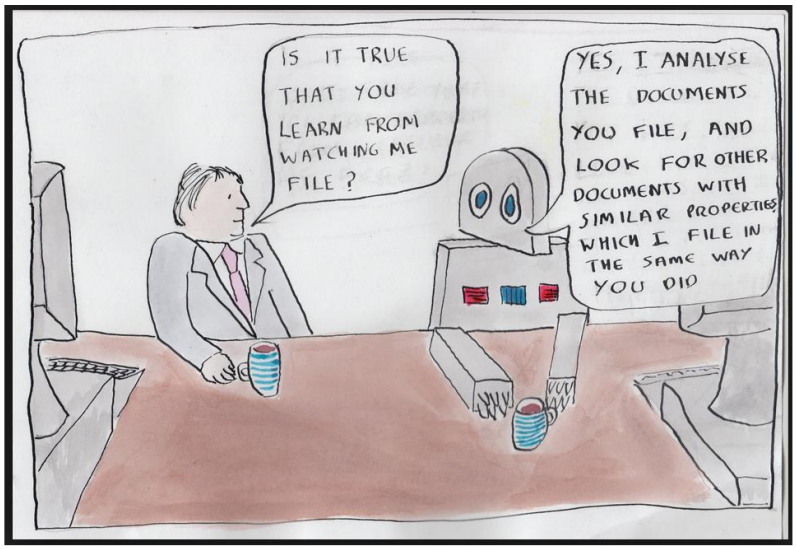
- Loyd's algorithm is the most famous method to find local minimum for the K-means cost function.



Machine Learning, Clustering

Some applications of clustering:

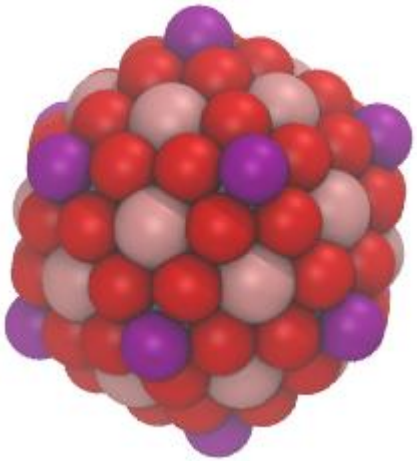
- Marketing (consumers into market segments).
- Biology (grouping of gene families).
- Sociology (finding communities in social networks).
- Climatology (finding weather regimes).
- Natural language processing (indentify topics in a corpus).
- Recommendation systems (e.g. predict user's preference based on preferences of other users in same cluster).
- …



Machine Learning, Coarse-Grained DFT Model

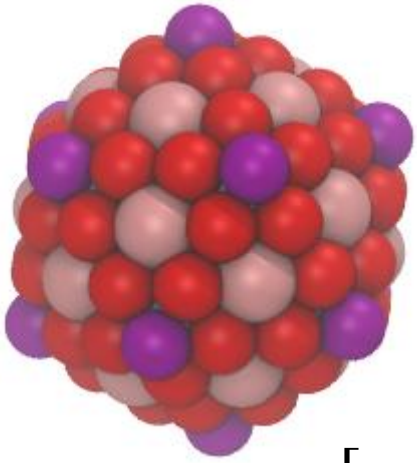
Example of a coarse-grained DFT modeling:

”How to predict ΔH_f from structural variables?”



Machine Learning, Coarse-Grained DFT Model

- Structural features are quantified and inductive bias is included.

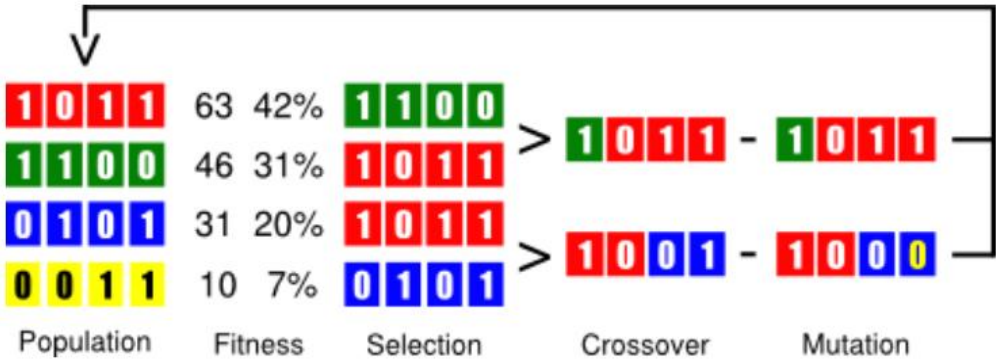


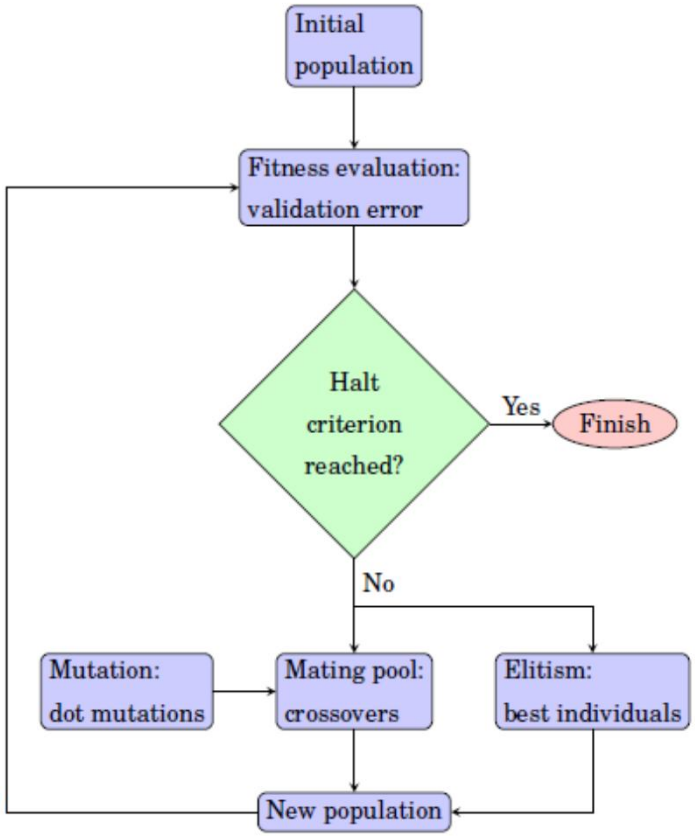
$$x_i = |0.9048, -0.667, -0.1429 \ldots|$$

$$H_f(\sigma) = V^0 + \sum_{\alpha} \left[V_{\alpha}^{(1,f)} \langle \sigma_{\alpha;i} \rangle^{(\sigma)} + \sum_{\alpha',f} V_{\alpha\alpha'}^{(2,f)} \langle \sigma_{\alpha;i}\sigma_{\alpha';j} \rangle^{(\sigma)} + \sum_{\alpha',\alpha'' f} V_{\alpha\alpha'\alpha''}^{(3,f)} \langle \sigma_{\alpha;i}\sigma_{\alpha';j}\sigma_{\alpha'';k} \rangle^{(\sigma)} + \sum_{\alpha',\alpha'',\alpha''' f} V_{\alpha\alpha'\alpha''\alpha'''}^{(4,f)} \langle \sigma_{\alpha;i}\sigma_{\alpha';j}\sigma_{\alpha'';k}\sigma_{\alpha''';l} \rangle^{(\sigma)} \right]$$

Machine Learning, Coarse-Grained DFT Model

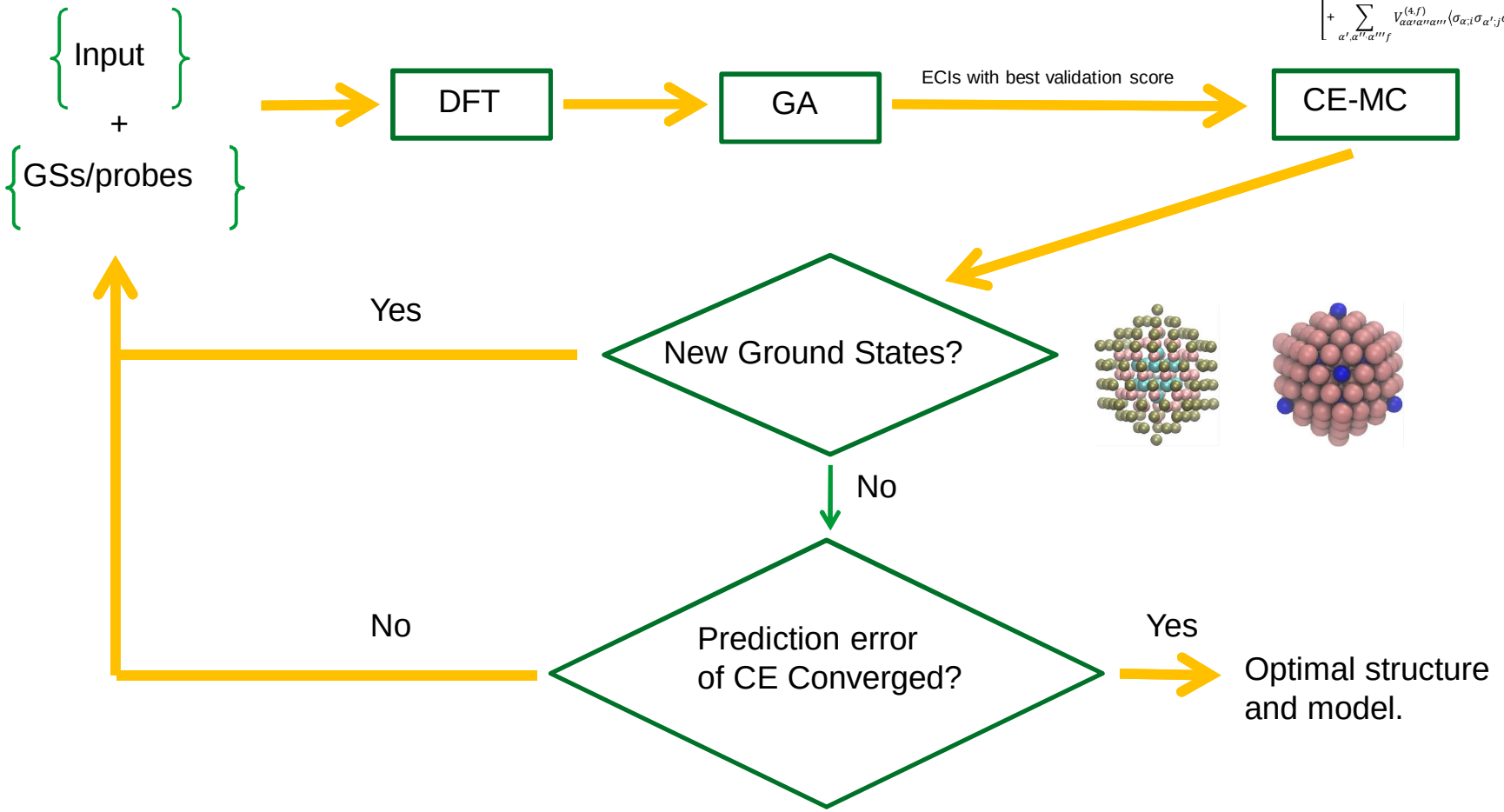
- As in the previous regression problem, there is limited amount of learning data.
- There must be at least one data point for each parameter. Overfitting is easy.
- An accurate parameters can be found by using genetic algorithms that mimic natural evolution.





Machine Learning, Coarse-Grained DFT Model

$$H_f(\sigma) = V^0 + \sum_{\alpha} \begin{bmatrix} V_{\alpha}^{(1,f)} \langle \sigma_{\alpha;i} \rangle^{(\sigma)} + \sum_{\alpha',f} V_{\alpha\alpha'}^{(2,f)} \langle \sigma_{\alpha;i} \sigma_{\alpha';j} \rangle^{(\sigma)} \\ + \sum_{\alpha',\alpha''f} V_{\alpha\alpha'\alpha''}^{(3,f)} \langle \sigma_{\alpha;i} \sigma_{\alpha';j} \sigma_{\alpha'';k} \rangle^{(\sigma)} \\ + \sum_{\alpha',\alpha''\alpha'''f} V_{\alpha\alpha'\alpha''\alpha'''}^{(4,f)} \langle \sigma_{\alpha;i} \sigma_{\alpha';j} \sigma_{\alpha'';k} \sigma_{\alpha''';l} \rangle^{(\sigma)} \end{bmatrix}$$



Machine Learning, Conclusion

Every chemist should know basic principles of machine learning.

Further info:
T-61.3050 Machine Learning: Basic Principles.
Ethem Alpaydin, Introduction to Machine Learning.