Machine learning

Courses: Industrial automation, Automatizace pro průmyslovou praxi

CTU, FS, U12110

Matouš Cejnek

Contents

- Introduction
- Classification & Clustering
- Cross-validation

Introduction

Machine learning (ML)

A branch of artificial intelligence (AI) where **computers learn from data to improve their performance** on a task.

Or in other words:

The **ability of computer systems to learn patterns** and make decisions without explicit programming.

Machine learning utilization

- **Data analysis**:
  - Pre-processing
  - Core-analysis
  - Post-processing
- Applied data science
  - Data product deployment in production

ML data analysis

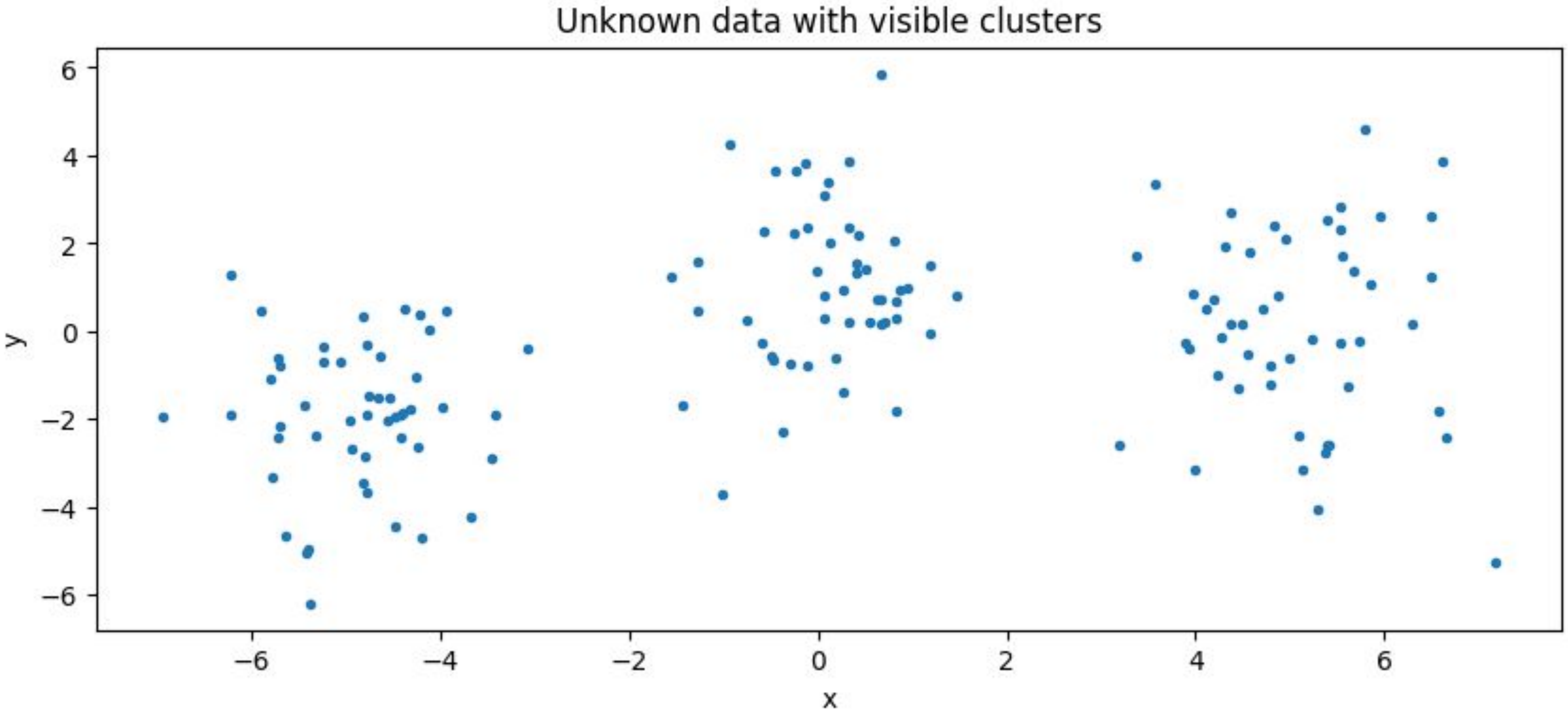
Most common:

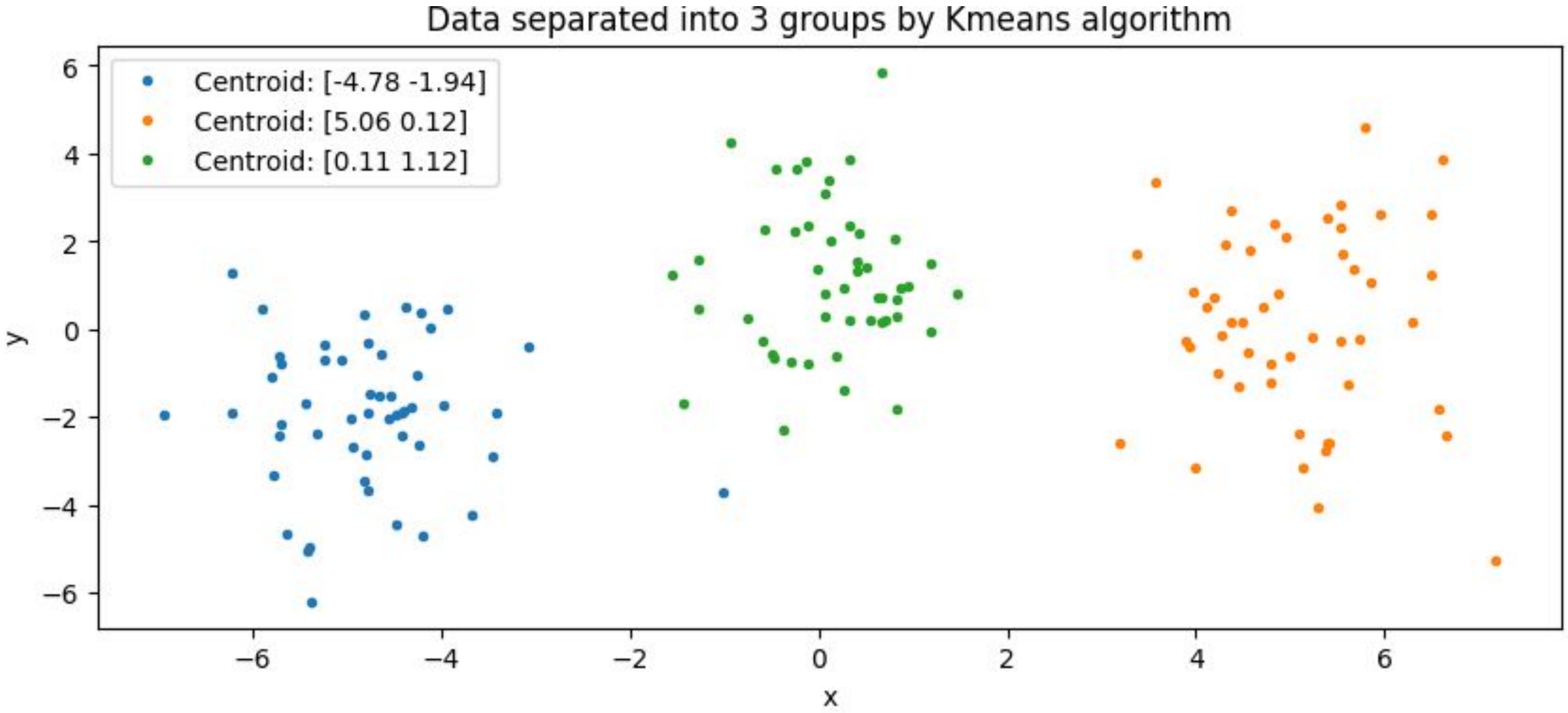
- Regression
- **Classification**
- **Clustering**

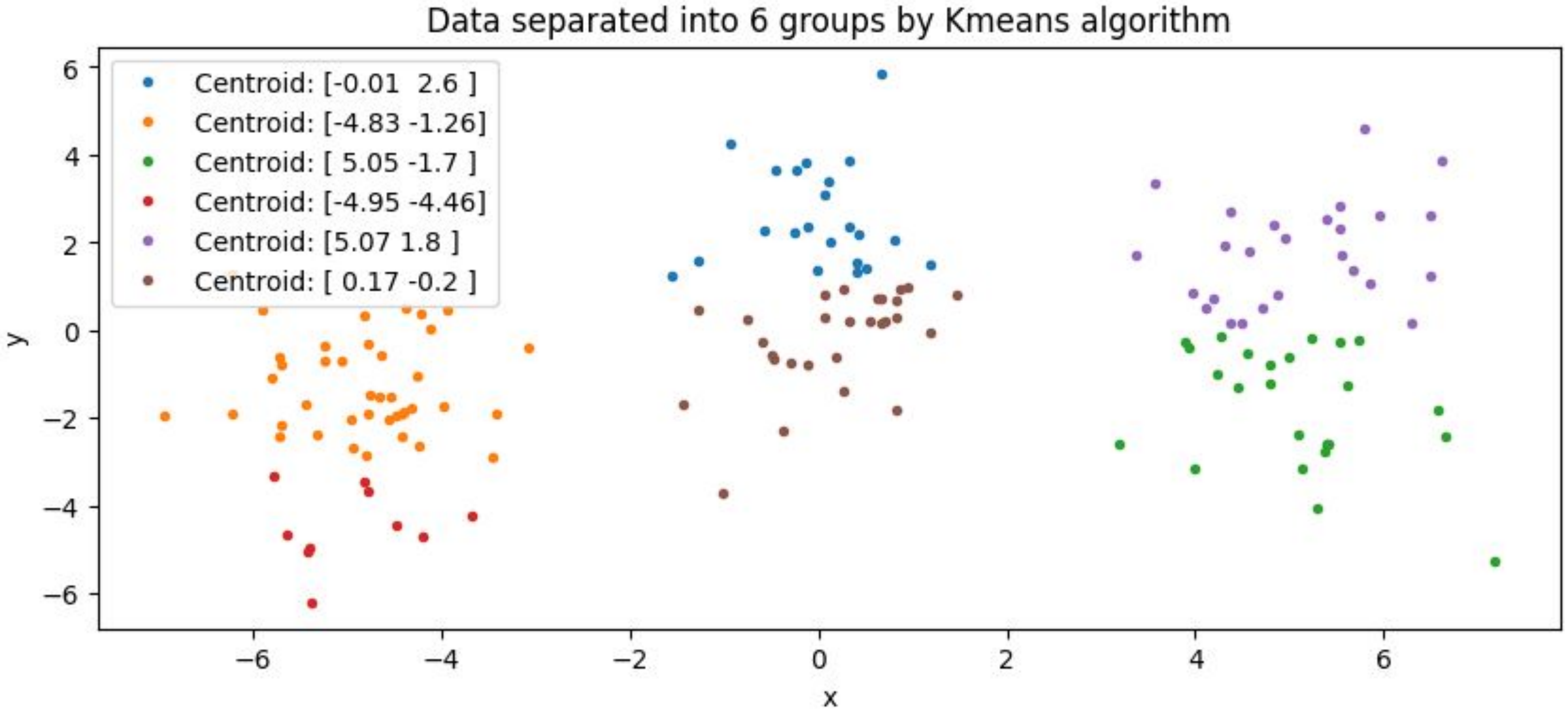
Clustering

Clustering is an **unsupervised** learning technique in machine learning that involves grouping similar data points into distinct subsets or clusters.

- We're unsure about cluster shapes or their quantity


Unknown data with visible clusters
6
4
2
0
−2
−4
−6
y
−6
−4
−2
0
2
4
6
x


Data separated into 3 groups by Kmeans algorithm
● Centroid: [-4.78 -1.94]
● Centroid: [5.06 0.12]
● Centroid: [0.11 1.12]
y
6
4
2
0
−2
−4
−6
x
−6
−4
−2
0
2
4
6


Data separated into 6 groups by Kmeans algorithm
● Centroid: [-0.01 2.6]
● Centroid: [-4.83 -1.26]
● Centroid: [5.05 -1.7]
● Centroid: [-4.95 -4.46]
● Centroid: [5.07 1.8]
● Centroid: [0.17 -0.2]
6
4
2
0
−2
−4
−6
y
−6
−4
−2
0
2
4
6
x

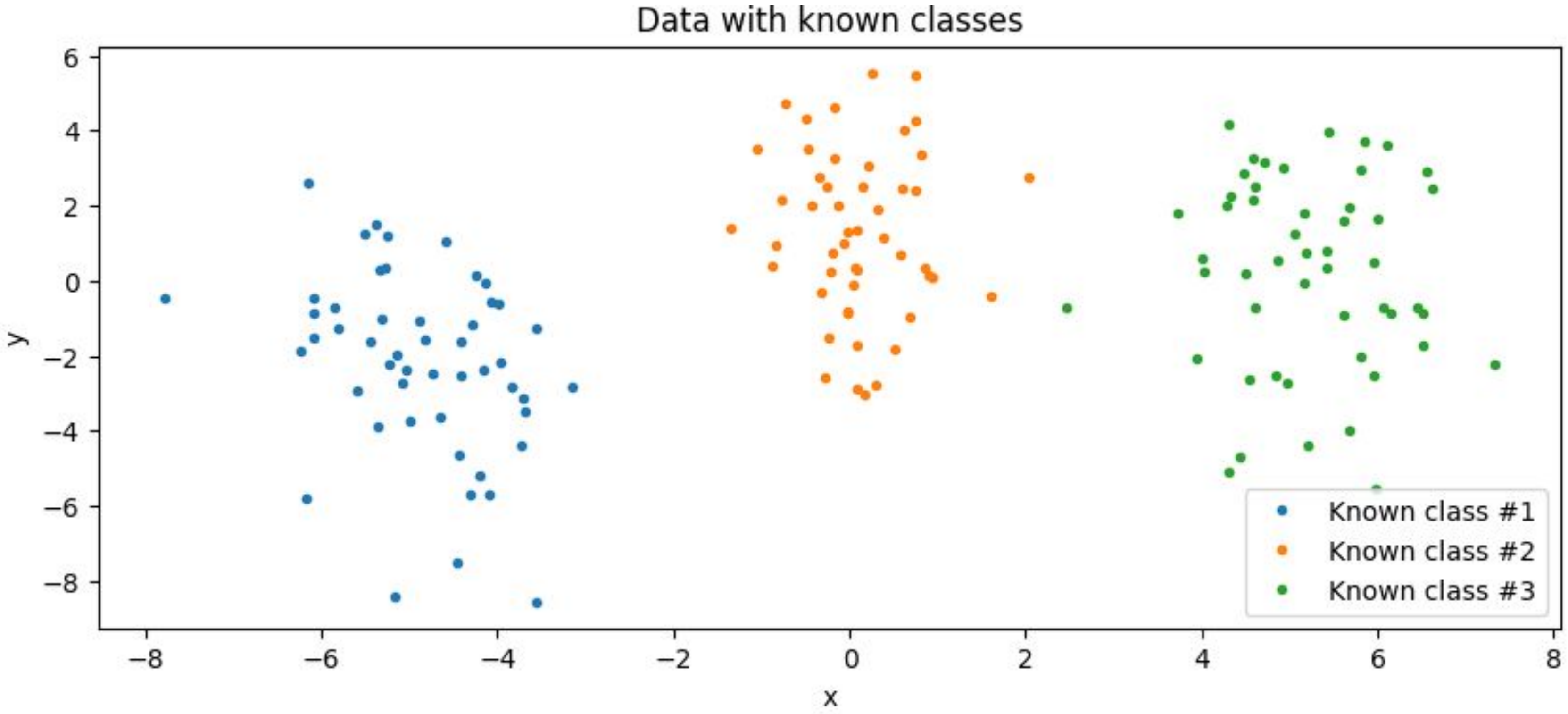
Popular clustering algorithms

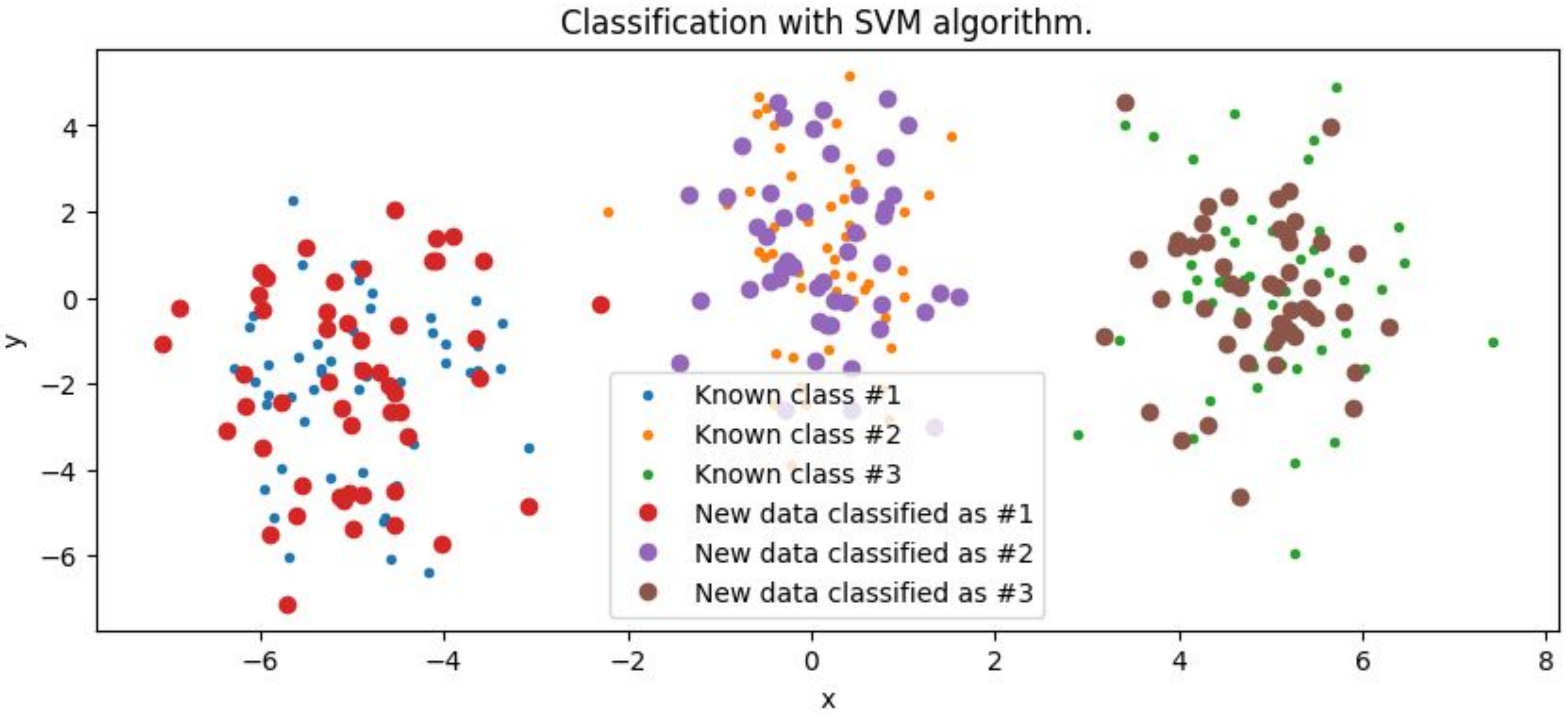
- Kmeans
- DBscan
- Self organizing map (SOM)
- …

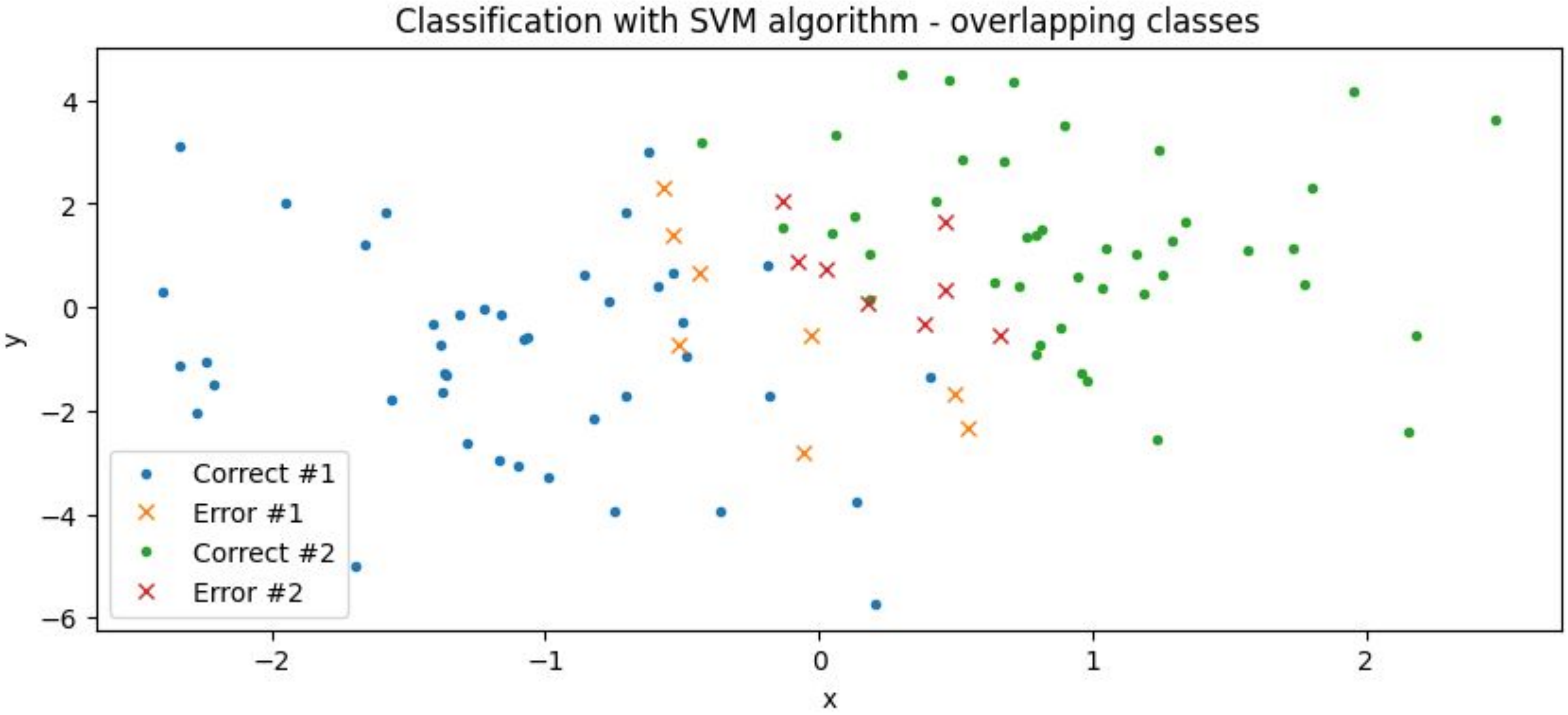
Classification

Classification is a **supervised** learning task in machine learning where the goal is to assign predefined labels or categories to input data based on its features.

- We have defined classes and clear expectations for each.


Data with known classes
6
4
2
0
−2
−4
−6
−8
y
−8
−6
−4
−2
0
2
4
6
8
x
Known class #1
Known class #2
Known class #3


Classification with SVM algorithm.
y
x
4
2
0
−2
−4
−6
−6
−4
−2
0
2
4
6
8
● Known class #1
● Known class #2
● Known class #3
● New data classified as #1
● New data classified as #2
● New data classified as #3


Classification with SVM algorithm - overlapping classes
y
4
2
0
−2
−4
−6
x
−2
−1
0
1
2
● Correct #1
× Error #1
● Correct #2
× Error #2

Popular classification algorithms

- K-Nearest Neighbors
- Support vector machines (SVM)
- Decision trees
- …

Clustering vs classification

- **Classification:** Assigns predefined labels to data points based on their inherent properties.
- **Clustering:** Seeks optimal label groupings within data, uncovering natural patterns.

Clustering vs classification

- **Classification:** We have defined classes and clear expectations for each.
- **Clustering:** We're unsure about cluster shapes or their quantity

Classifier evaluation

Confusion matrix

Confusion matrix is a specific table layout that allows visualization of the performance of an algorithm.

		Predicted condition	
	Total population (P+ N)	Positive (PP)	Negative (PN)
Actual condition	Positive (P)	True positive (TP)	False negative (FN)
	Negative (N)	False positive (FP)	True negative (TN)

Accuracy and Precision

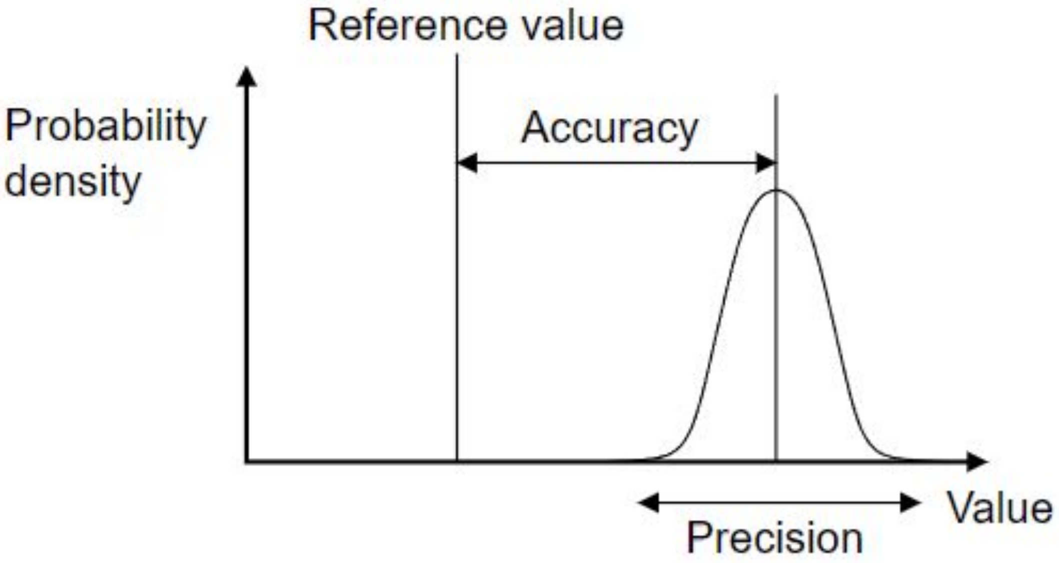
- **Accuracy:**

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{P} + \mathrm{N}}$$

- **Precision:**

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

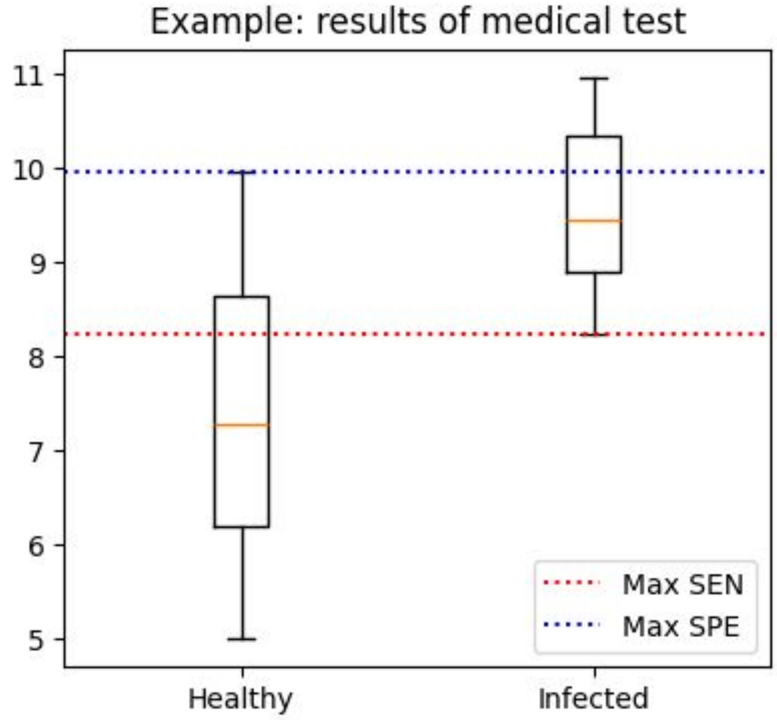
Precision vs Accuracy



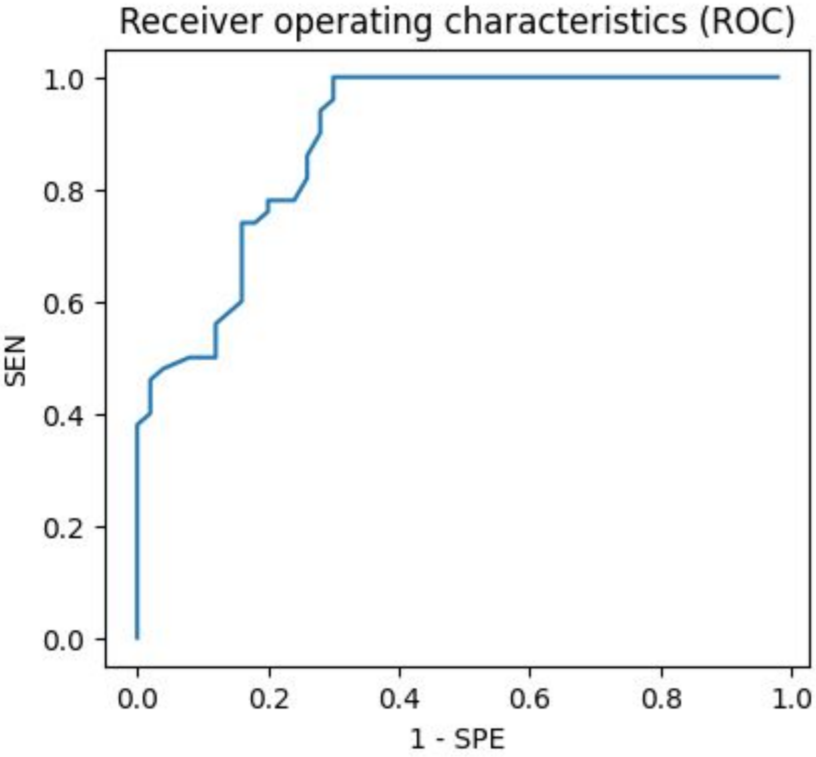
Sensitivity and specificity

- **Sensitivity:**
$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{P}} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$
Also known as: Recall, Hit rate, True positive rate

- **Specificity:**
$$\mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{N}} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}$$
Also known as: Selectivity, True negative rate



Example: results of medical test



Receiver operating characteristics (ROC)

Receiver Operating Characteristics (ROC)

- ROC curve, is a graphical plot that illustrates the performance of a classifier at varying threshold values.
- **AUROC** (area under the ROC curve) is very useful metric

Cross-validation

Validation

- Validation is a method for evaluating a predictive model performance by splitting the dataset into training and testing sets.
- In the ideal scenario, the prediction error on the training set should ideally be equal to or smaller than the training set error.

Cross-validation

Cross-validation is a method for evaluating a predictive model performance by systematically splitting the dataset into training and testing sets multiple times.

- N-fold cross-validation
- Leave-p-out cross-validation

Cross-validation

Cross-validation is a method for evaluating a predictive model performance by systematically splitting the dataset into training and testing sets multiple times.

- N-fold cross-validation
- Leave-p-out cross-validation

Cross-validation

N-fold cross-validation:

You train your model **n** times, each time using a different fold as the test set and the remaining folds as the training set.

Leave-p-out cross-validation:

You train your model multiple times, each time leaving out just **p**-number of data point for testing.

N-fold cross-validation

1. **Split data** into n-folds
2. Use **one fold for testing** and other folds for model training
3. **Repeat step 2** with different folds (train from scratch)
4. **Average the results** over all tested folds

- Folds are chosen at the beginning.
- Folds are not overlapping.

Leave-p-out cross-validation

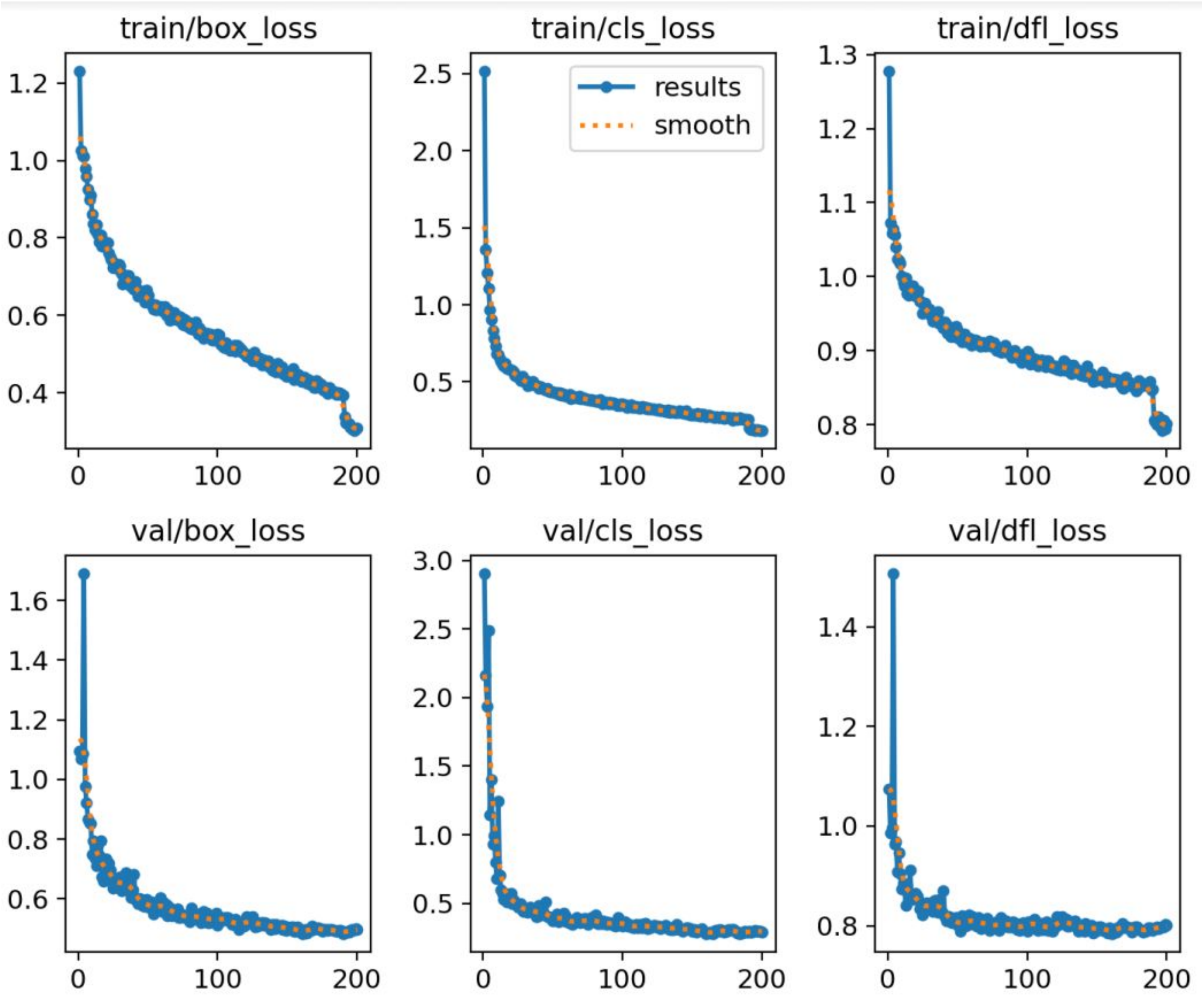
1. **Select p** number of **samples** from data
2. **Use selected samples for testing**, other data for training
3. **Repeat steps 1 and 2** (train from scratch)
4. **Average the results** over all tests

Overtraining

Overtraining (overfitting)

Overtraining, occurs when a model learns the training data too well, including its noise and outliers, to the extent that it negatively impacts the model's performance on new, unseen data.

- Training dataset do not represent the process fully
- We need models that can generalize


train/box_loss
1.2
1.0
0.8
0.6
0.4
0
100
200
train/cls_loss
2.5
2.0
1.5
1.0
0.5
0
100
200
results
smooth
train/dfl_loss
1.3
1.2
1.1
1.0
0.9
0.8
0
100
200
val/box_loss
1.6
1.4
1.2
1.0
0.8
0.6
0
100
200
val/cls_loss
3.0
2.5
2.0
1.5
1.0
0.5
0
100
200
val/dfl_loss
1.4
1.2
1.0
0.8
0
100
200