

Lecture 7

Machine Learning for Intelligent Systems

Introduction, Clustering, Classification, Regression, Evaluation

COMP 474/6741, Winter 2021

Machine Learning
Primer

History
ML Types
Process

Clustering Documents

Motivation
k-Means Clustering
Application Example

Classifications &
Predictions

Introduction
Classification with kNN
Regression with kNN

Machine Learning
Evaluation

Evaluation Methodology
Evaluation Metrics
Error Analysis
Overfitting
Underfitting
Cross-Validation

Notes and Further
Reading

René Witte
Department of Computer Science
and Software Engineering
Concordia University

Outline

1 Machine Learning Primer

- History
- ML Types
- Process

2 Clustering Documents

- Motivation
- k-Means Clustering
- Application Example

3 Classifications & Predictions

- Introduction
- Classification with kNN
- Regression with kNN

4 Machine Learning Evaluation

- Evaluation Methodology
- Evaluation Metrics
- Error Analysis
- Overfitting
- Underfitting
- Cross-Validation

5 Notes and Further Reading

René Witte



Machine Learning
Primer

History
ML Types
Process

Clustering Documents
Motivation
k-Means Clustering
Application Example

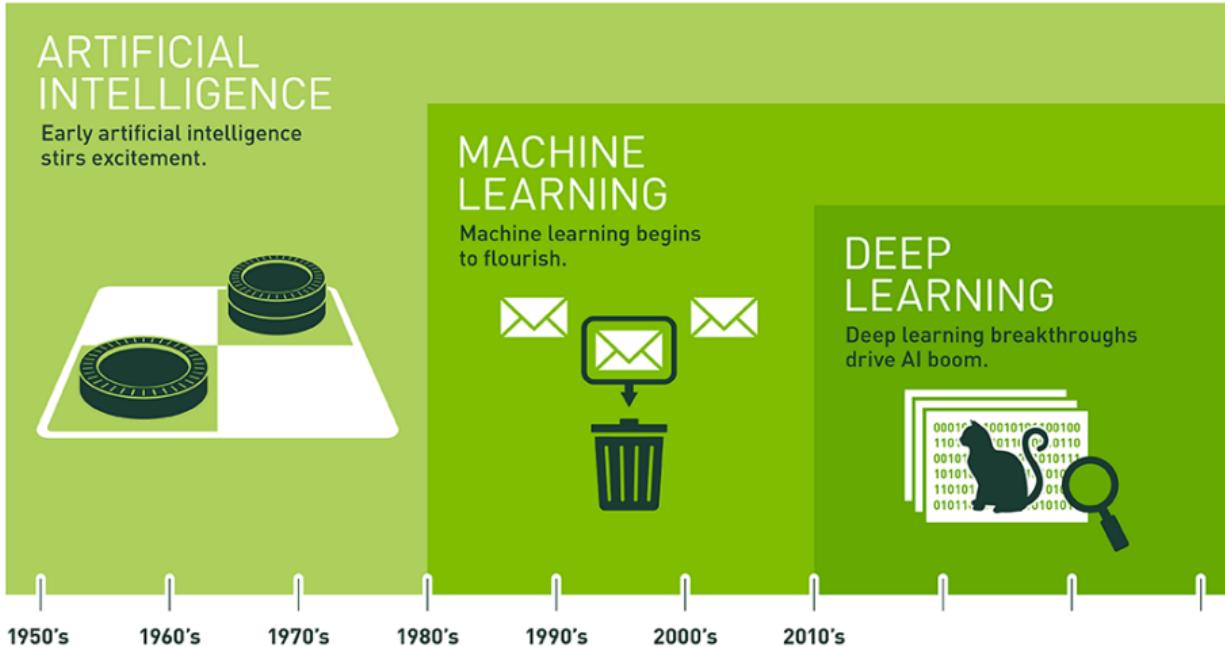
Classifications &
Predictions

Introduction
Classification with kNN
Regression with kNN

Machine Learning
Evaluation

Evaluation Methodology
Evaluation Metrics
Error Analysis
Overfitting
Underfitting
Cross-Validation

Notes and Further
Reading



Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

Machine Learning Primer

History

ML Types

Process

Clustering Documents

Motivation

k-Means Clustering

Application Example

Classifications & Predictions

Introduction

Classification with kNN

Regression with kNN

Machine Learning Evaluation

Evaluation Methodology

Evaluation Metrics

Error Analysis

Overfitting

Underfitting

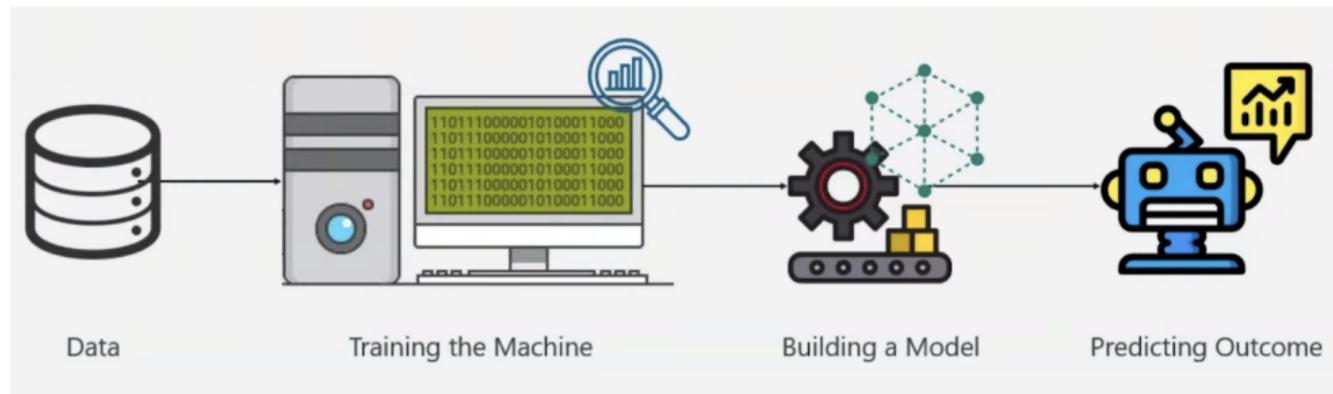
Cross-Validation

Notes and Further Reading

Learn from experience

In 1959, Arthur Samuel first proposed the concept
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."



Machine Learning
Primer

History

ML Types

Process

Clustering Documents

Motivation

k-Means Clustering

Application Example

Classifications &
Predictions

Introduction

Classification with kNN

Regression with kNN

Machine Learning
Evaluation

Evaluation Methodology

Evaluation Metrics

Error Analysis

Overfitting

Underfitting

Cross-Validation

Notes and Further
Reading

Inference

Process of deriving new facts from a set of premises

Types of logical inference

- ① Deduction
- ② Abduction
- ③ Induction

Machine Learning Primer

History

ML Types

Process

Clustering Documents

Motivation

k-Means Clustering

Application Example

Classifications & Predictions

Introduction

Classification with kNN

Regression with kNN

Machine Learning Evaluation

Evaluation Methodology

Evaluation Metrics

Error Analysis

Overfitting

Underfitting

Cross-Validation

Notes and Further Reading

aka Natural Deduction

- Conclusion follows necessarily from the premises.
- From $A \Rightarrow B$ and **A**, we conclude that **B**
- We conclude from the general case to a specific example of the general case
- Example:
 - ① All men are mortal.
 - ② Socrates is a man.
 - ③ from ① \wedge ② \Rightarrow Socrates is mortal.
- Our subclass inference in RDFS also falls into this category.

Machine Learning
Primer

History

ML Types

Process

Clustering Documents

Motivation

k-Means Clustering

Application Example

Classifications &
Predictions

Introduction

Classification with kNN

Regression with kNN

Machine Learning
Evaluation

Evaluation Methodology

Evaluation Metrics

Error Analysis

Overfitting

Underfitting

Cross-Validation

Notes and Further
Reading

Abductive Reasoning

- Conclusion is one hypothetical (most probable) explanation for the premises
- From $A \Rightarrow B$ and B , we conclude A
- Example:
 - 1 Drunk people do not walk straight.
 - 2 John does not walk straight.
 - 3 from 1 \wedge 2 \Rightarrow John is drunk.
- Not sound... but may be most likely explanation for B
- Used in medicine...
 - 1 in reality: disease \Rightarrow symptoms
 - 2 patient complains about some symptoms... doctor concludes a disease

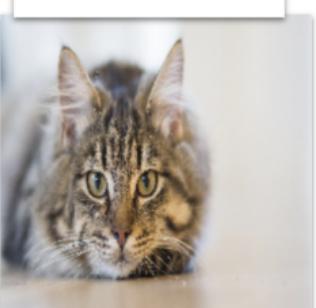
Inductive Reasoning

- Conclusion about all members of a class from the examination of only a few member of the class.
- From $A \wedge C \Rightarrow B$ and $A \wedge D \Rightarrow B$, we conclude $A \Rightarrow B$
- We construct a general explanation based on specific cases
- Example:
 - All CS students in COMP 474 are smart.
 - All CS students on vacation are smart.
 - from ① \wedge ② \Rightarrow All CS students are smart.
- Not sound
- But, can be seen as hypothesis construction or generalisation

Learning from examples

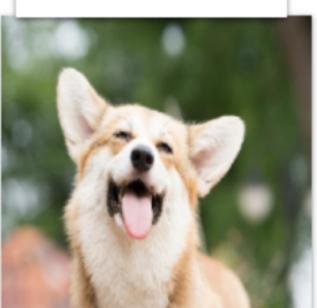
- Most work in ML
- Examples are given (positive and/or negative) to train a system in a classification (or regression) task
- Extrapolate from the training set to make accurate predictions about future examples
- Given a new instance X you have never seen, you must find an estimate of the function $f(X)$ where $f(X)$ is the desired output

100% cat



```
print("""cat: {
    np.round(model.predict(cat),2)
}""")
cat: {[1. 0. 0.]}
```

97% dog



```
print("""dog: {
    np.round(model.predict(dog),2)
}""")
dog: {[0.02 0.97 0.01]}
```

14% dog
85% Elon Musk



```
print("""elon: {
    np.round(model.predict(elon_with_disguise),2)
}""")
elon: {[0. 0.14 0.85]}
```

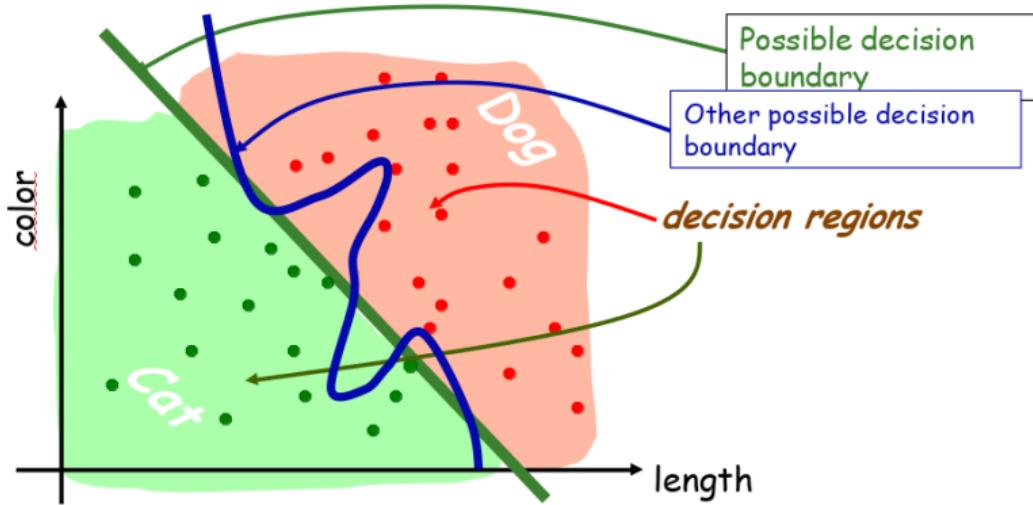
100% Elon Musk



```
print("""elon: {
    np.round(model.predict(elon_without_disguise),2)
}""")
elon: {[0. 0. 1.]}
```

Example

- Given pairs $(X, f(X))$ (the training set – the data points)
- Find a function f that fits the training set well
- So that given a new X , you can predict its $f(X)$ value



Note: choosing one function over another beyond just looking at the training set is called **inductive bias** (eg. prefer “smoother” functions)

Feature Vectors

- Input data are represented by a **vector of features**, X
- Each vector X is a list of (attribute, value) pairs.
- Ex: $X = [\text{nose:big}, \text{teeth:big}, \text{eyes:big}, \text{moustache:no}]$
- The number of attributes is fixed (positive, finite)
- Each attribute has a fixed, finite number of possible values
- Each example can be interpreted as a point in a n -dimensional feature space, where n is the number of attributes (features)

Machine Learning
Primer

History

ML Types

Process

Clustering Documents

Motivation

k-Means Clustering

Application Example

Classifications &
Predictions

Introduction

Classification with kNN

Regression with kNN

Machine Learning
Evaluation

Evaluation Methodology

Evaluation Metrics

Error Analysis

Overfitting

Underfitting

Cross-Validation

Notes and Further
Reading

Probabilistic Methods

- e.g., Naïve Bayes Classifier

Decision Trees

- Use only discriminating features as questions in a big *if-then-else* tree

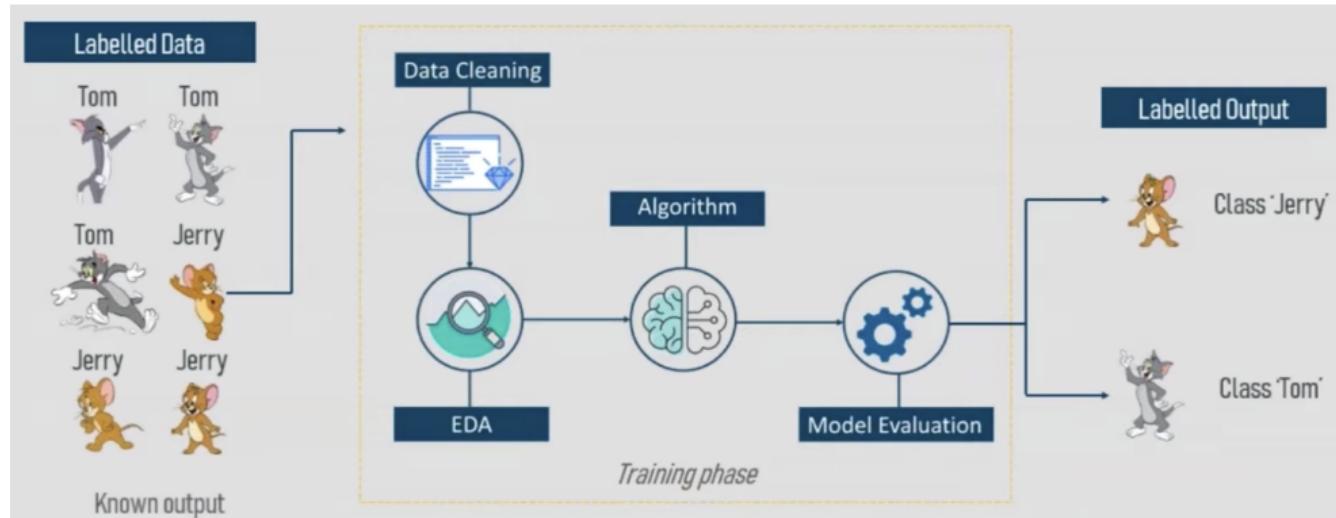
Neural Networks

- Also called parallel distributed processing or connectionist systems
- Intelligence arise from having a large number of simple computational units

NB: Deep Learning \approx Neural Networks “on steroids”

Supervised Learning

René Witte



Labeled Data

In **Supervised Learning**, we train a system using data with known labels.

Machine Learning Primer

History

ML Types

Process

Clustering Documents

Motivation

k-Means Clustering

Application Example

Classifications & Predictions

Introduction

Classification with kNN

Regression with kNN

Machine Learning Evaluation

Evaluation Methodology

Evaluation Metrics

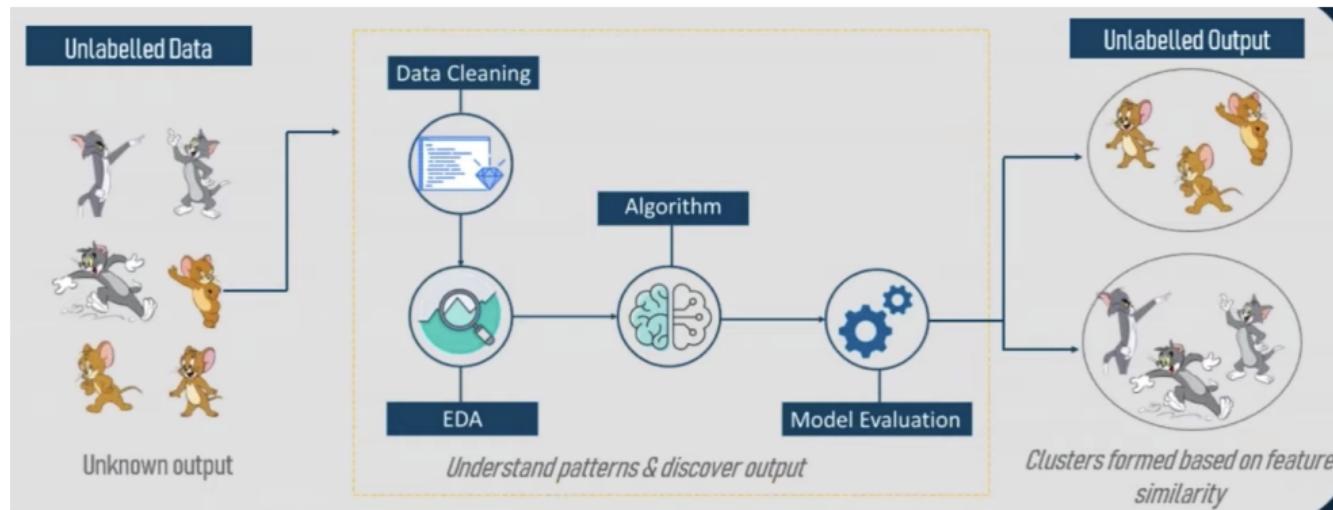
Error Analysis

Overfitting

Underfitting

Cross-Validation

Notes and Further Reading



Unlabeled Data

In [Unsupervised Learning](#), we have only unlabeled data and train a system without guidance from an expected output.

Machine Learning
Primer

History

ML Types

Process

Clustering Documents

Motivation

k-Means Clustering

Application Example

Classifications &
Predictions

Introduction

Classification with kNN

Regression with kNN

Machine Learning
Evaluation

Evaluation Methodology

Evaluation Metrics

Error Analysis

Overfitting

Underfitting

Cross-Validation

Notes and Further
Reading

Reinforcement Learning

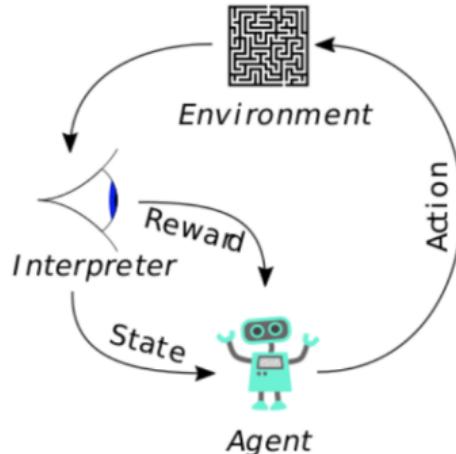
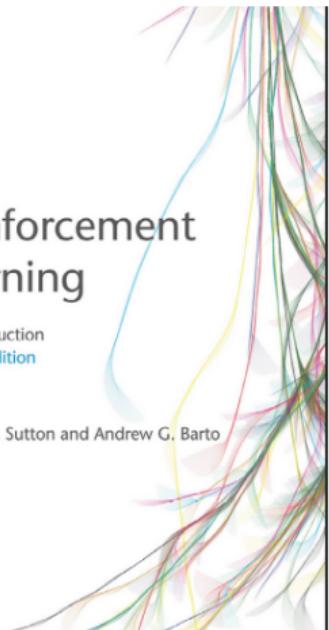
René Witte



Reinforcement Learning

An Introduction
second edition

Richard S. Sutton and Andrew G. Barto



The typical 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.

Machine Learning Primer

History

ML Types

Process

Clustering Documents

Motivation

k-Means Clustering

Application Example

Classifications & Predictions

Introduction

Classification with kNN

Regression with kNN

Machine Learning Evaluation

Evaluation Methodology

Evaluation Metrics

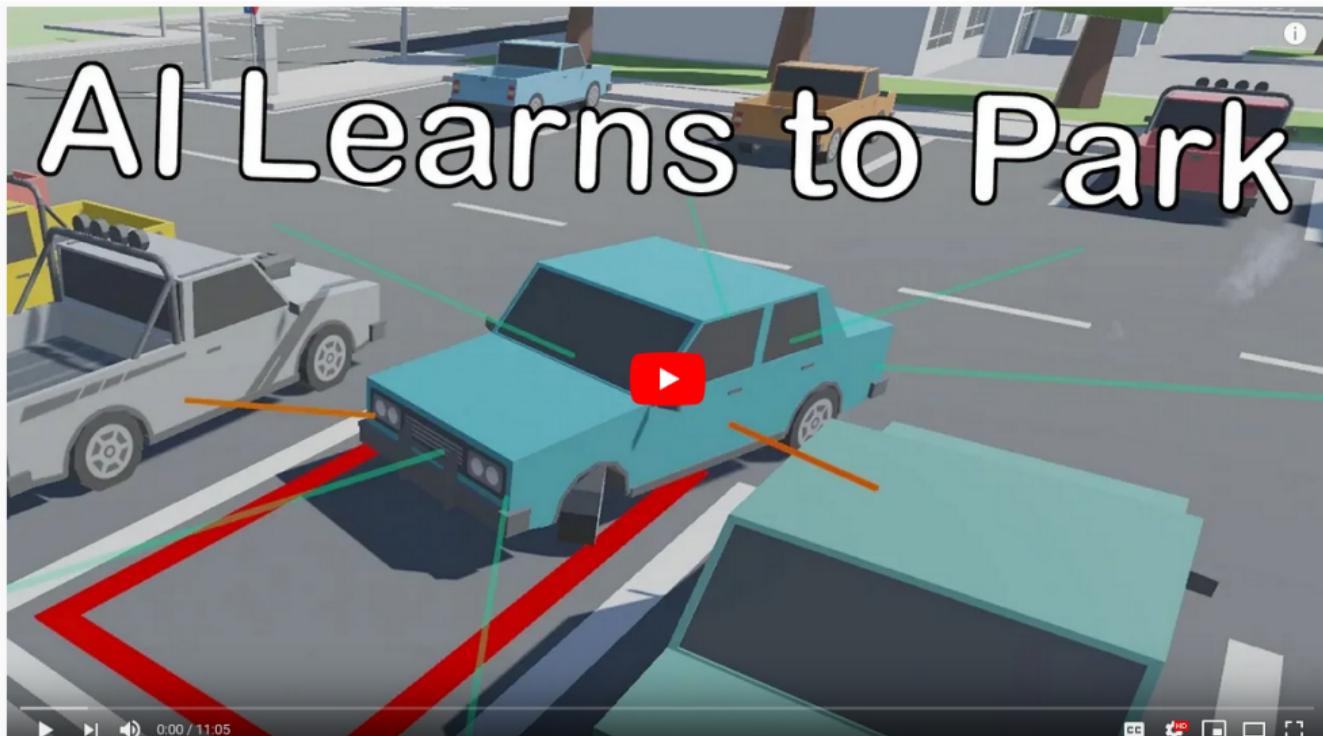
Error Analysis

Overfitting

Underfitting

Cross-Validation

Notes and Further Reading



#ArtificialIntelligence #MachineLearning #ReinforcementLearning

AI Learns to Park - Deep Reinforcement Learning

https://www.youtube.com/watch?v=VMp6pq6_QjI

Machine Learning
Primer

History

ML Types

Process

Clustering Documents

Motivation

k-Means Clustering

Application Example

Classifications &
Predictions

Introduction

Classification with kNN

Regression with kNN

Machine Learning
Evaluation

Evaluation Methodology

Evaluation Metrics

Error Analysis

Overfitting

Underfitting

Cross-Validation

Notes and Further
Reading

Machine Learning Categories

René Witte



| | Supervised Learning | Unsupervised Learning | Reinforcement Learning |
|---------------------------|--|---|---|
| Definition | The machine learns by using labelled data | The machine is trained on unlabeled data without any guidance | An agent interacts with its environment by producing actions & discovers errors and rewards |
| Types of problems | Regression & Classification | Association & Clustering | Reward based |
| Type of data | Labelled data | Unlabelled data | No pre-defined data |
| Training | External supervision | No supervision | No supervision |
| Approach | Map labelled input to known output | Understand patterns and discover output | Follow trial and error method |
| Popular Algorithms | Linear Regression, Logistic Regression, KNN, etc | K-means, C-means, etc | Q-learning, etc |

Machine Learning Primer

History

ML Types

Process

Clustering Documents

Motivation

k-Means Clustering

Application Example

Classifications & Predictions

Introduction

Classification with kNN

Regression with kNN

Machine Learning Evaluation

Evaluation Methodology

Evaluation Metrics

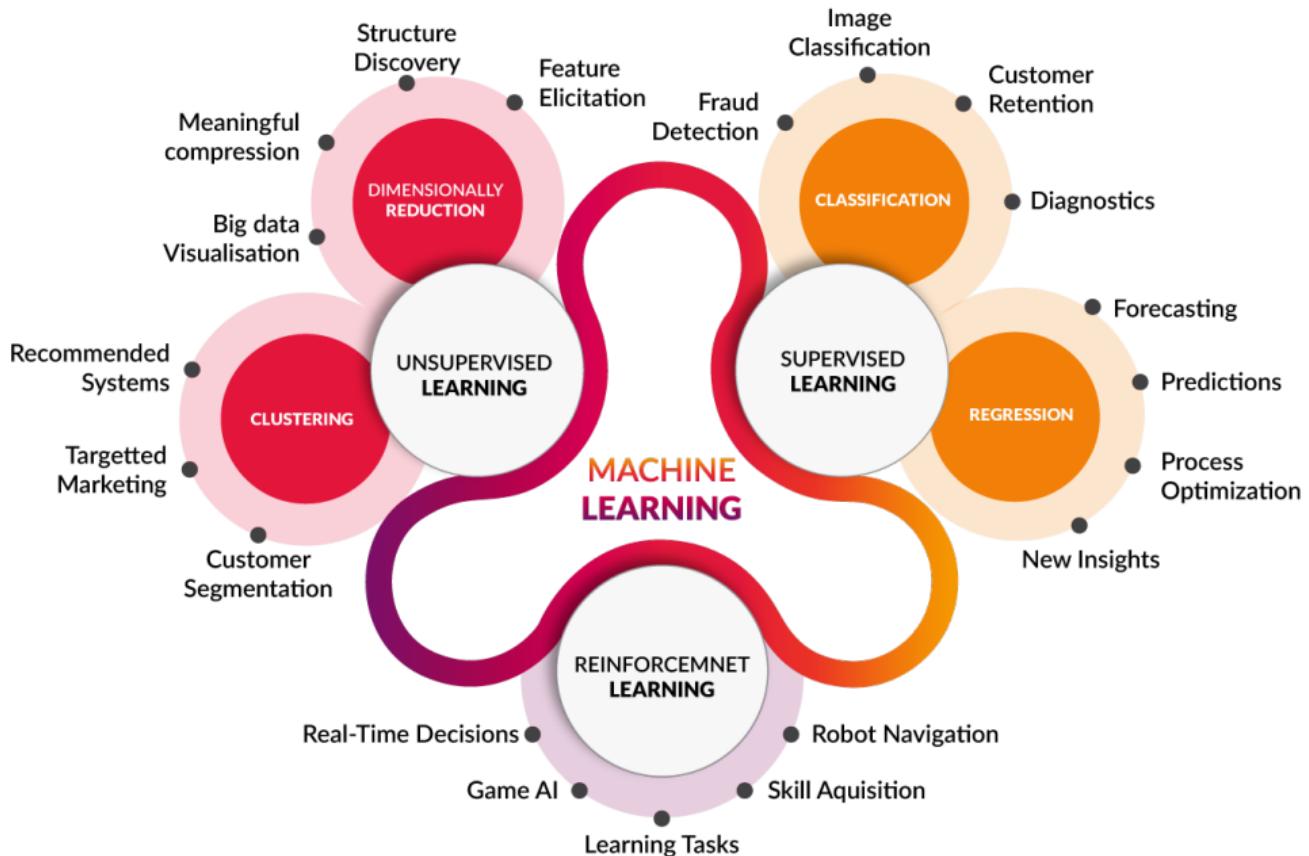
Error Analysis

Overfitting

Underfitting

Cross-Validation

Notes and Further Reading



Machine Learning
Primer

History

ML Types

Process

Clustering Documents

Motivation

k-Means Clustering

Application Example

Classifications &
Predictions

Introduction

Classification with kNN

Regression with kNN

Machine Learning
Evaluation

Evaluation Methodology

Evaluation Metrics

Error Analysis

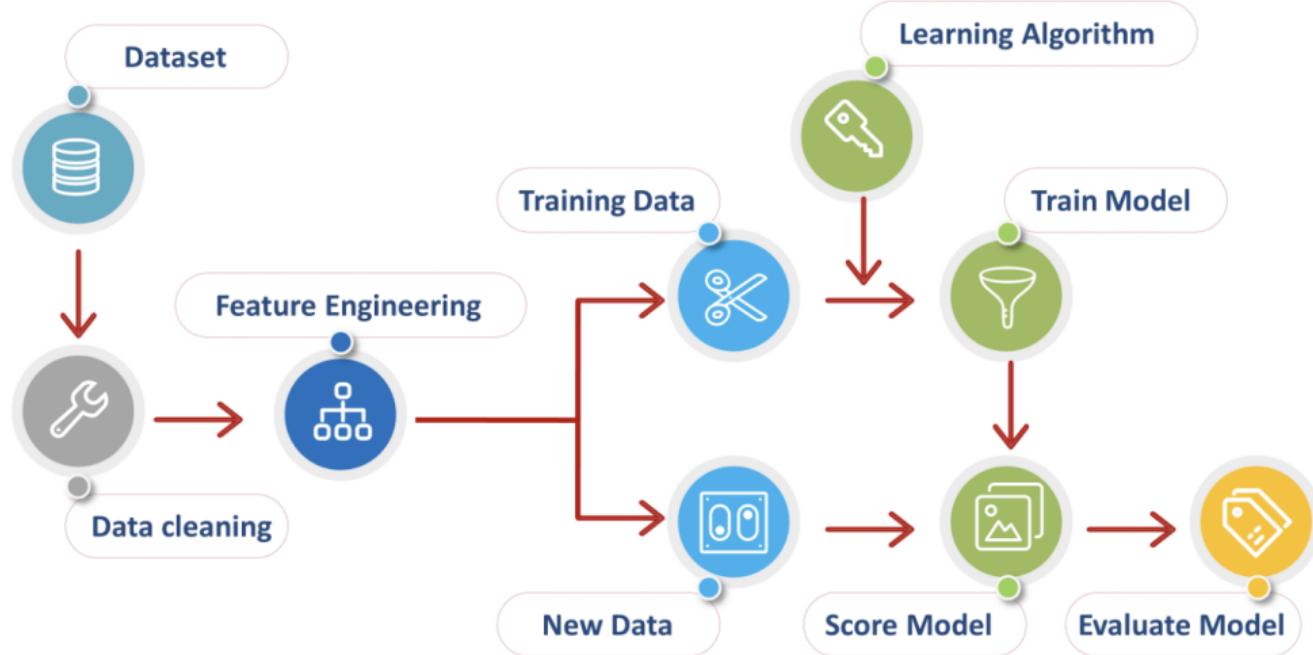
Overfitting

Underfitting

Cross-Validation

Notes and Further
Reading

General machine learning process



→ Worksheet #6: Task 1

1 Machine Learning Primer

Machine Learning
Primer
History
ML Types
Process

2 Clustering Documents

Motivation
k-Means Clustering
Application Example

Clustering Documents
Motivation
k-Means Clustering
Application Example

3 Classifications & Predictions

Classifications &
Predictions
Introduction
Classification with kNN
Regression with kNN

4 Machine Learning Evaluation

Machine Learning
Evaluation
Evaluation Methodology
Evaluation Metrics
Error Analysis
Overfitting
Underfitting
Cross-Validation

5 Notes and Further Reading

Notes and Further
Reading

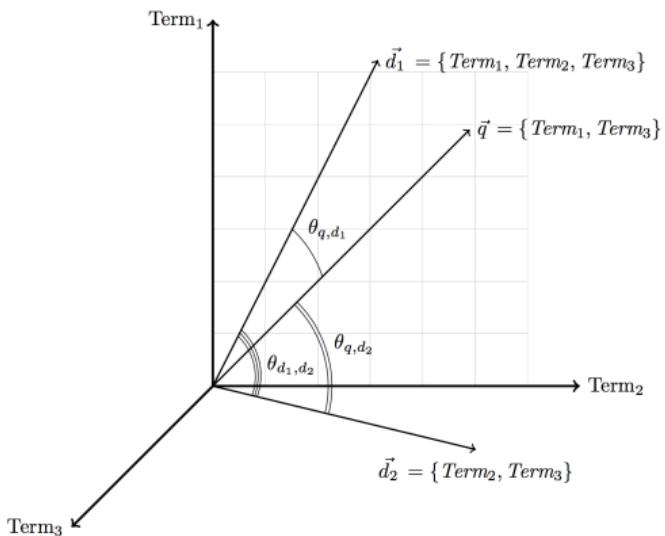
Vector Space Model

- A mathematical model to portray an n -dimensional space
- Entities are described by vectors with n coordinates in a real space \mathbb{R}^n
- Given two vectors, we can compute a similarity coefficient between them
- Cosine of the angle between two vectors reflects their degree of similarity

$$tf = 1 + \log(tf_{t,d}) \quad (1)$$

$$idf = \log \frac{N}{df_t} \quad (2)$$

$$\cos(\vec{q}, \vec{d}) = \frac{\sum_{i=1}^{|v|} q_i \cdot d_i}{\sqrt{\sum_{i=1}^{|v|} q_i^2} \cdot \sqrt{\sum_{i=1}^{|v|} d_i^2}} \quad (3)$$



Machine Learning
Primer

History
ML Types
Process

Clustering Documents

Motivation
k-Means Clustering
Application Example

Classifications &
Predictions

Introduction
Classification with kNN
Regression with kNN

Machine Learning
Evaluation

Evaluation Methodology
Evaluation Metrics
Error Analysis
Overfitting
Underfitting
Cross-Validation

Notes and Further
Reading

Intelligent Systems for Investigative Journalism

Organize large, unstructured document collections:

- Enron email dataset – ca. 500,000 emails from management
- Wikileaks – often releases millions of documents
 - Guantanamo Bay Files, TPP Agreements, CIA Documents, German BND-NSA Inquiry, ...
- Facebook internal documents leaks (Cambridge Analytica scandal, 7000 documents)
- Luanda Leaks (715,000 emails, charts, contracts, audits, etc.)
- Paradise Papers (13.4 million confidential papers regarding offshore investments)

Machine Learning
Primer

History
ML Types
Process

Clustering Documents

Motivation

k-Means Clustering
Application Example

Classifications & Predictions

Introduction
Classification with kNN
Regression with kNN

Machine Learning Evaluation

Evaluation Methodology
Evaluation Metrics
Error Analysis
Overfitting
Underfitting
Cross-Validation

Notes and Further Reading

Canada Revenue Agency launches 100 audits after Paradise Papers leak

René Witte



By **Alex Boutilier** Ottawa Bureau
▲ Tues., Jan. 29, 2019 | 2 min. read



<https://www.thestar.com/news/paradise-papers/2019/01/29/>

<canada-revenue-agency-launches-100-audits-after-paradise-papers-leak.html>

Machine Learning
Primer

History
ML Types
Process

Clustering Documents

Motivation
k-Means Clustering
Application Example

Classifications &
Predictions

Introduction
Classification with kNN
Regression with kNN

Machine Learning
Evaluation

Evaluation Methodology
Evaluation Metrics
Error Analysis
Overfitting
Underfitting
Cross-Validation

Notes and Further
Reading

[HOME](#) / [USAGE](#)

/ HOW TO SEARCH, EXPLORE, ANALYZE, STRUCTURE, FILTER AND VISUALIZE LARGE DOCUMENT COLLECTIONS OR MANY SEARCH RESULTS

How to search, explore, analyze, structure, filter and visualize large document collections or many search results

Semantic search, exploratory search, interactive filters, data visualization, information retrieval, document discovery & text mining

The screenshot shows the Open Semantic Search interface. At the top, there's a navigation bar with links for 'New search', 'Newest documents', 'Advanced search', 'Alert', 'Search by list', 'Manage structure', 'Datasources', and 'Help'. Below the navigation is a search bar with the placeholder 'annotate' and a 'Search' button. To the right of the search bar are 'Search options' and a 'Sort' dropdown set to 'Relevance'. The main area displays search results for 'Paths' and 'File date'. A sidebar on the right lists 'Tags' such as 'Faceted search', 'Hypothesis', and 'Open Source'. At the bottom, there's a footer with the text 'How to search, explore, analyze, structure, filter and visualize large document collections or many search results' and a link to 'Open Semantic Search'.

Machine Learning
Primer

History
ML Types
Process

Clustering Documents

Motivation
k-Means Clustering
Application Example

Classifications &
Predictions

Introduction
Classification with kNN
Regression with kNN

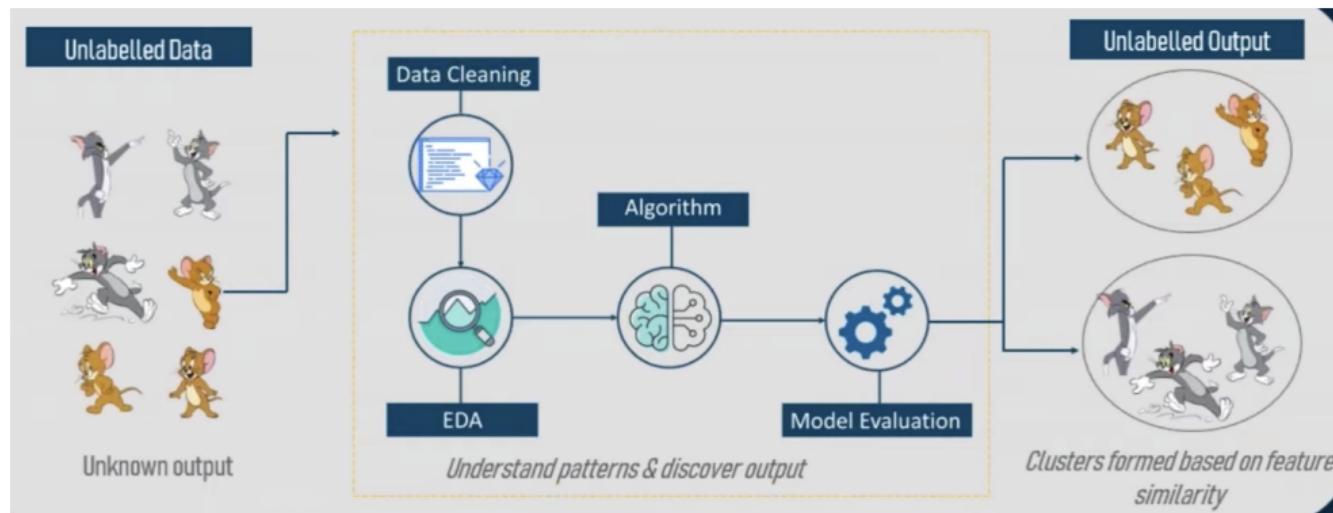
Machine Learning
Evaluation

Evaluation Methodology
Evaluation Metrics
Error Analysis
Overfitting
Underfitting
Cross-Validation

Notes and Further
Reading

Unsupervised Learning

- Remember, we do not “classify” documents (like in “spam vs. ham”)
- Rather, we group similar documents together
- Often used as a first exploratory step in data analysis
 - Data points (here: documents) in individual clusters can be further analyzed, possibly with different methods



Machine Learning Primer

History

ML Types

Process

Clustering Documents

Motivation

k-Means Clustering

Application Example

Classifications & Predictions

Introduction

Classification with kNN

Regression with kNN

Machine Learning Evaluation

Evaluation Methodology

Evaluation Metrics

Error Analysis

Overfitting

Underfitting

Cross-Validation

Notes and Further Reading

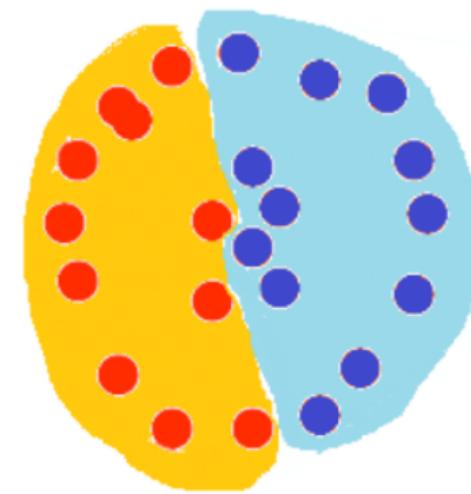
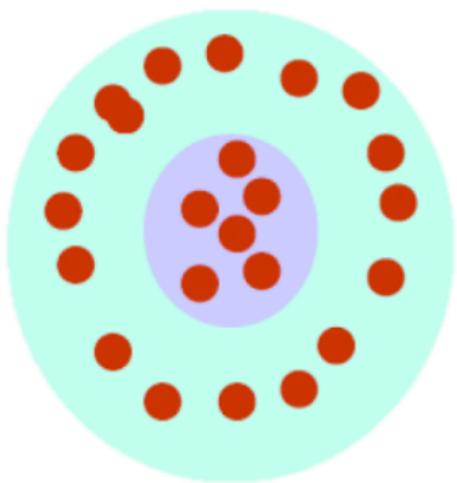
What are Clusters?

René Witte



Clustering

- The organization of unlabeled data into similarity groups, called **clusters**
- A cluster is a collection of data items which are “similar” between them, and “dissimilar” to data items in other clusters.
- Generally, there is no right or wrong answer to what the clusters in a dataset are.



Machine Learning Primer

History

ML Types

Process

Clustering Documents

Motivation

k-Means Clustering

Application Example

Classifications & Predictions

Introduction

Classification with kNN

Regression with kNN

Machine Learning Evaluation

Evaluation Methodology

Evaluation Metrics

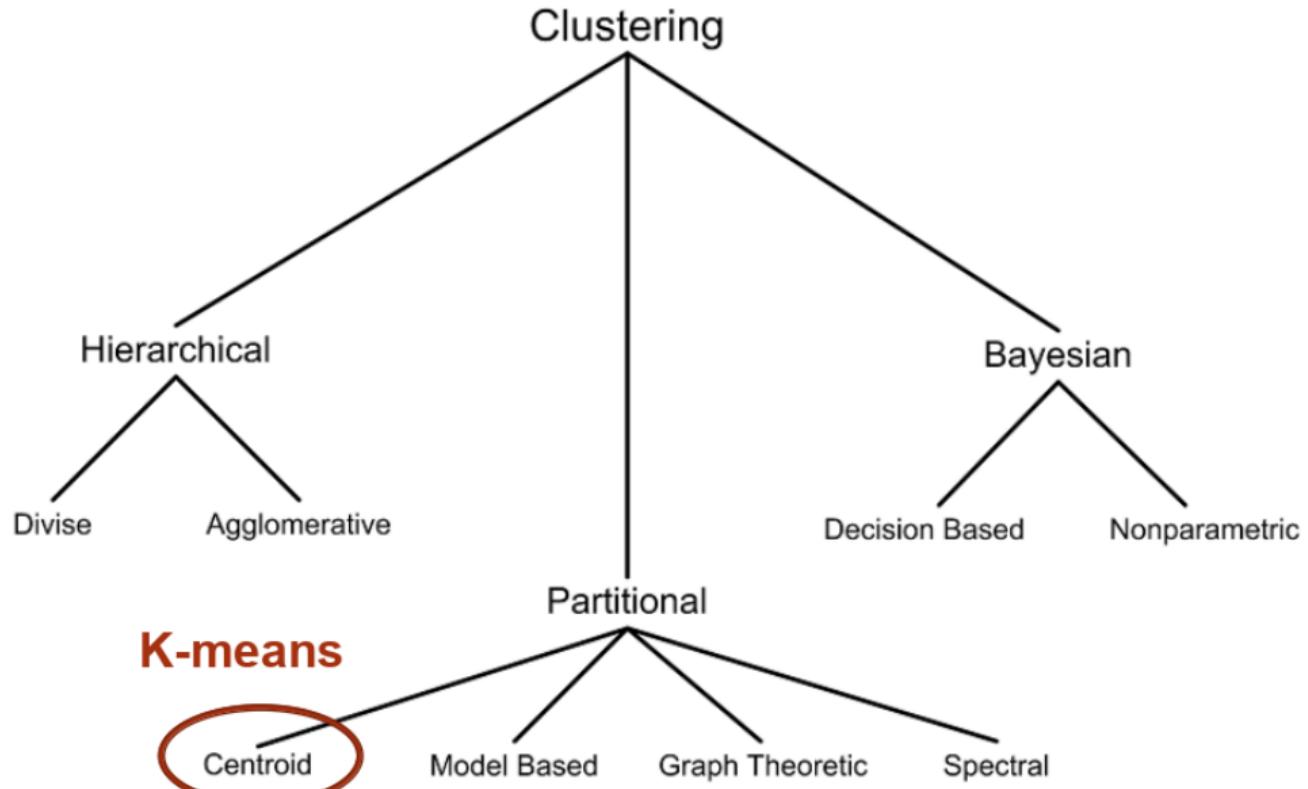
Error Analysis

Overfitting

Underfitting

Cross-Validation

Notes and Further Reading



Machine Learning Primer

History

ML Types

Process

Clustering Documents

Motivation

k-Means Clustering

Application Example

Classifications & Predictions

Introduction

Classification with kNN

Regression with kNN

Machine Learning Evaluation

Evaluation Methodology

Evaluation Metrics

Error Analysis

Overfitting

Underfitting

Cross-Validation

Notes and Further Reading

Partition-based Clustering

K-means (MacQueen, 1967) is a partitional clustering algorithm:

- Given m vectors in an n -dimensional space, $\vec{x}_1, \dots, \vec{x}_m \in \mathbb{R}^n$
- User defines k , the number of clusters

Algorithm

- ① Pick k points from the dataset (usually at random).
These points represent our initial group **centroïds**.
- ② Assign each data point \vec{x}_i to the nearest centroïd.
- ③ When all data points have been assigned, recalculate the positions of the k centroïds as the average of the cluster.
- ④ Repeat Steps 2 and 3 until none of the data instances change group
(or changes stay below a given convergence limit Δ).

Machine Learning
Primer

History

ML Types

Process

Clustering Documents

Motivation

k-Means Clustering

Application Example

Classifications &
Predictions

Introduction

Classification with kNN

Regression with kNN

Machine Learning
Evaluation

Evaluation Methodology

Evaluation Metrics

Error Analysis

Overfitting

Underfitting

Cross-Validation

Notes and Further
Reading

Euclidian Distance

René Witte



To find the nearest centroid:

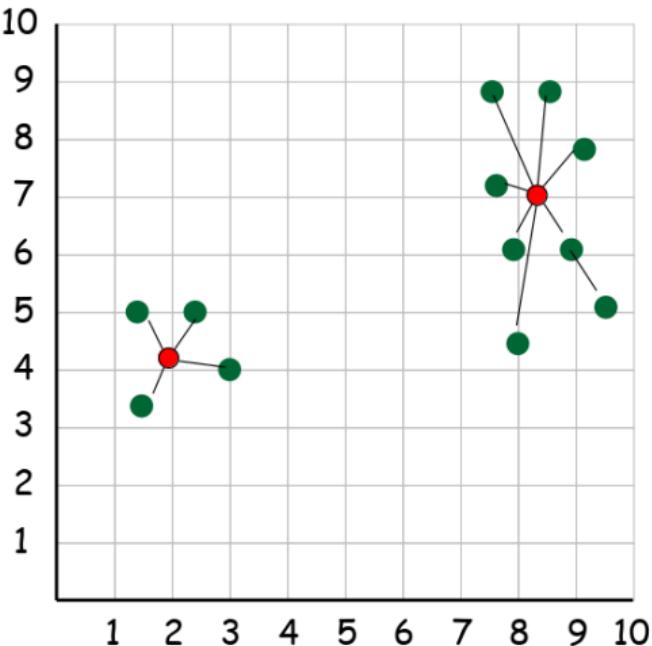
- a possible metric is the **Euclidean distance**
- distance d between 2 points p, q

$$p = (p_1, p_2, \dots, p_n)$$

$$q = (q_1, q_2, \dots, q_n)$$

$$d = \sqrt{\sum_{i=1}^n (p_i - q_i)^2}$$

- where to assign a data point \vec{x} ?
- → for all k clusters, choose the one where \vec{x} has the smallest distance



Machine Learning Primer

History

ML Types

Process

Clustering Documents

Motivation

k-Means Clustering

Application Example

Classifications & Predictions

Introduction

Classification with kNN

Regression with kNN

Machine Learning Evaluation

Evaluation Methodology

Evaluation Metrics

Error Analysis

Overfitting

Underfitting

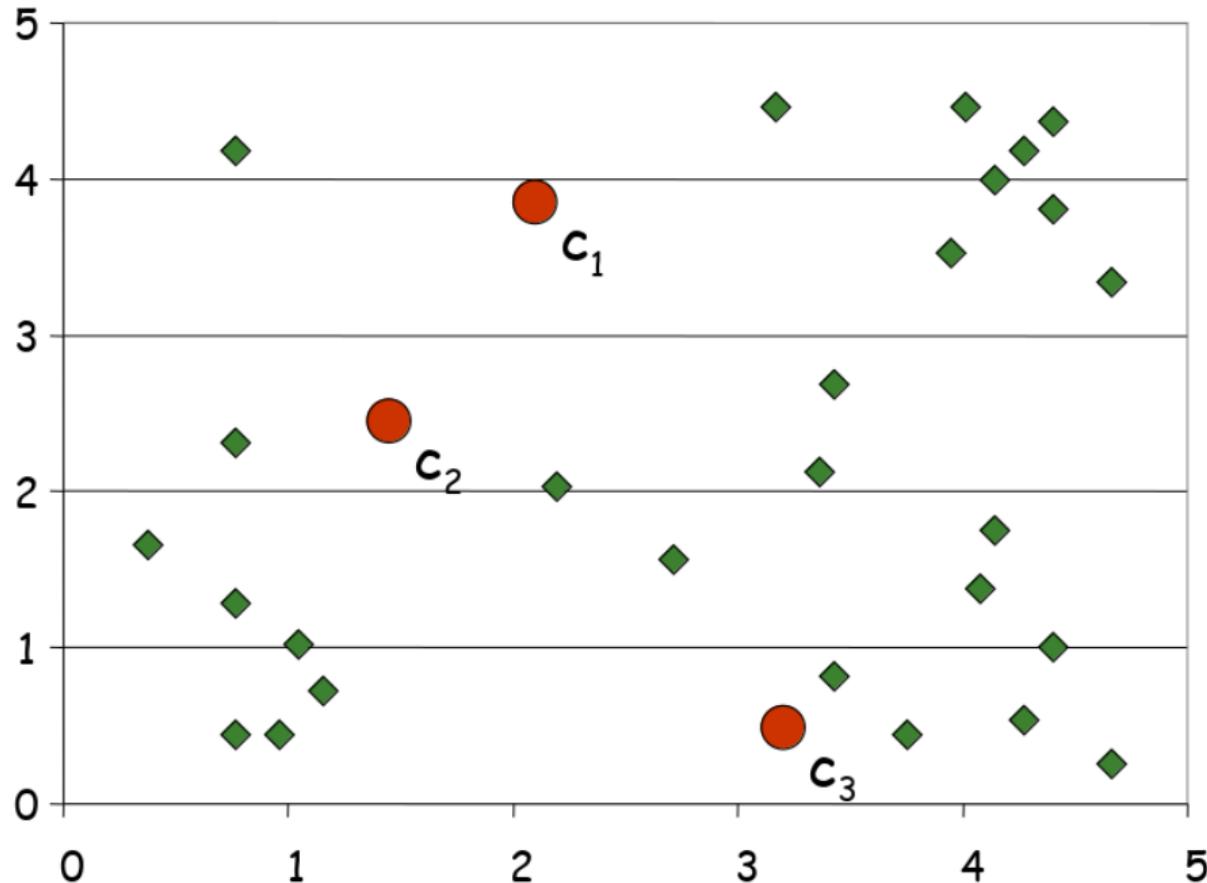
Cross-Validation

Notes and Further Reading

Example (1/5)

2D-vectors, k=3: Initialize random centroïds

René Witte



Machine Learning Primer

History

ML Types

Process

Clustering Documents

Motivation

k-Means Clustering

Application Example

Classifications & Predictions

Introduction

Classification with kNN

Regression with kNN

Machine Learning Evaluation

Evaluation Methodology

Evaluation Metrics

Error Analysis

Overfitting

Underfitting

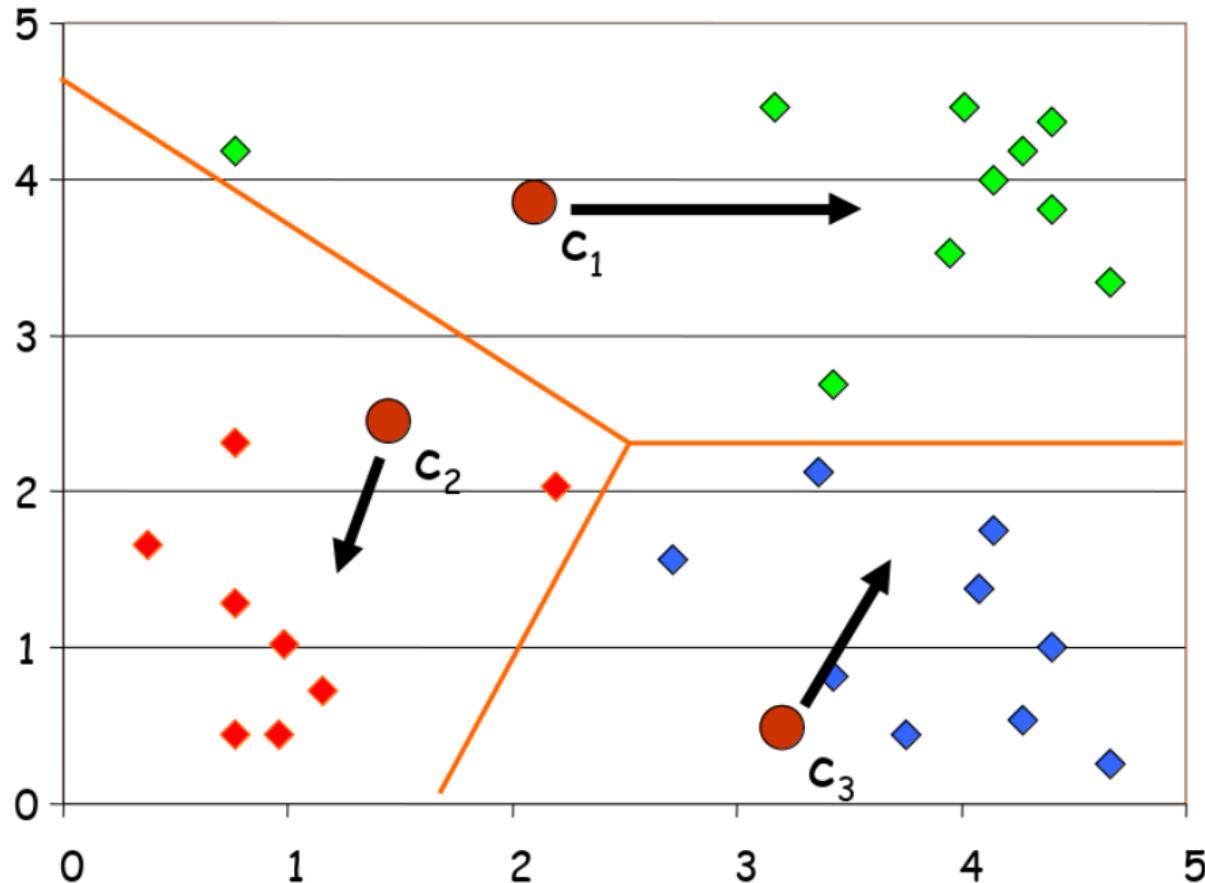
Cross-Validation

Notes and Further Reading

Example (2/5)

Partition data points to closest centroïds

René Witte



Machine Learning Primer

History

ML Types

Process

Clustering Documents

Motivation

k-Means Clustering

Application Example

Classifications & Predictions

Introduction

Classification with kNN

Regression with kNN

Machine Learning Evaluation

Evaluation Methodology

Evaluation Metrics

Error Analysis

Overfitting

Underfitting

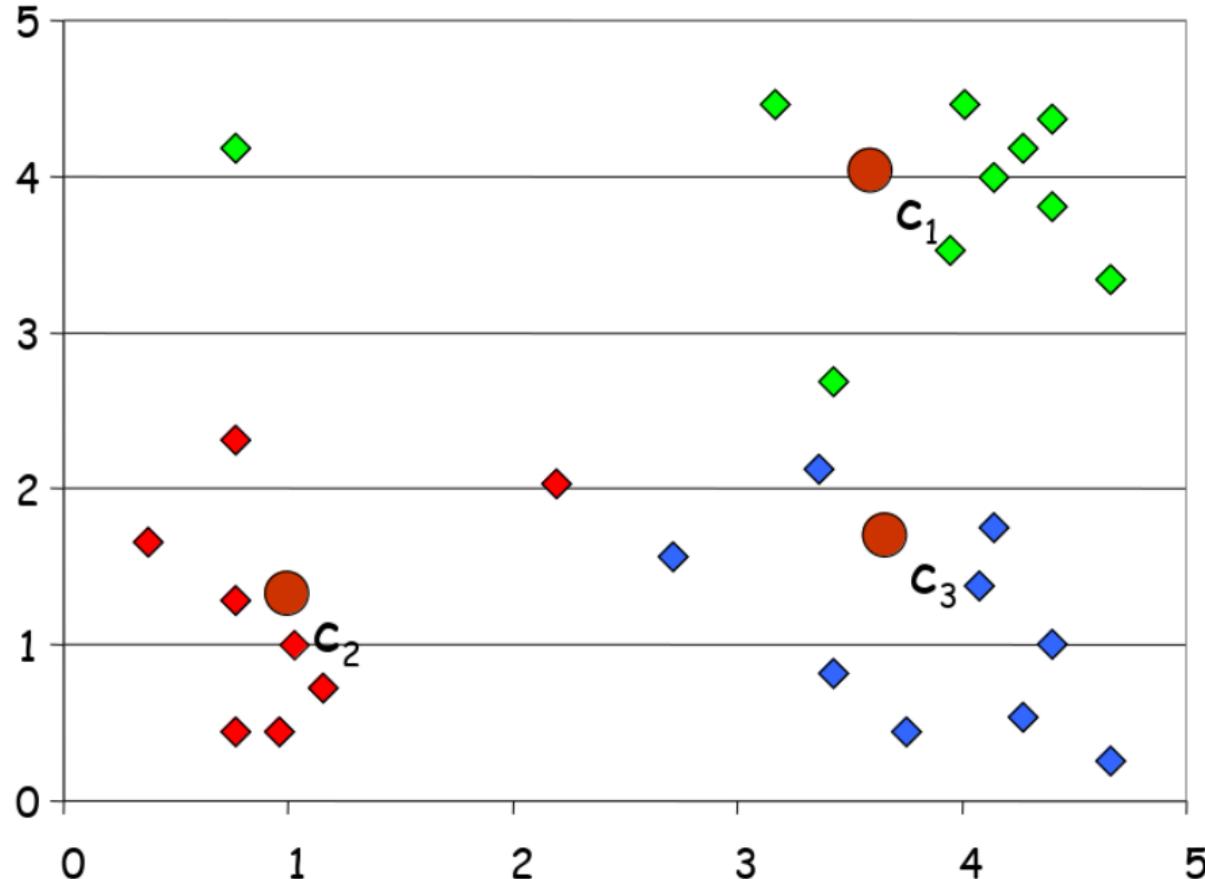
Cross-Validation

Notes and Further Reading

Example (3/5)

Compute new centroids

René Witte



Machine Learning Primer

History

ML Types

Process

Clustering Documents

Motivation

k-Means Clustering

Application Example

Classifications & Predictions

Introduction

Classification with kNN

Regression with kNN

Machine Learning Evaluation

Evaluation Methodology

Evaluation Metrics

Error Analysis

Overfitting

Underfitting

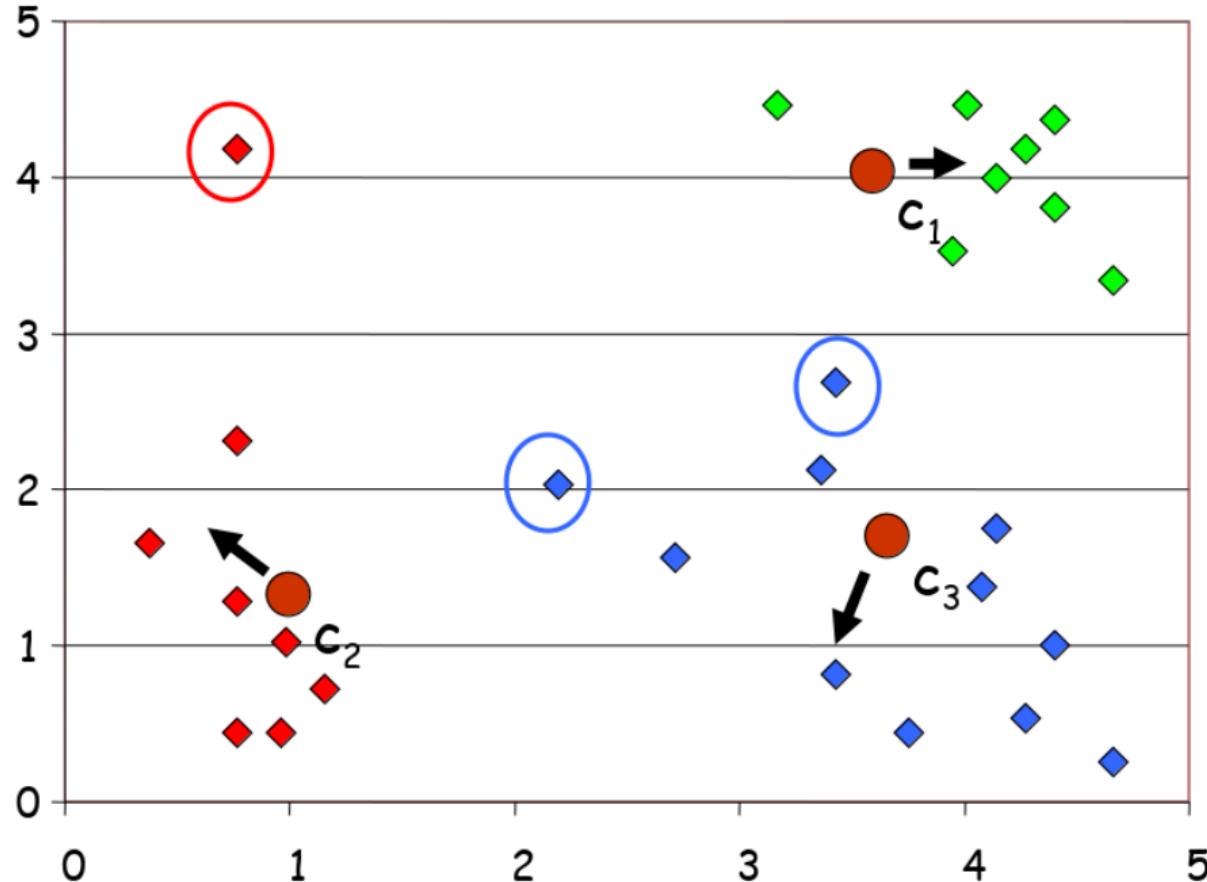
Cross-Validation

Notes and Further Reading

Example (4/5)

Re-assign data points to closest new centroïds

René Witte



Machine Learning Primer

History

ML Types

Process

Clustering Documents

Motivation

k-Means Clustering

Application Example

Classifications & Predictions

Introduction

Classification with kNN

Regression with kNN

Machine Learning Evaluation

Evaluation Methodology

Evaluation Metrics

Error Analysis

Overfitting

Underfitting

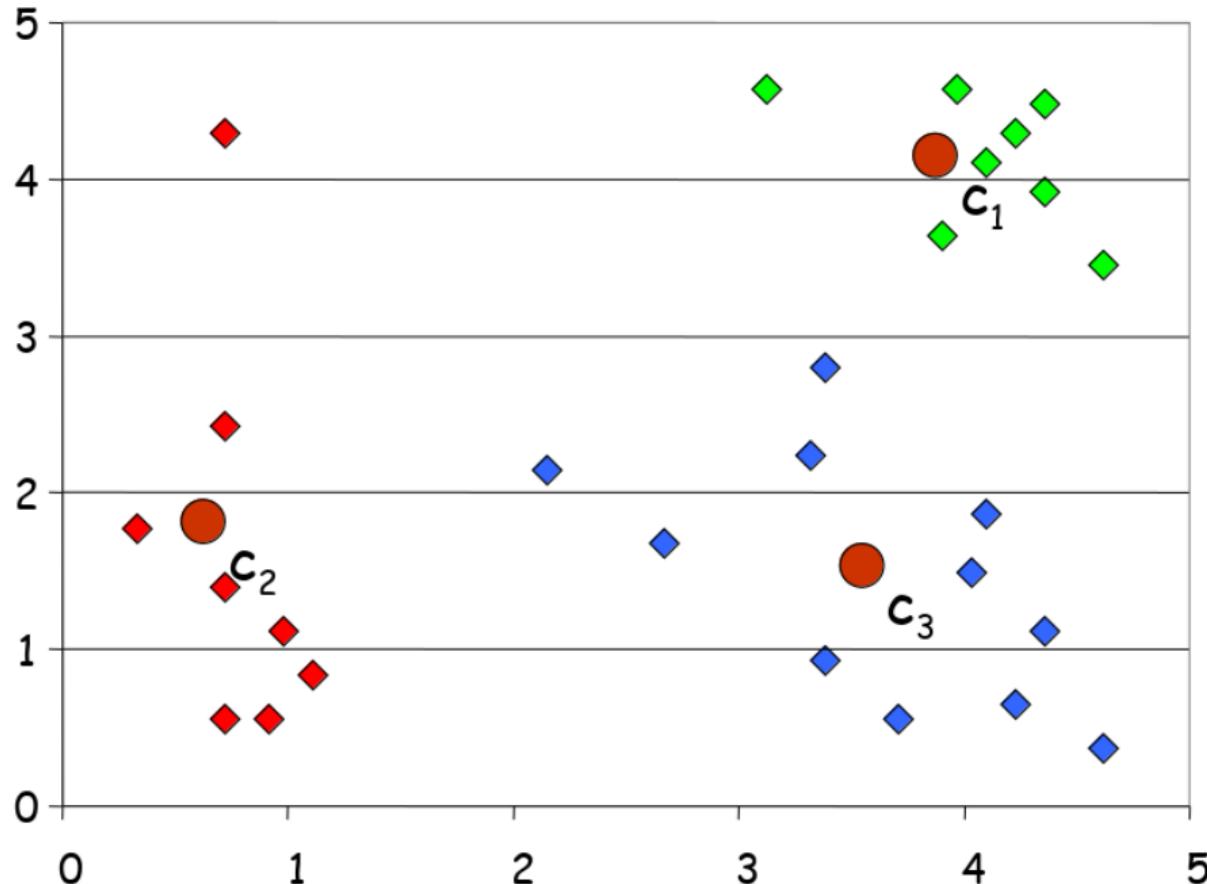
Cross-Validation

Notes and Further Reading

Example (5/5)

Repeat until clusters stabilize

René Witte



Machine Learning Primer

History

ML Types

Process

Clustering Documents

Motivation

k-Means Clustering

Application Example

Classifications & Predictions

Introduction

Classification with kNN

Regression with kNN

Machine Learning Evaluation

Evaluation Methodology

Evaluation Metrics

Error Analysis

Overfitting

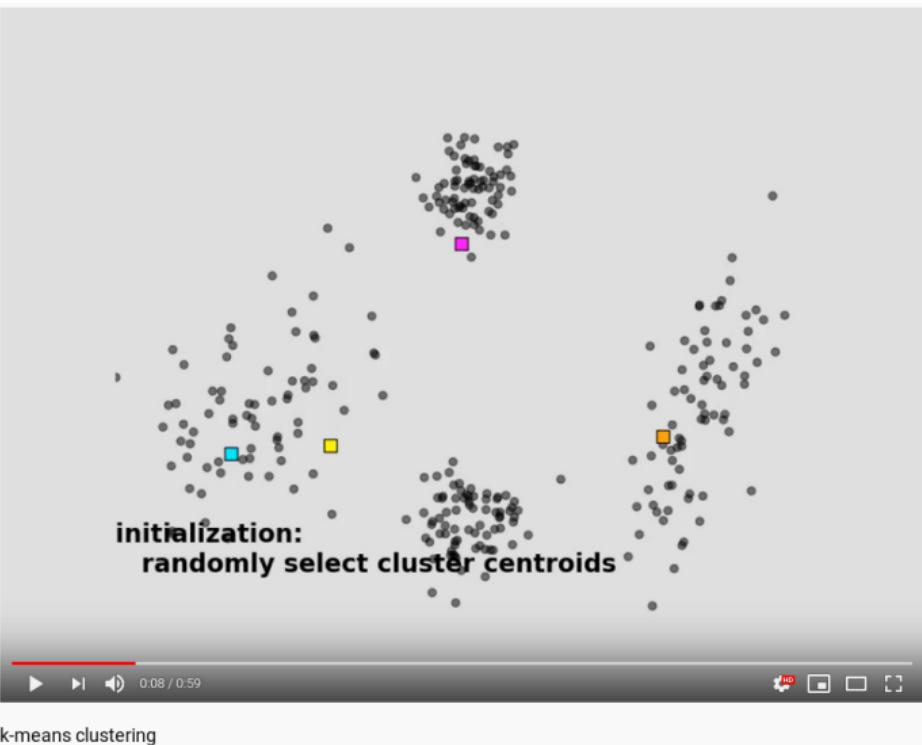
Underfitting

Cross-Validation

Notes and Further Reading

k-Means Clustering Illustrated

René Witte



Machine Learning
Primer

History

ML Types

Process

Clustering Documents

Motivation

k-Means Clustering

Application Example

Classifications &
Predictions

Introduction

Classification with kNN

Regression with kNN

Machine Learning
Evaluation

Evaluation Methodology

Evaluation Metrics

Error Analysis

Overfitting

Underfitting

Cross-Validation

Notes and Further
Reading

<https://www.youtube.com/watch?v=5I3Ei69I40s>

→ Worksheet #6: Task 2

Pros

- Simple, easy to understand and implement
 - Converges very fast
 - Efficient: Time complexity $O(t \cdot k \cdot n)$, with
 - n number of data points
 - k number of clusters
 - t number of iterations
- considered linear for practical purposes

Cons

- User needs to choose k (usually not known)
- Sensitive to outliers
- Different results on same dataset, based on initial (random) centroids

Machine Learning
Primer

History

ML Types

Process

Clustering Documents

Motivation

k-Means Clustering

Application Example

Classifications &
Predictions

Introduction

Classification with kNN

Regression with kNN

Machine Learning
Evaluation

Evaluation Methodology

Evaluation Metrics

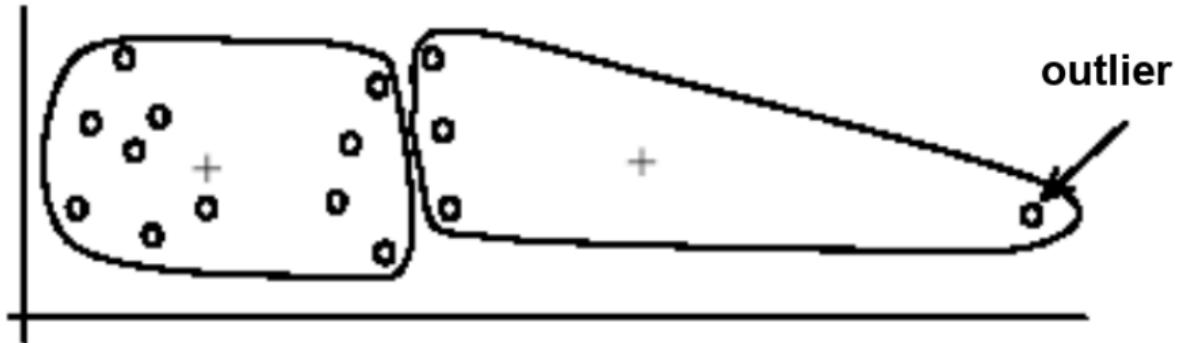
Error Analysis

Overfitting

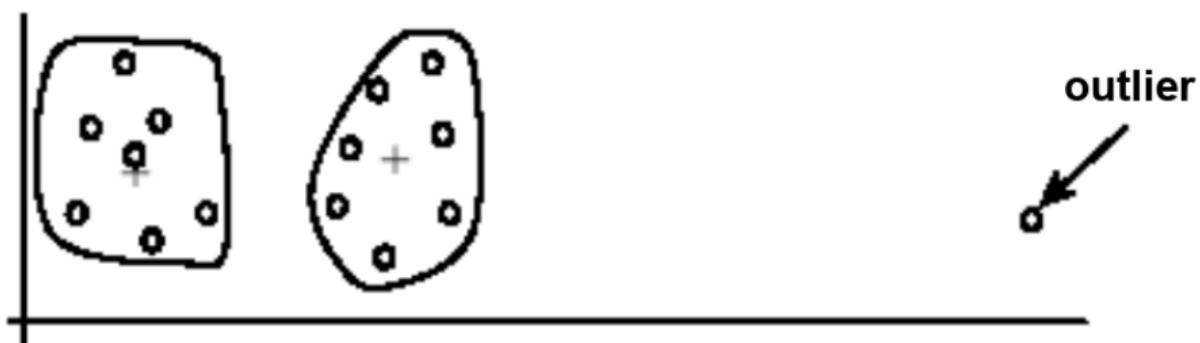
Underfitting

Cross-Validation

Notes and Further
Reading



(A) Undesirable clusters



(B) Ideal clusters

[Machine Learning Primer](#)

[History](#)

[ML Types](#)

[Process](#)

[Clustering Documents](#)

[Motivation](#)

[k-Means Clustering](#)

[Application Example](#)

[Classifications & Predictions](#)

[Introduction](#)

[Classification with kNN](#)

[Regression with kNN](#)

[Machine Learning Evaluation](#)

[Evaluation Methodology](#)

[Evaluation Metrics](#)

[Error Analysis](#)

[Overfitting](#)

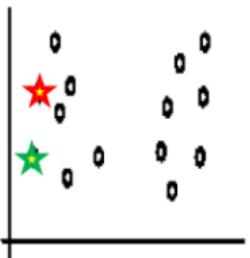
[Underfitting](#)

[Cross-Validation](#)

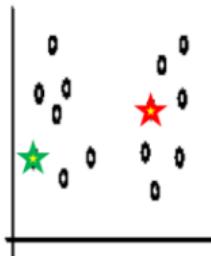
[Notes and Further Reading](#)

k-Means: Sensitivity to Initial Seeds

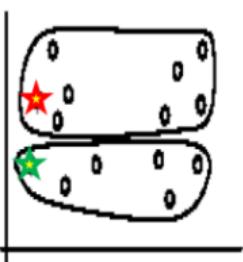
René Witte



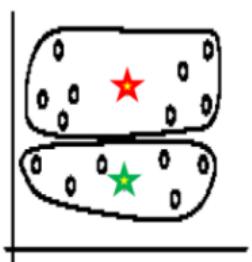
Random selection of seeds (centroids)



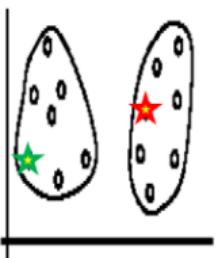
Random selection of seeds (centroids)



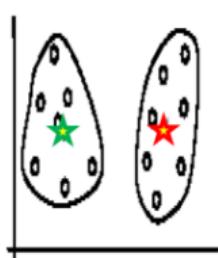
Iteration 1



Iteration 2



Iteration 1



Iteration 2

Machine Learning Primer

History

ML Types

Process

Clustering Documents

Motivation

k-Means Clustering

Application Example

Classifications & Predictions

Introduction

Classification with kNN

Regression with kNN

Machine Learning Evaluation

Evaluation Methodology

Evaluation Metrics

Error Analysis

Overfitting

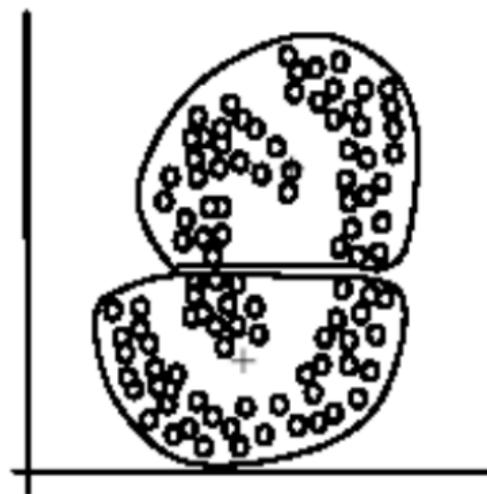
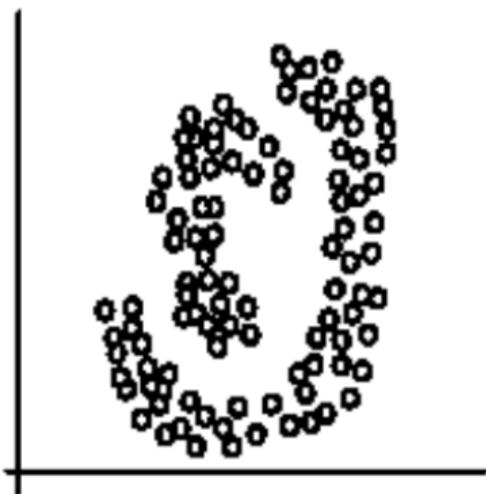
Underfitting

Cross-Validation

Notes and Further Reading

Summary

- Despite weaknesses, k-means is still one of the most popular algorithms, due to its simplicity and efficiency
- No clear evidence that any other clustering algorithm performs better in general
- Comparing different clustering algorithms is a difficult task:
No one knows the correct clusters!



Machine Learning
Primer

History

ML Types

Process

Clustering Documents

Motivation

k-Means Clustering

Application Example

Classifications &
Predictions

Introduction

Classification with kNN

Regression with kNN

Machine Learning
Evaluation

Evaluation Methodology

Evaluation Metrics

Error Analysis

Overfitting

Underfitting

Cross-Validation

Notes and Further
Reading

Document Clustering Example: Analyzing NFS Research Grants

René Witte



Analyzing US National Science Foundation research proposals using Lingo4G text clustering engine

<https://www.youtube.com/watch?v=85fZcK5EpnA>

Machine Learning Primer

History

ML Types

Process

Clustering Documents

Motivation

k-Means Clustering

Application Example

Classifications & Predictions

Introduction

Classification with kNN

Regression with kNN

Machine Learning Evaluation

Evaluation Methodology

Evaluation Metrics

Error Analysis

Overfitting

Underfitting

Cross-Validation

Notes and Further Reading

1 Machine Learning Primer

Machine Learning
Primer
History
ML Types
Process

2 Clustering Documents

Clustering Documents
Motivation
k-Means Clustering
Application Example

3 Classifications & Predictions

Introduction
Classification with kNN
Regression with kNN

Classifications &
Predictions
Introduction
Classification with kNN
Regression with kNN

4 Machine Learning Evaluation

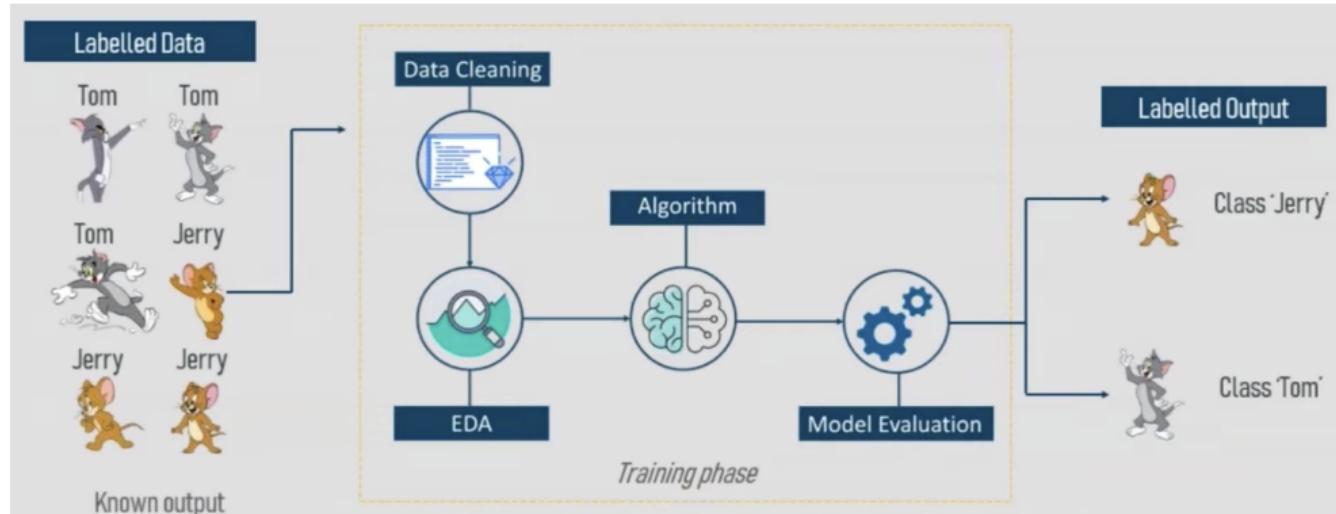
Machine Learning
Evaluation
Evaluation Methodology
Evaluation Metrics
Error Analysis
Overfitting
Underfitting
Cross-Validation

5 Notes and Further Reading

Notes and Further
Reading

Classification of Data

René Witte



Machine Learning Primer

History

ML Types

Process

Clustering Documents

Motivation

k-Means Clustering

Application Example

Classifications & Predictions

Introduction

Classification with kNN
Regression with kNN

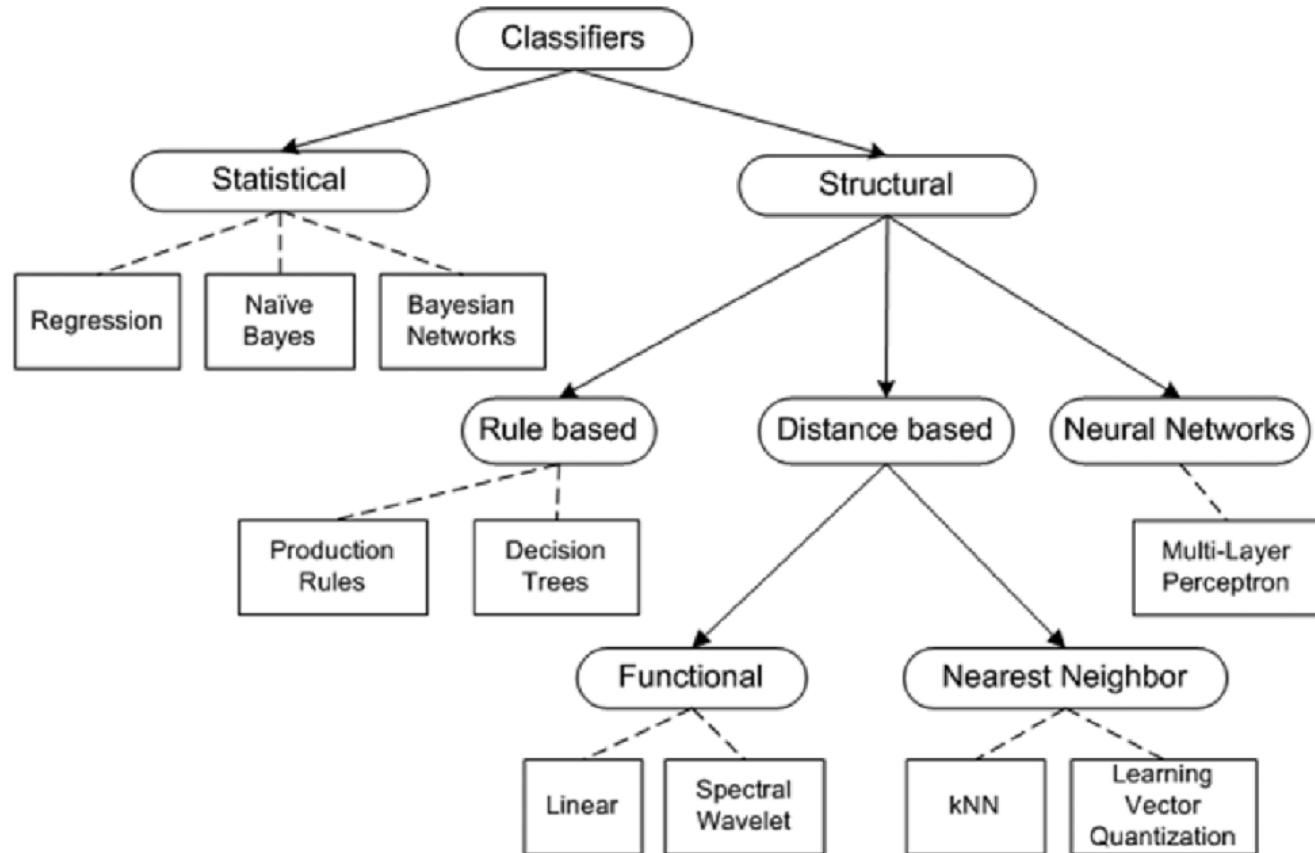
Machine Learning Evaluation

Evaluation Methodology
Evaluation Metrics
Error Analysis
Overfitting
Underfitting
Cross-Validation

Notes and Further Reading

Classification Algorithms

René Witte



Machine Learning Primer

History

ML Types

Process

Clustering Documents

Motivation

k-Means Clustering

Application Example

Classifications & Predictions

Introduction

Classification with kNN

Regression with kNN

Machine Learning Evaluation

Evaluation Methodology

Evaluation Metrics

Error Analysis

Overfitting

Underfitting

Cross-Validation

Notes and Further Reading

k-Nearest-Neighbor (kNN) Classification

René Witte

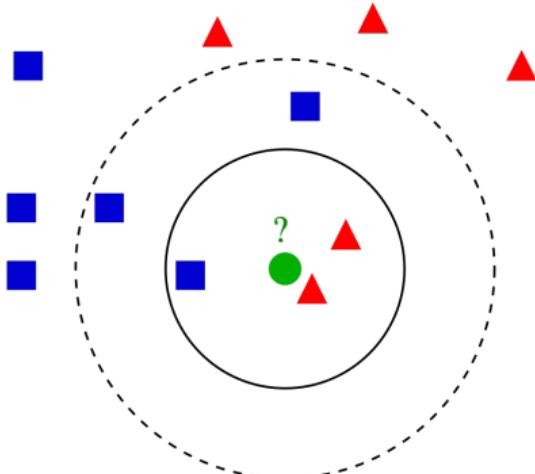


kNN Algorithm

Training: only store feature vectors + class labels

Testing: Find the k data points nearest (e.g., Euclidian distance) to the new value. Resulting class is decided by majority vote.

Note: in this simple form, kNN has no training effort, but large testing effort (so-called [lazy learning](#))



Copyright Anti Ajanki (<https://commons.wikimedia.org/wiki/File:KnnClassification.svg>), "KnnClassification", licensed under <https://creativecommons.org/licenses/by-sa/3.0/legalcode>

Machine Learning Primer

History

ML Types

Process

Clustering Documents

Motivation

k-Means Clustering

Application Example

Classifications & Predictions

Introduction

Classification with kNN

Regression with kNN

Machine Learning Evaluation

Evaluation Methodology

Evaluation Metrics

Error Analysis

Overfitting

Underfitting

Cross-Validation

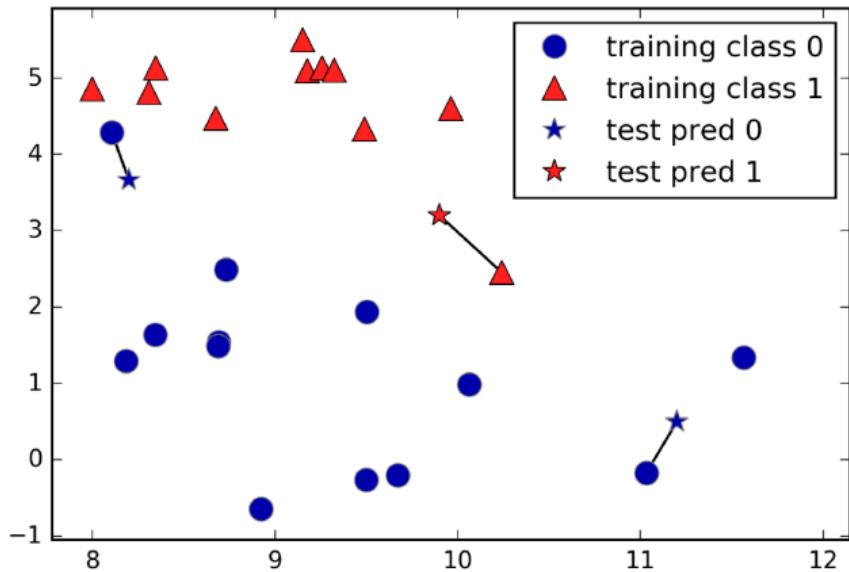
Notes and Further Reading

kNN Classification

René Witte

With $k = 1$

- Compute the distance of the unknown sample to all existing samples
- Assign the class of the *closest* neighbor to the new sample
 - Distance can be computed with different metrics, e.g., Euclidean distance or Manhattan distance



Machine Learning
Primer

History

ML Types
Process

Clustering Documents

Motivation

k-Means Clustering
Application Example

Classifications &
Predictions

Introduction

Classification with kNN
Regression with kNN

Machine Learning
Evaluation

Evaluation Methodology
Evaluation Metrics
Error Analysis
Overfitting
Underfitting
Cross-Validation

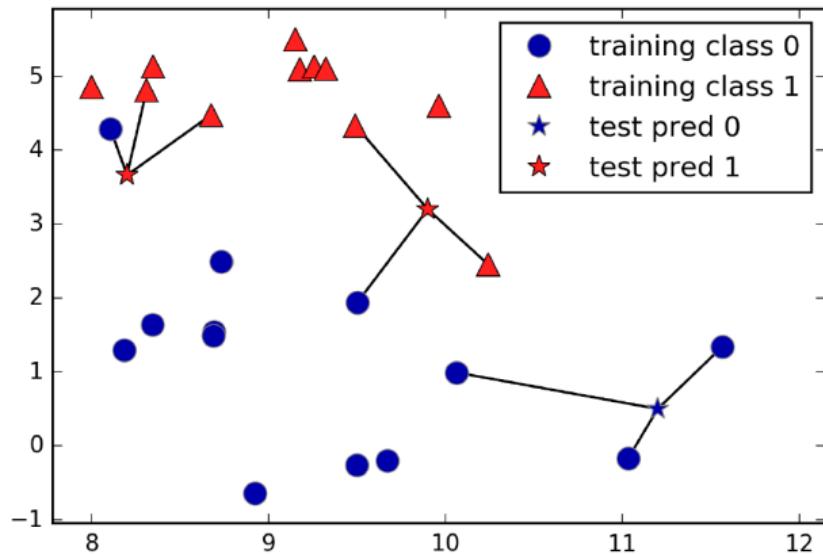
Notes and Further
Reading

kNN Classification: General case

René Witte

With arbitrary k

- kNN classification becomes a [voting algorithm](#)
- assign the same class as the [majority](#) of the k closest neighbors to the new sample
- Choice of k is dependent on data set



[Machine Learning Primer](#)

[History](#)

[ML Types](#)

[Process](#)

[Clustering Documents](#)

[Motivation](#)

[k-Means Clustering](#)

[Application Example](#)

[Classifications & Predictions](#)

[Introduction](#)

[Classification with kNN](#)

[Regression with kNN](#)

[Machine Learning Evaluation](#)

[Evaluation Methodology](#)

[Evaluation Metrics](#)

[Error Analysis](#)

[Overfitting](#)

[Underfitting](#)

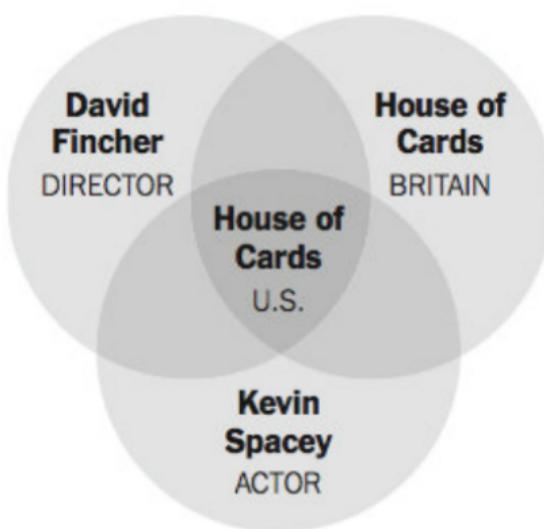
[Cross-Validation](#)

[Notes and Further Reading](#)

Netflix: Predict Success of Original Content

René Witte

In 2013, Netflix decided to commission two seasons of the U.S. remake of the British series *House of Cards* based on an analysis of its customers' data



Machine Learning
Primer

History

ML Types

Process

Clustering Documents

Motivation

k-Means Clustering

Application Example

Classifications &
Predictions

Introduction

Classification with kNN

Regression with kNN

Machine Learning
Evaluation

Evaluation Methodology

Evaluation Metrics

Error Analysis

Overfitting

Underfitting

Cross-Validation

Notes and Further
Reading

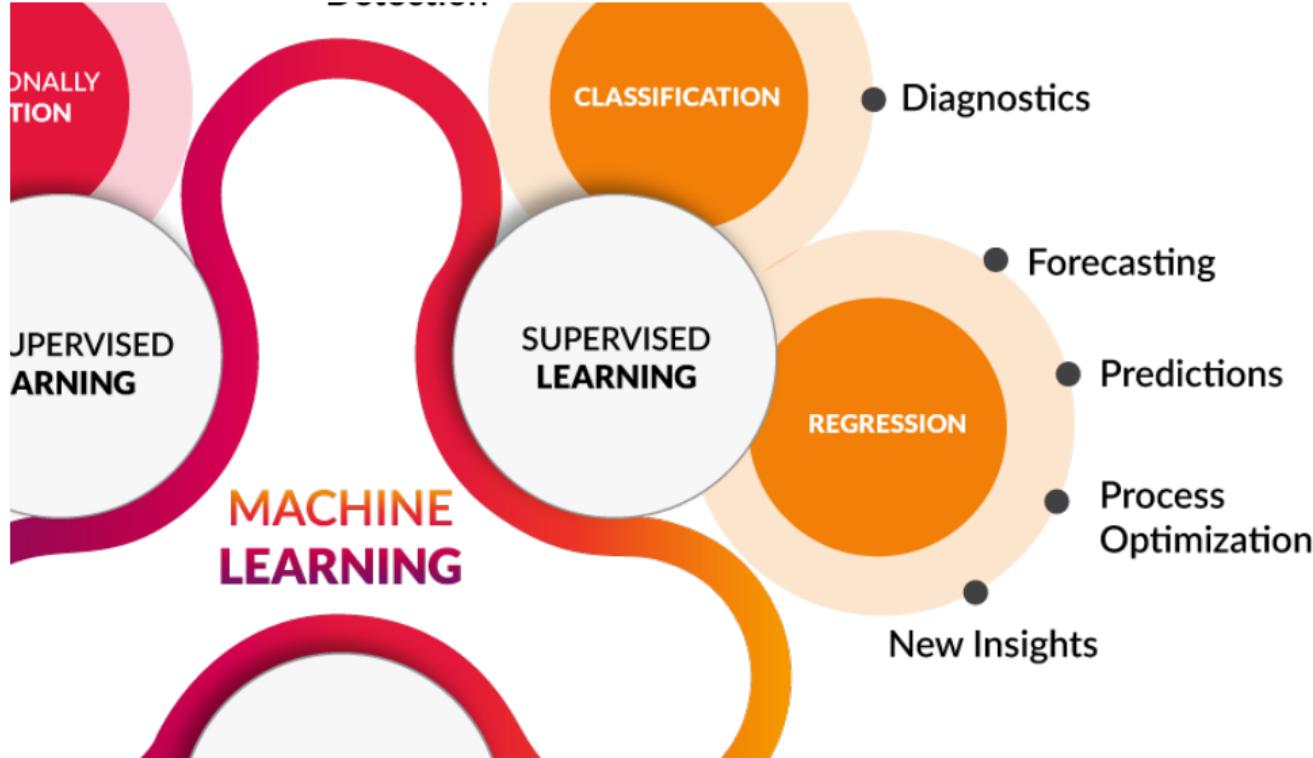
[https://informationstrategysm.wordpress.com/2014/10/19/
big-data-analytics-house-of-cards-and-future-of-television-creation-consumption/](https://informationstrategysm.wordpress.com/2014/10/19/big-data-analytics-house-of-cards-and-future-of-television-creation-consumption/)

→ Worksheet #6: Task 3

Regression

Forecasting or predicting a value: e.g., house price, movie rating, temperature at noon, ...

René Witte



Machine Learning Primer

History

ML Types

Process

Clustering Documents

Motivation

k-Means Clustering

Application Example

Classifications & Predictions

Introduction

Classification with kNN

Regression with kNN

Machine Learning Evaluation

Evaluation Methodology

Evaluation Metrics

Error Analysis

Overfitting

Underfitting

Cross-Validation

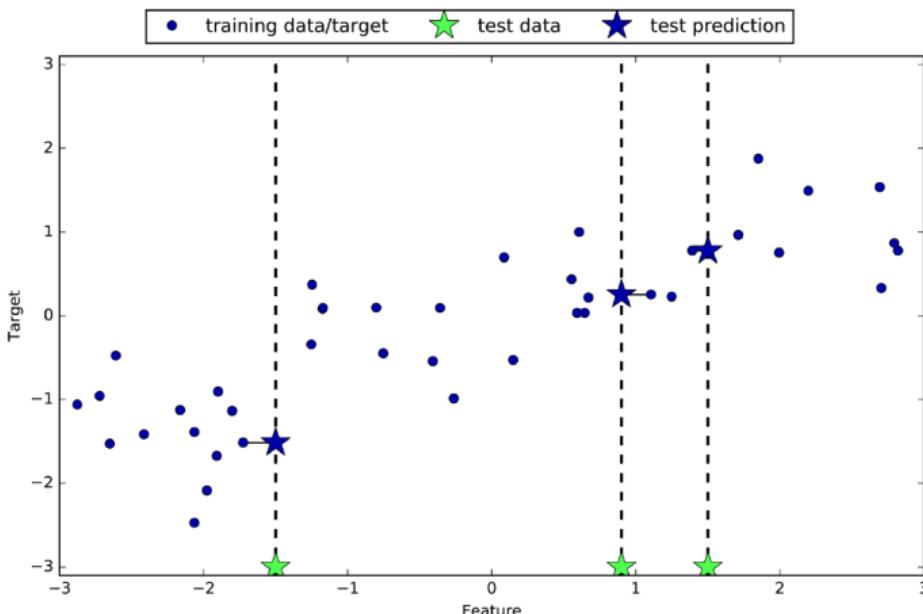
Notes and Further Reading

kNN Regression

René Witte

With $k = 1$

- Find the **nearest** existing data point to a new sample as before
- Assign the value of this point (e.g., *price*, *rating*, ...) to the new instance
 - Note: given n -dimensional vectors, we are using $n - 1$ dimensions for the similarity and the final for the predicted value



Machine Learning Primer
History
ML Types
Process

Clustering Documents
Motivation
k-Means Clustering
Application Example

Classifications & Predictions

Introduction
Classification with kNN
Regression with kNN

Machine Learning Evaluation

Evaluation Methodology
Evaluation Metrics
Error Analysis
Overfitting
Underfitting
Cross-Validation

Notes and Further Reading

kNN Regression: General Case

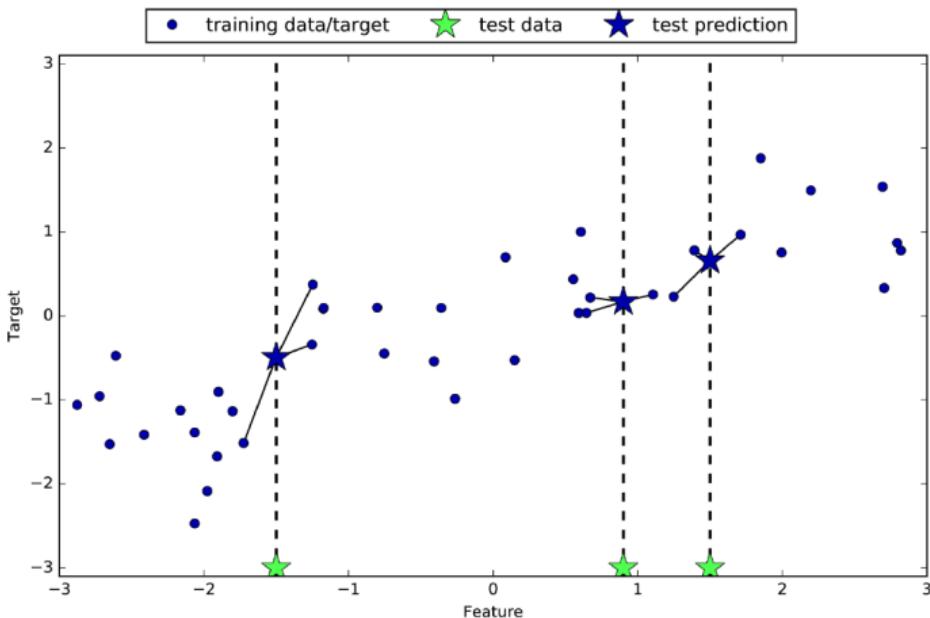
René Witte



Find the k nearest existing data points

Assign the average of their values to the new point

- Note that this algorithm cannot extrapolate



Copyright 2017 by O'Reilly Media, Inc. [MG17]

Machine Learning Primer
History
ML Types
Process

Clustering Documents
Motivation
k-Means Clustering
Application Example

Classifications & Predictions
Introduction
Classification with kNN
Regression with kNN

Machine Learning Evaluation
Evaluation Methodology
Evaluation Metrics
Error Analysis
Overfitting
Underfitting
Cross-Validation

Notes and Further Reading

→ Worksheet #6: Task 4

Machine Learning at Netflix

René Witte

A large digital wall composed of numerous small, rectangular screens arranged in a grid. Each screen displays a different machine learning dashboard or interface, likely related to Netflix's research. The screens are primarily green and red. The overall setup is modern and high-tech, with several pendant lights hanging from the ceiling above the wall.

0:23 / 3:00

▶ ▶ 🔍 0:23 / 3:00

HD CC

SHARE SAVE ...

Netflix Research: Machine Learning

1,297 views • Aug 8, 2018

15 0 SHARE SAVE ...

<https://www.youtube.com/watch?v=X9ZES-fsxgU>

Machine Learning
Primer

History

ML Types

Process

Clustering Documents

Motivation

k-Means Clustering

Application Example

Classifications &
Predictions

Introduction

Classification with kNN

Regression with kNN

Machine Learning
Evaluation

Evaluation Methodology

Evaluation Metrics

Error Analysis

Overfitting

Underfitting

Cross-Validation

Notes and Further
Reading

Outline

René Witte



1 Machine Learning Primer

Machine Learning Primer

History

ML Types

Process

2 Clustering Documents

Clustering Documents

Motivation

k-Means Clustering

Application Example

3 Classifications & Predictions

Classifications & Predictions

Introduction

Classification with kNN

Regression with kNN

4 Machine Learning Evaluation

Evaluation Methodology

Evaluation Metrics

Error Analysis

Overfitting

Underfitting

Cross-Validation

Machine Learning Evaluation

Evaluation Methodology

Evaluation Metrics

Error Analysis

Overfitting

Underfitting

Cross-Validation

5 Notes and Further Reading

Notes and Further Reading

Methodology

- How do you know if what you learned is correct?
- You run your classifier on a data set of **unseen** examples (that you did not use for training) for which you know the correct classification (“gold standard”)

Training vs. testing data

- Split data into **training** (80%) and **testing** (20%) sets
- Depending on ML algorithm, the training set can be further split into:
 - Actual training set (80%)
 - Validation set (20%)

Machine Learning
Primer

History
ML Types
Process

Clustering Documents
Motivation
k-Means Clustering
Application Example

Classifications &
Predictions
Introduction
Classification with kNN
Regression with kNN

Machine Learning
Evaluation

Evaluation Methodology
Evaluation Metrics
Error Analysis
Overfitting
Underfitting
Cross-Validation

Notes and Further
Reading

- ① Collect a large set of examples (all with correct classifications)
- ② Divide collection into **training**, **validation** and **test** set
- ③ Apply learning algorithm to training set to learn the parameters
- ④ Measure performance with the validation set, and adjust *hyper-parameters* to improve performance
- ⑤ Performance not good enough? \Rightarrow ③
- ⑥ Measure performance with the test set

DO NOT LOOK AT THE TEST SET

until you arrived at Step 6.

Parameters

Basic values learned by the ML model, e.g.:

- for NB: prior & conditional probabilities
- for DTs: features to split
- for ANNs: weights

Hyper-Parameters

Parameters used to set up the ML model, e.g.:

- for NB: value of delta for smoothing
- for DTs: pruning level
- for ANNs: # of hidden layers, # of nodes per layer...

Machine Learning
Primer

History
ML Types
Process

Clustering Documents
Motivation
k-Means Clustering
Application Example

Classifications &
Predictions

Introduction
Classification with kNN
Regression with kNN

Machine Learning
Evaluation

Evaluation Methodology
Evaluation Metrics
Error Analysis
Overfitting
Underfitting
Cross-Validation

Notes and Further
Reading

Accuracy

- % of instances of the test set the algorithm correctly classifies
- when all classes are equally important and represented

Recall & Precision

- when one class is more important than the others

F-Measure

- Combined Precision & Recall (harmonic mean)

Machine Learning
Primer
History
ML Types
Process

Clustering Documents
Motivation
k-Means Clustering
Application Example

Classifications &
Predictions
Introduction
Classification with kNN
Regression with kNN

Machine Learning
Evaluation
Evaluation Methodology
Evaluation Metrics
Error Analysis
Overfitting
Underfitting
Cross-Validation

Notes and Further
Reading

Evaluation of Classifiers

What kind of errors can we make?

| | | Reality says... | |
|-------------------|----------|---------------------|---------------------|
| | | Positive | Negative |
| Model predicts... | Positive | True Positive (TP) | False Positive (FP) |
| | Negative | False Negative (FN) | True Negative (TN) |

This is a so-called (binary) confusion matrix

Error Types

- False positive classification: **Type I error**
(“convict the innocent!”)
- False negative classification: **Type II error**
(“free the guilty!”)

Important realization: not all errors are created equal!

Voltaire: “*It is better to risk saving a guilty man than to condemn an innocent one.*”

Machine Learning
Primer

History

ML Types

Process

Clustering Documents

Motivation

k-Means Clustering

Application Example

Classifications &
Predictions

Introduction

Classification with kNN

Regression with kNN

Machine Learning
Evaluation

Evaluation Methodology

Evaluation Metrics

Error Analysis

Overfitting

Underfitting

Cross-Validation

Notes and Further
Reading

Evaluation Metrics

René Witte



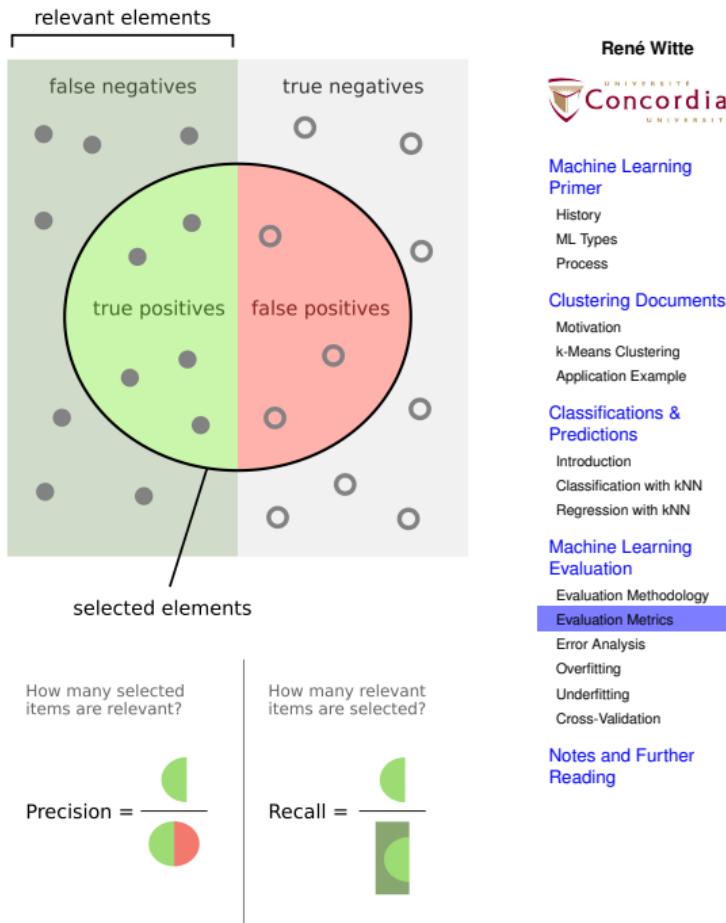
Commonly used

- Accuracy = $(TP + TN)/(P + N)$
- Recall = $TP/(TP + FN)$
- Precision = $TP/(TP + FP)$
- F_1 -score = $\frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$ (harmonic mean)

Mind the evaluation task

Precision, recall etc. are defined slightly differently for:

- Information retrieval tasks
- Classification tasks
- Ranked retrieval tasks
- Information extraction tasks



Machine Learning Primer

History

ML Types
Process

Clustering Documents

Motivation
k-Means Clustering
Application Example

Classifications & Predictions

Introduction
Classification with kNN
Regression with kNN

Machine Learning Evaluation

Evaluation Methodology
Evaluation Metrics
Error Analysis
Overfitting
Underfitting
Cross-Validation

Notes and Further Reading

Confusion Matrix

René Witte



- Where did the learner go wrong ?
- Use a [confusion matrix](#) (contingency table)

| correct class (that should have been assigned) | classes assigned by the learner | | | | | | | Total |
|--|---------------------------------|--------------------|--------------------|--------------------|--------------------|--------------------|---------------------|-----------------------|
| | C1 | C2 | C3 | C4 | C5 | C6 | ... | |
| C1 | 94 | 3 | 0 | 0 | 3 | 0 | 100 | |
| C2 | 0 | 93 | 3 | 4 | 0 | 0 | 100 | |
| C3 | 0 | 1 | 94 | 2 | 1 | 2 | 100 | |
| C4 | 0 | 1 | 3 | 94 | 2 | 0 | 100 | |
| C5 | 0 | 0 | 3 | 2 | 92 | 3 | 100 | |
| C6 | 0 | 0 | 5 | 0 | 10 | 85 | 100 | |
| ... | | | | | | | | |

[Machine Learning
Primer](#)

History

ML Types

Process

[Clustering Documents](#)

Motivation

k-Means Clustering

Application Example

[Classifications &
Predictions](#)

Introduction

Classification with kNN

Regression with kNN

[Machine Learning
Evaluation](#)

Evaluation Methodology

Evaluation Metrics

Error Analysis

Overfitting

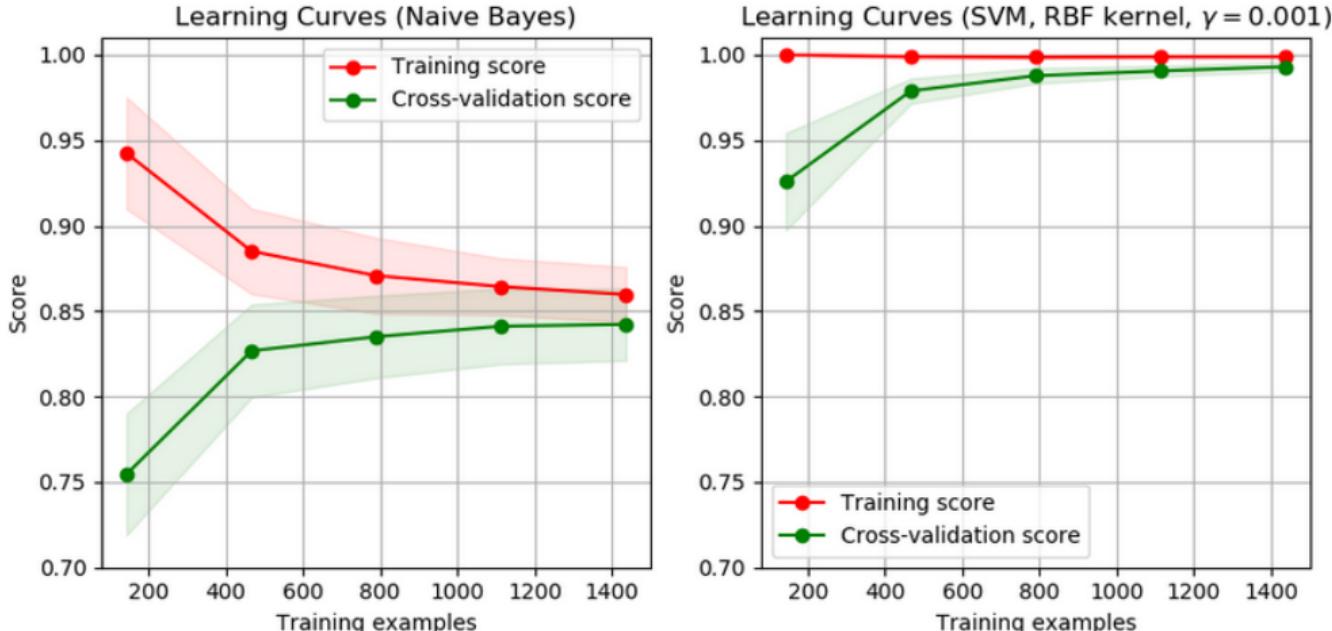
Underfitting

Cross-Validation

Notes and Further
Reading

Learning Curve

René Witte



Copyright 2007–2019, scikit-learn developers (BSD License), https://scikit-learn.org/stable/auto_examples/model_selection/plot_learning_curve.html

Plot evaluation metric vs. size of training set

- the more, the better
- but after a while, not much improvement...

Machine Learning Primer

History
ML Types
Process

Clustering Documents

Motivation
k-Means Clustering
Application Example

Classifications & Predictions

Introduction
Classification with kNN
Regression with kNN

Machine Learning Evaluation

Evaluation Methodology
Evaluation Metrics

Error Analysis

Overfitting
Underfitting
Cross-Validation

Notes and Further Reading

Some Words on Training...

René Witte



Watch out for:

- Noisy Data
- Overfitting/Underfitting

Machine Learning
Primer
History
ML Types
Process

Clustering Documents
Motivation
k-Means Clustering
Application Example

Classifications &
Predictions
Introduction
Classification with kNN
Regression with kNN

Machine Learning
Evaluation
Evaluation Methodology
Evaluation Metrics

Error Analysis
Overfitting
Underfitting
Cross-Validation

Notes and Further
Reading

Common issues

- Two examples have the same feature-value pairs, but different outputs
- Some values of features are incorrect or missing (ex. errors in the data acquisition)
- Some relevant attributes are not taken into account in the data set

| Size | Color | Shape | Output |
|------|-------|--------|--------|
| Big | Red | Circle | + |
| Big | Red | Circle | - |

Machine Learning
Primer

History
ML Types
Process

Clustering Documents
Motivation
k-Means Clustering
Application Example

Classifications &
Predictions
Introduction
Classification with kNN
Regression with kNN

Machine Learning
Evaluation
Evaluation Methodology
Evaluation Metrics

Error Analysis
Overfitting
Underfitting
Cross-Validation

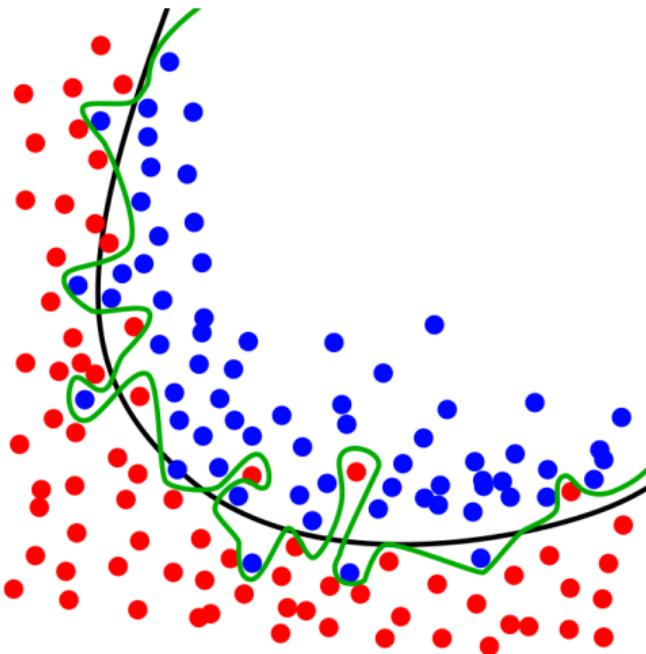
Notes and Further
Reading

Overfitting

René Witte



- If a large number of irrelevant features are there, we may find meaningless regularities in the data that are particular to the training data but irrelevant to the problem.
- Complicated boundaries **overfit** the data (a.k.a. *overtraining*)
- they are too tuned to the particular training data at hand
- They do not **generalize** well to the new data
- Extreme case: “rote learning”
 - Training error is low
 - Testing error is high



Copyright by Chabacano (<https://commons.wikimedia.org/wiki/File:Overfitting.svg>) license under the Creative Commons Attribution-Share Alike 4.0 International license, <https://creativecommons.org/licenses/by-sa/4.0/legalcode>

Machine Learning Primer

History

ML Types

Process

Clustering Documents

Motivation

k-Means Clustering

Application Example

Classifications & Predictions

Introduction

Classification with kNN

Regression with kNN

Machine Learning Evaluation

Evaluation Methodology

Evaluation Metrics

Error Analysis

Overfitting

Underfitting

Cross-Validation

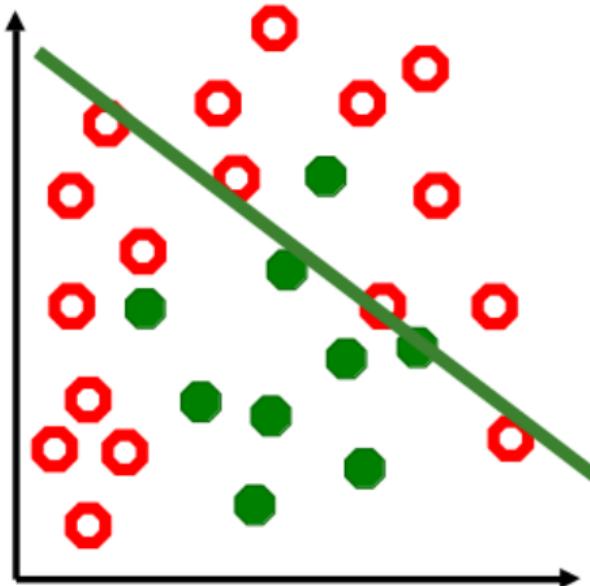
Notes and Further Reading

Underfitting

René Witte



- We can also underfit data, i.e. use too simple decision boundary
- Model is not expressive enough (not enough features)
- a.k.a. Undertraining
- There is no way to fit a linear decision boundary so that the training examples are well separated
 - Training error is high
 - Testing error is high



Machine Learning Primer

History

ML Types

Process

Clustering Documents

Motivation

k-Means Clustering

Application Example

Classifications & Predictions

Introduction

Classification with kNN
Regression with kNN

Machine Learning Evaluation

Evaluation Methodology
Evaluation Metrics

Error Analysis

Overfitting

Underfitting

Cross-Validation

Notes and Further Reading

Example: Animal Classification

René Witte

Features

What about cat vs. dog?

| has-hair? | has-scales? | has-feathers? | flies? | lives in water? | lays eggs? | |
|-----------|-------------|---------------|--------|-----------------|------------|-----------|
| 1 | 0 | 0 | 0 | 0 | 0 | Dog |
| 1 | 0 | 0 | 0 | 0 | 0 | Cat |
| 1 | 0 | 0 | 1 | 0 | 0 | Bat |
| 1 | 0 | 0 | 0 | 1 | 0 | Whale |
| 0 | 0 | 1 | 1 | 0 | 1 | Canary |
| 0 | 0 | 1 | 1 | 0 | 1 | Robin |
| 0 | 0 | 1 | 1 | 0 | 1 | Ostrich |
| 0 | 1 | 0 | 0 | 0 | 1 | Snake |
| 0 | 1 | 0 | 0 | 0 | 1 | Lizard |
| 0 | 1 | 0 | 0 | 1 | 1 | Alligator |

Machine Learning
Primer

History

ML Types

Process

Clustering Documents

Motivation

k-Means Clustering

Application Example

Classifications &
Predictions

Introduction

Classification with kNN

Regression with kNN

Machine Learning
Evaluation

Evaluation Methodology

Evaluation Metrics

Error Analysis

Overfitting

Underfitting

Cross-Validation

Notes and Further
Reading

Data Scarcity

- there is never enough training data
- so testing data is precious as well

k-fold Cross-Validation

'Re-use' different parts of the training data for testing. E.g., 10-fold cross-validation:

- split data into 10 equal parts
- train on 9 of these, test on the 10th
- repeat 10 times, resulting in 10 different performance results
- average these for overall performance

| | | | | | | | | | | |
|-------|-------|--|--|--|--|--|------|-------|------|-------|
| exp1: | train | | | | | | | | test | |
| exp2: | train | | | | | | | | test | train |
| exp3: | train | | | | | | test | train | | |
| ... | ... | | | | | | | | | |

Machine Learning
Primer

History

ML Types

Process

Clustering Documents

Motivation

k-Means Clustering

Application Example

Classifications &
Predictions

Introduction

Classification with kNN

Regression with kNN

Machine Learning
Evaluation

Evaluation Methodology

Evaluation Metrics

Error Analysis

Overfitting

Underfitting

Cross-Validation

Notes and Further
Reading

- 1 Machine Learning Primer
- 2 Clustering Documents
- 3 Classifications & Predictions
- 4 Machine Learning Evaluation
- 5 Notes and Further Reading

Machine Learning
Primer
History
ML Types
Process

Clustering Documents
Motivation
k-Means Clustering
Application Example

Classifications &
Predictions
Introduction
Classification with kNN
Regression with kNN

Machine Learning
Evaluation
Evaluation Methodology
Evaluation Metrics
Error Analysis
Overfitting
Underfitting
Cross-Validation

Notes and Further
Reading

Required

- [MG17, Chapters 2, 3, 5] (kNN, k-Means, Evaluation)

Supplemental

- [PS12, Chapter 7] (ML Training)
- [PS12, Chapter 8] (Testing and Evaluation)

Machine Learning
Primer
History
ML Types
Process

Clustering Documents
Motivation
k-Means Clustering
Application Example

Classifications &
Predictions
Introduction
Classification with kNN
Regression with kNN

Machine Learning
Evaluation
Evaluation Methodology
Evaluation Metrics
Error Analysis
Overfitting
Underfitting
Cross-Validation

Notes and Further
Reading

References

René Witte



- [MG17] Andreas C Müller and Sarah Guido.
Introduction to Machine Learning with Python.
O'Reilly, 2017.
<https://concordiauniversity.on.worldcat.org/oclc/960211579>.
- [PS12] James Pustejovsky and Amber Stubbs.
Natural Language Annotation for Machine Learning.
O'Reilly, 2012.
<https://concordiauniversity.on.worldcat.org/oclc/801812987>.

Machine Learning
Primer

History
ML Types
Process

Clustering Documents
Motivation
k-Means Clustering
Application Example

Classifications &
Predictions
Introduction
Classification with kNN
Regression with kNN

Machine Learning
Evaluation
Evaluation Methodology
Evaluation Metrics
Error Analysis
Overfitting
Underfitting
Cross-Validation

Notes and Further
Reading