Machine Learning

Introduction

Karol Przystalski

03.03.2020

Department of Information Technologies, Jagiellonian University

Course overview

Who?

Overview

2016 – now - assistant professor @ Jagiellonian University

2015 – obtained a Ph.D. in Computer Science @ Polish Science Academy and Jagiellonian University

2010 until now – CTO @ Codete

Recent research papers

Multispectral skin patterns analysis using fractal methods, K. Przystalski and M. J.Ogorzalek. Expert Systems with Applications, 2017

`https://www.sciencedirect.com/science/article/pii/S0957417417304803`

Contact

kprzystalski@gmail.com

Skype: kprzystalski

Who?

Overview

Obtained the professor degree in physics in 2013, with a specialty in statistical mechanics, and classical field theory.

Contact piotr.bialas@uj.edu.pl
room C-2-28

When?

- 03.03 – K Introduction to Machine Learning
- 10.03 – K Clustering
- 17.03 – K Linear methods
- 24.03 – K Decision Trees
- 31.03 – K Ensemble methods
- 14.04 – K SVM
- 21.04 – K xAI
- 28.04 – K Natural Language Processing
- 05.05 – P Probability-based methods I
- 12.05 – P Probability-based methods I
- 19.05 – P Probability-based methods III
- 26.05 – P Neural Networks I
- 02.06 – P Neural Networks II
- 09.06 – P Recommendation systems

How to pass?

The course consist of fifteen lectures and laboratories and will be given entirely in English. Each lab ends with a exercise (homework) that needs to be finished by the student within two weeks. Final score is an average of each exercise. You can fail one exercise or pass the deadline by two exercises.

The exam is set of 40 multiple choice questions. To pass the exam you need to answer correctly for 28 questions (70%).

Introduction

AI is the new electricity

AI is the new UI

Buzzwords, buzzwords and more buzzwords

Machine learning became a buzzword a few years ago. Like deep learning, blockchain or data science, each buzzword is often used by startup to show the innovative approach.

There are many projects/challenges where machine learning shouldn't be the solution or at least shouldn't be the first choice.

The hype

Forty percent of "AI startups" in Europe don't actually use AI

The State of AI 2019: Divergence

Startups labelled as being in AI attract 15% to 50% more funding than other technology firms.

The State of AI 2019: Divergence

AI

If it's written in PowerPoint, it is definitely **Artificial Intelligence**. However, if it's written in Python/R/Scala/whatever, it is probably **Machine Learning**. ML is just one of the attempts to achieve AI – the best we currently have, but surely not good enough to reach it at any point.

Many forms of Government have been tried, and will be tried in this world of sin and woe. No one pretends that democracy is perfect or all-wise. Indeed it has been said that democracy is the worst form of Government except for all those other forms that have been tried from time to time.

Winston Churchill

Tools

There are a few levels of tools:

- language,
- math libraries,
- general ML tools,
- shallow method libraries,
- neural network libraries,
- deep learning frameworks,

There are also online API solution

Hardware

There are several solutions available to decrease the time of calculation like:

- GPU on local machines or neural sticks – CUDA,
- GPU on the cloud – AWS, GCP, Azure,
- TPU,
- Nvidia boxes.

Use cases

Machine learning (or AI) has many use cases in the **process automation** in the fields like:

- security,
- medical diagnosis,
- customer service,
- financial analytics - i.e. risk management, insurance prediction,
- blockchain,
- self-driving cars,
- test automation,
- and many more.

Roles

At Netflix there are many roles related to data:

- Business Analyst,
- Data Analyst,
- Quantitative Analyst,
- Algorithm Engineer,
- Analytics Engineer,
- Data Engineer,
- Data Scientist,
- Machine Learning Scientist,
- Research Scientist.

Data Analyst

Data Analyst responsibilities and skills:

- data wrangling,
- basic descriptive statistics,
- data visualization,
- SQL experience,
- knowledge of R/Python.

Machine Learning Scientist

Machine Learning Scientist responsibilities and skills:

- using ML algorithms to utilize data, learn from it and forecast future,
- data m odelling and evaluation,
- probability and statistics knowledge,
- programming skills.

(Big) Data Engineer

(Big) Data Engineer responsibilities and skills:

- building infrastructure and architecture for Big Data,
- using databases,
- designing large-scale processing systems,
- integrate different data sources into Data Lake,
- knowledge of Hadoop ecosystem: HDFS, Spark, Hive, Kafka, Druid, etc.
- data importing.

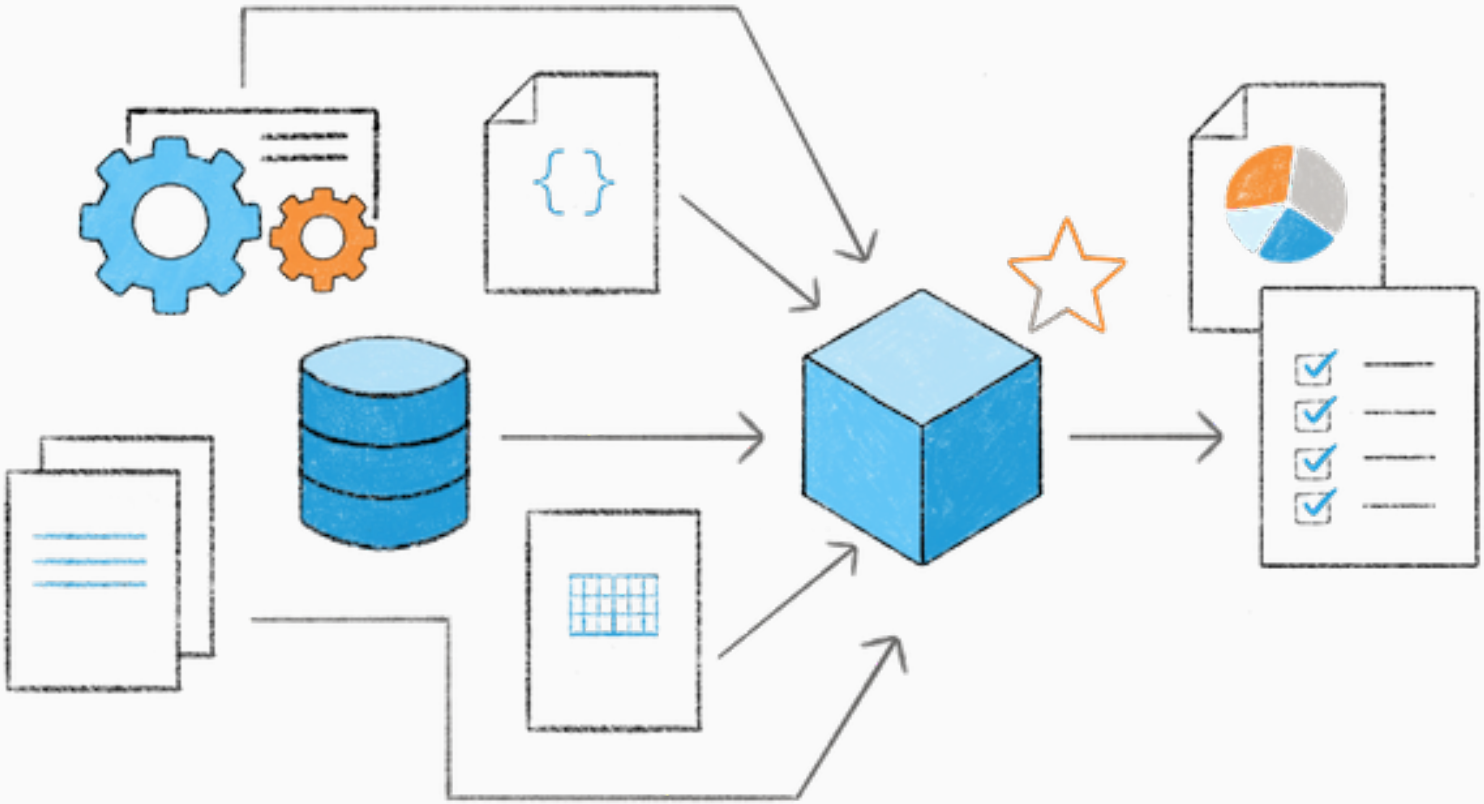
Data Scientist

Data Scientist responsibilities and skills:

- business and data understanding,
- statistical modelling and machine learning,
- reporting and visualization,
- and many more ....

Failures

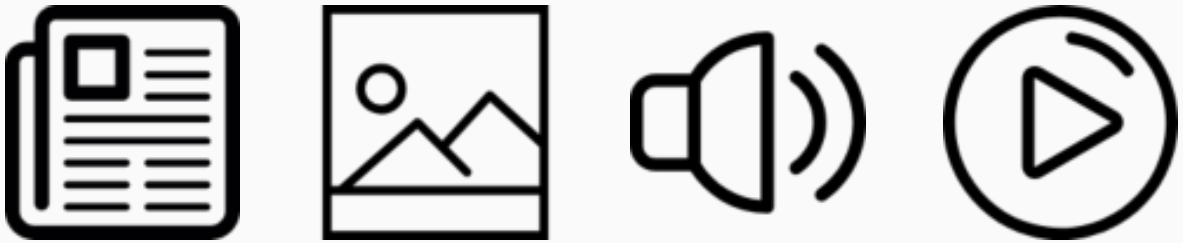
Data



Structured vs. unstructured

Structured data:

Unstructured data:



A real-world story

Actors:

- Sweet little girl and little red cap
- Her grandmother
- Big bad wolf

Goal:

- A piece of cake and a bottle of wine to be delivered

A real-world story

How to distinguish grandma from a wolf?

A real-world story

There is a well-known example of a Machine Learning system designed for classifying the images of wolves and huskies. The system gives about 90% accuracy.

Failure #1: Overfitting and/or data set issue

It's a well-known problem where the data set are now prepared/cleaned up properly. In the example of wolves vs. huskies, the system takes as the main feature the background, because the wolves photos are taken usually in the wild, usually with snow in the background.

(a) Husky classified as wolf

(b) Explanation

Failure #2 The Process

It's hard to combine data science projects into a Scrum model. There are many problems that needs to be solved, one of the most problematic is to divide the tasks properly. Properly means:

- avoid setting tasks where we use fixed values of quality metrics,
- use specific metrics, usually more just one, avoid using accuracy,
- divide tasks into data acquisition, preprocessing, model strategy, and quality metrics.

Perform due diligence before stepping into a ML/DS projects.

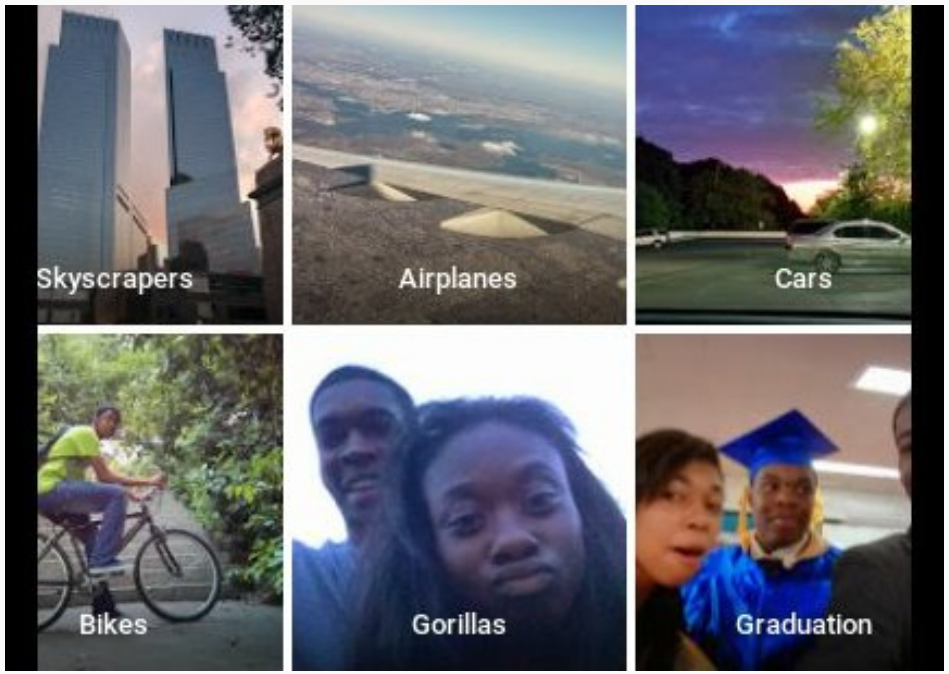
Failure #3 Set a strategy

A company using machine learning methods doesn't make your company an AI company. There are many challenges that needs to be fulfilled to become an AI company. Just to point out a few challenges:

- data acquisition strategy,
- unified data storage,
- pervasive automation,
- setup new data roles structure.

Failure #4 Use proper data

The data is crucial and using a data set that isn't prepared properly, it might end with such cases like the Google one.



Failure #5 Security

The retraining part of the machine learning process is a good approach, but without some additional supporting tools it can be a fail.

Other security issue related to machine learning are adversarial attacks.

Taxonomy

History of AI

Before **1950** Bayes theorem, Markov chains, (...)

1957 Rosenblatt's Perceptron

1967 Nearest Neighbor

1985 Sejnowski's NetTalk

1986 Backpropagation

1989 Reinforcement learning

1995 Random forest

1995 SVM

1997 Deep Blue beats Kasparov

1998 MNIST database

2006 Deep learning

2006 Netflix challenge

2010 Kaggle

2011 Watson beats Jeopardy competitors

2012 Google Xlab

2015 Stephen Hawking, Elon Musk and al. letter

Definition

"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."

Tom M. Mitchell

"A branch of artificial intelligence in which a computer generates rules underlying or based on raw data that has been fed into it."

Dictionary.com

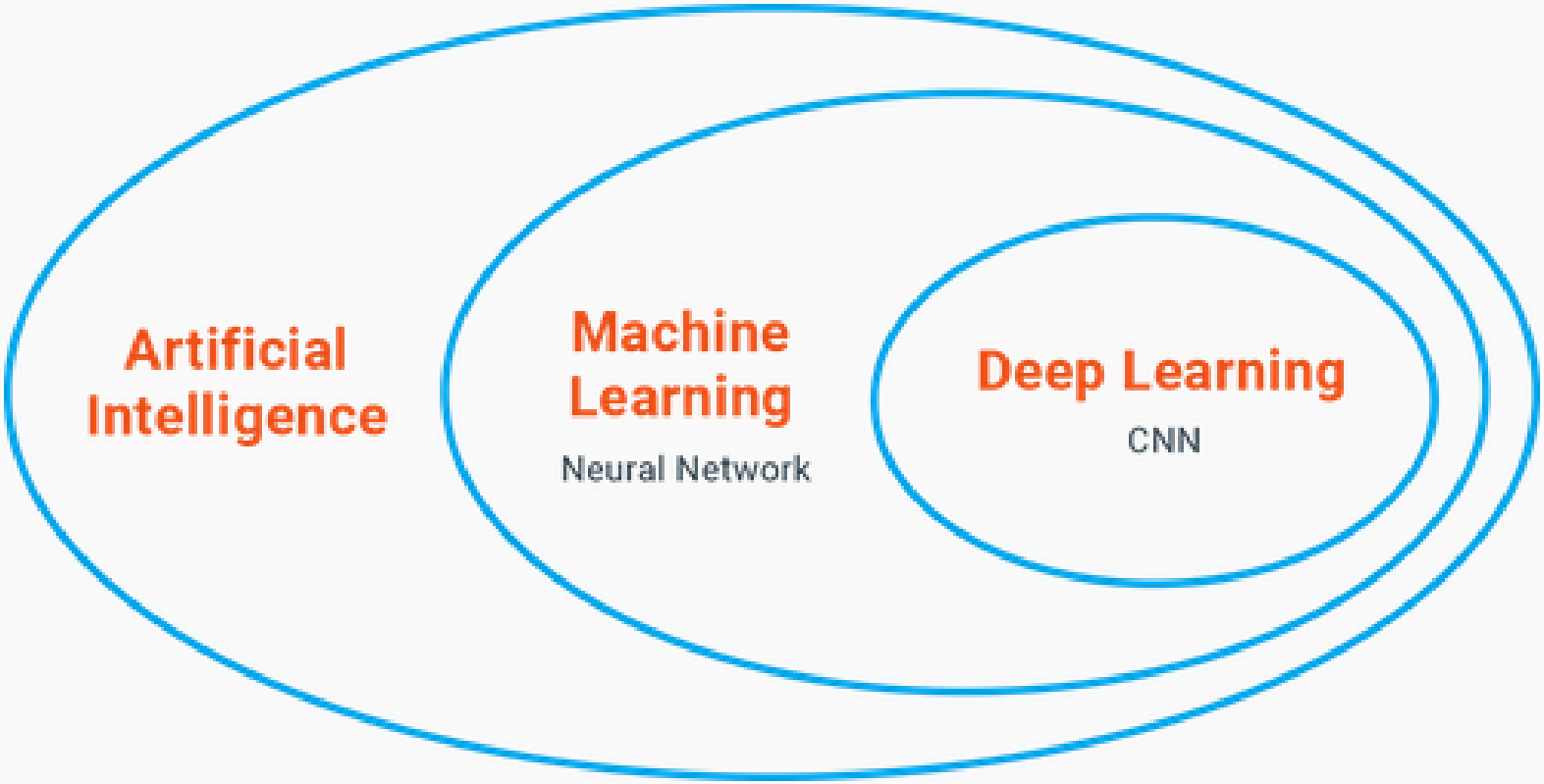
"Machine learning is the training of a model from data that generalizes a decision against a performance measure."

Jason Brownlee

"Machine learning is a scientific discipline that is concerned with the design and development of algorithms that allow computers to evolve behaviors based on empirical data, such as from sensor data or databases."

Wikipedia

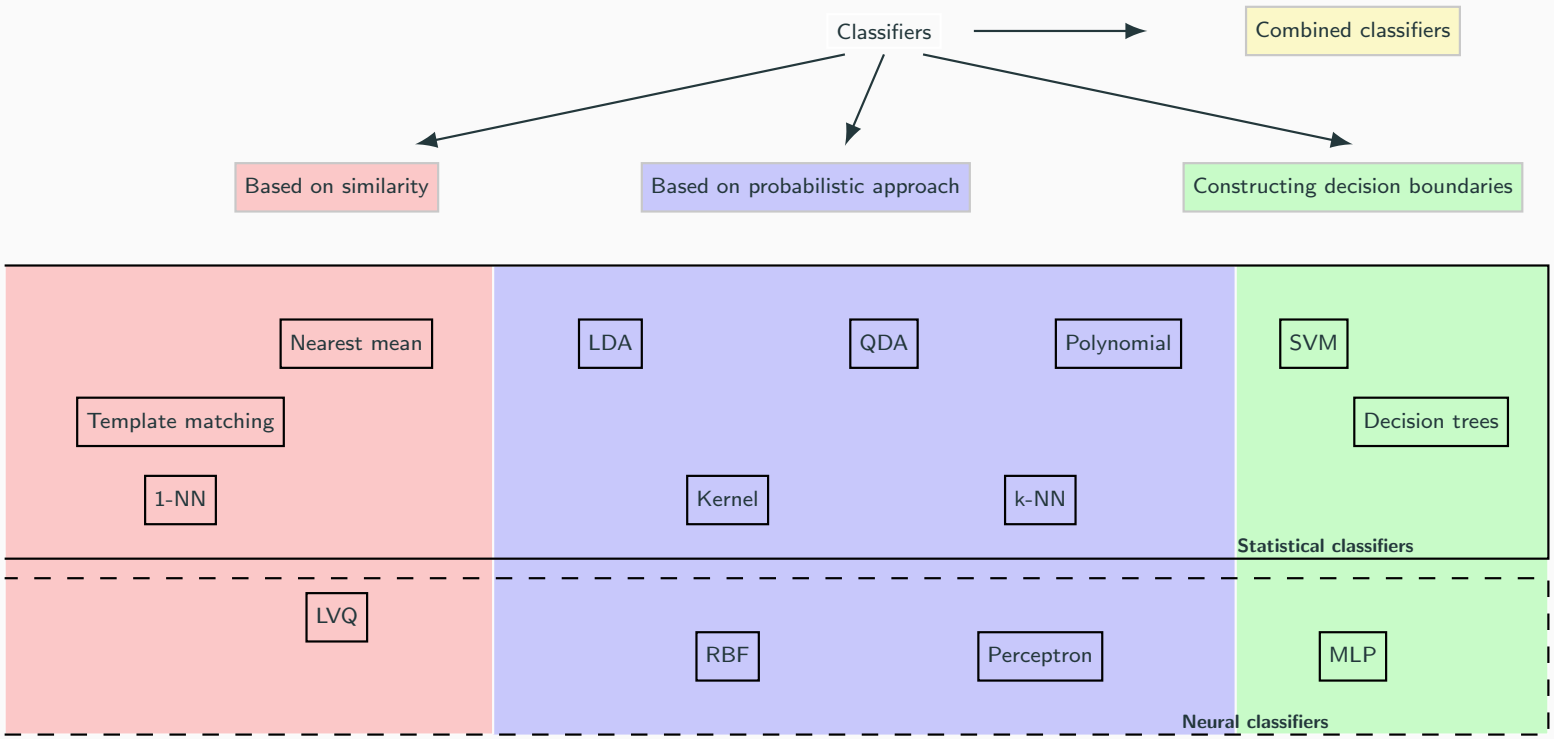
AI


Artificial Intelligence
Machine Learning
Neural Network
Deep Learning
CNN

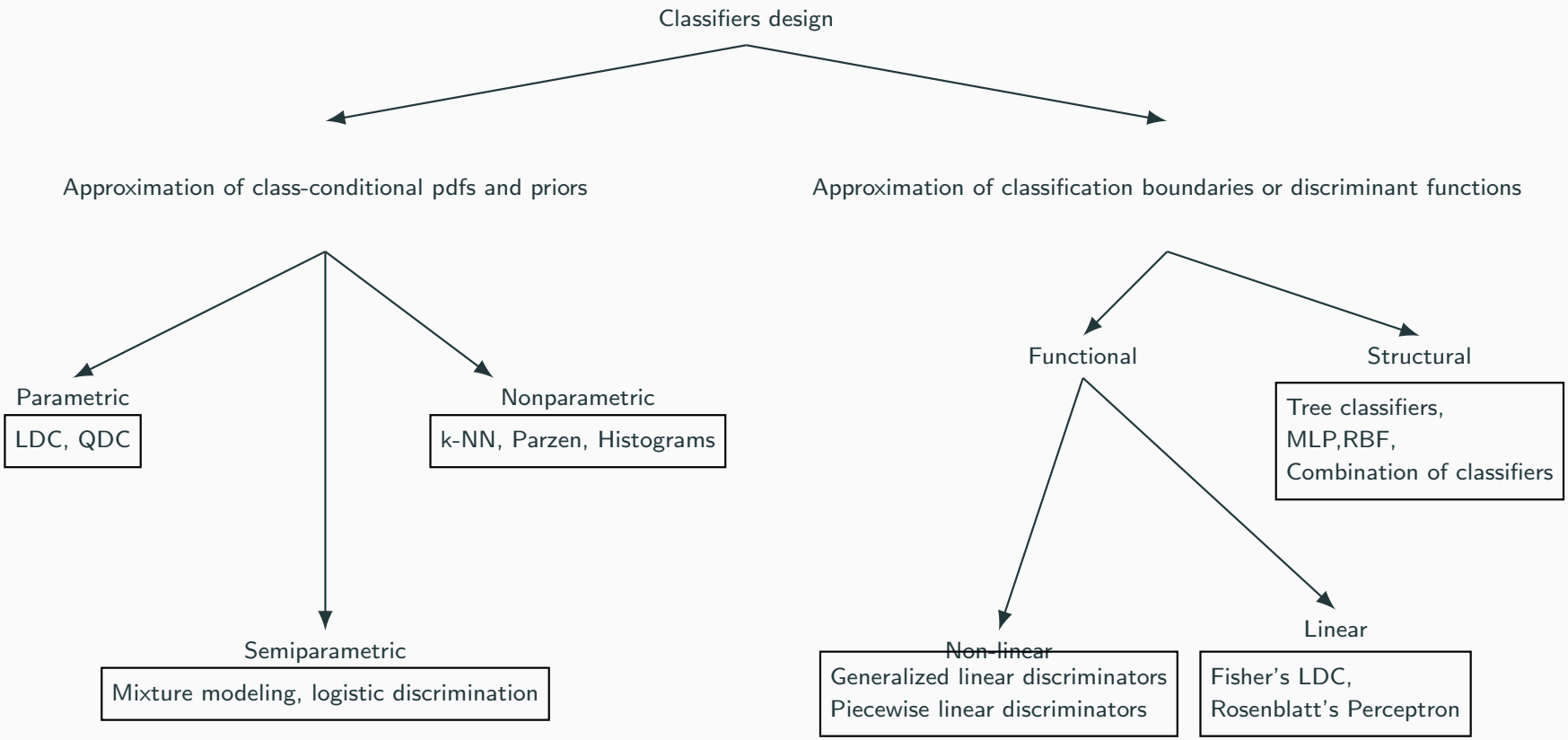
Taxonomy

1. Is this A or B? **Classification**
2. Is this weird? **Anomaly detection**
3. How much / how many ? **Regression**
4. How is this organized? **Clustering**
5. What should I do next? **Reinforcement learning**

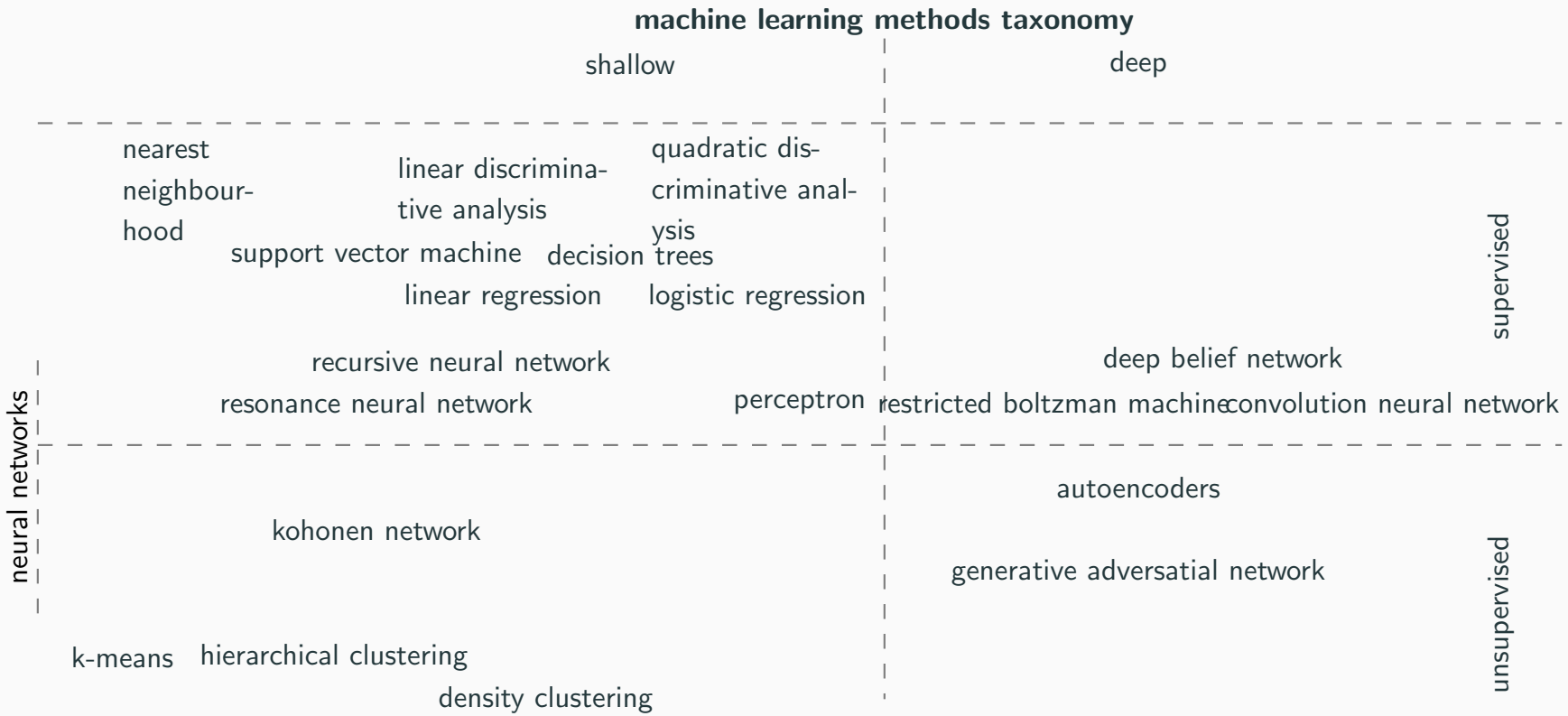
Taxonomy

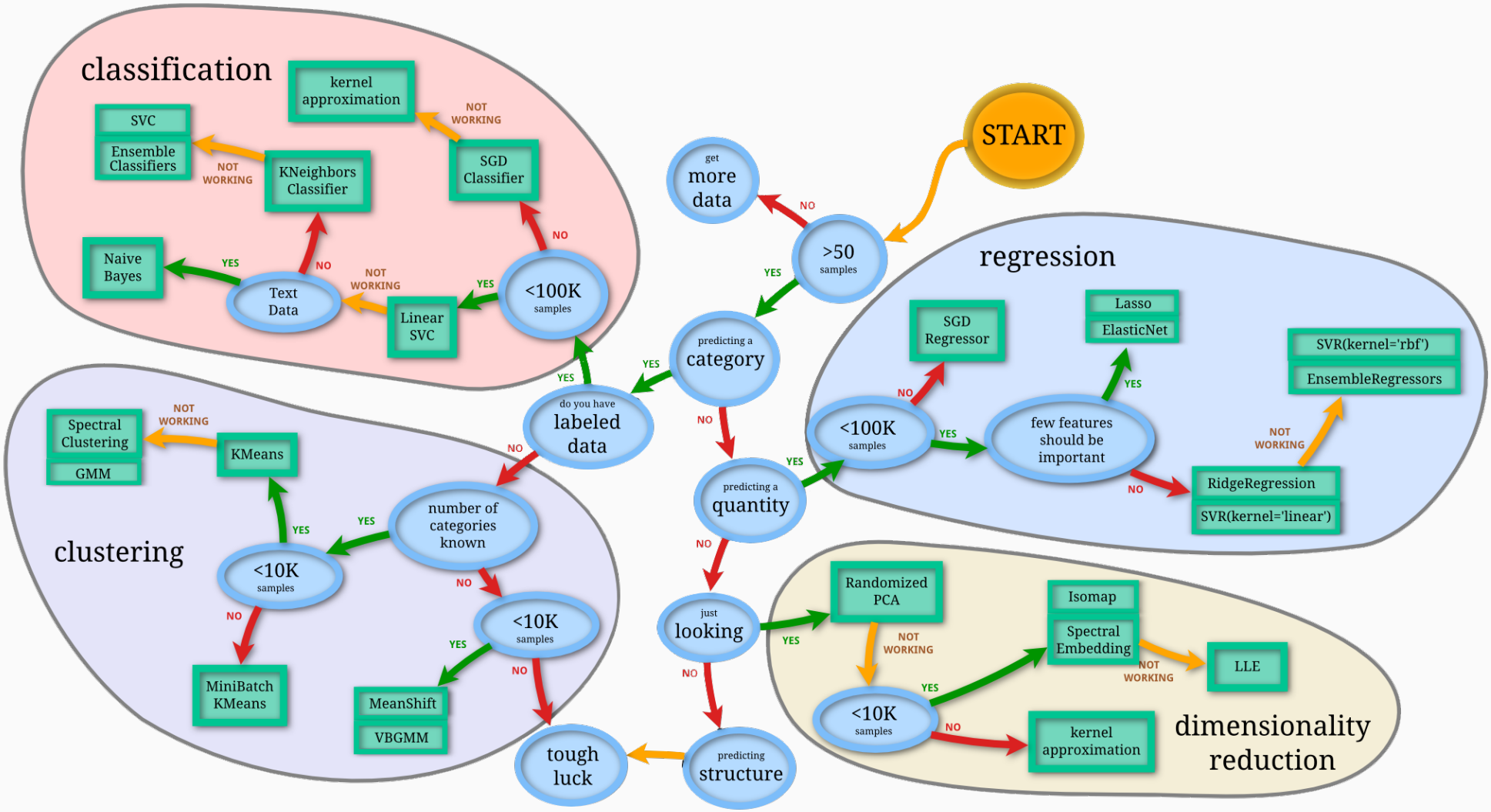

Classifiers
Combined classifiers
Based on similarity
Based on probabilistic approach
Constructing decision boundaries
Nearest mean
LDA
QDA
Polynomial
SVM
Template matching
Decision trees
1-NN
Kernel
k-NN
Statistical classifiers
LVQ
RBF
Perceptron
MLP
Neural classifiers

Taxonomy

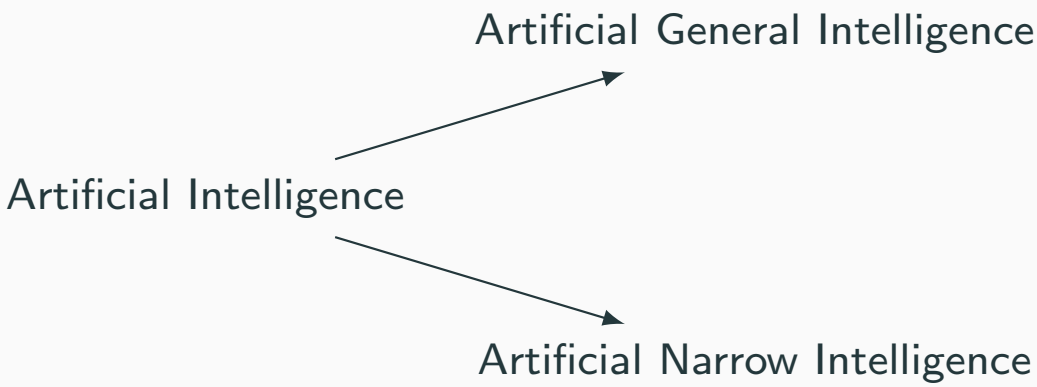

Classifiers design
Approximation of class-conditional pdfs and priors
Approximation of classification boundaries or discriminant functions
Parametric
LDC, QDC
Nonparametric
k-NN, Parzen, Histograms
Semiparametric
Mixture modeling, logistic discrimination
Functional
Structural
Tree classifiers,
MLP,RBF,
Combination of classifiers
Non-linear
Generalized linear discriminators
Piecewise linear discriminators
Linear
Fisher's LDC,
Rosenblatt's Perceptron

Taxonomy II

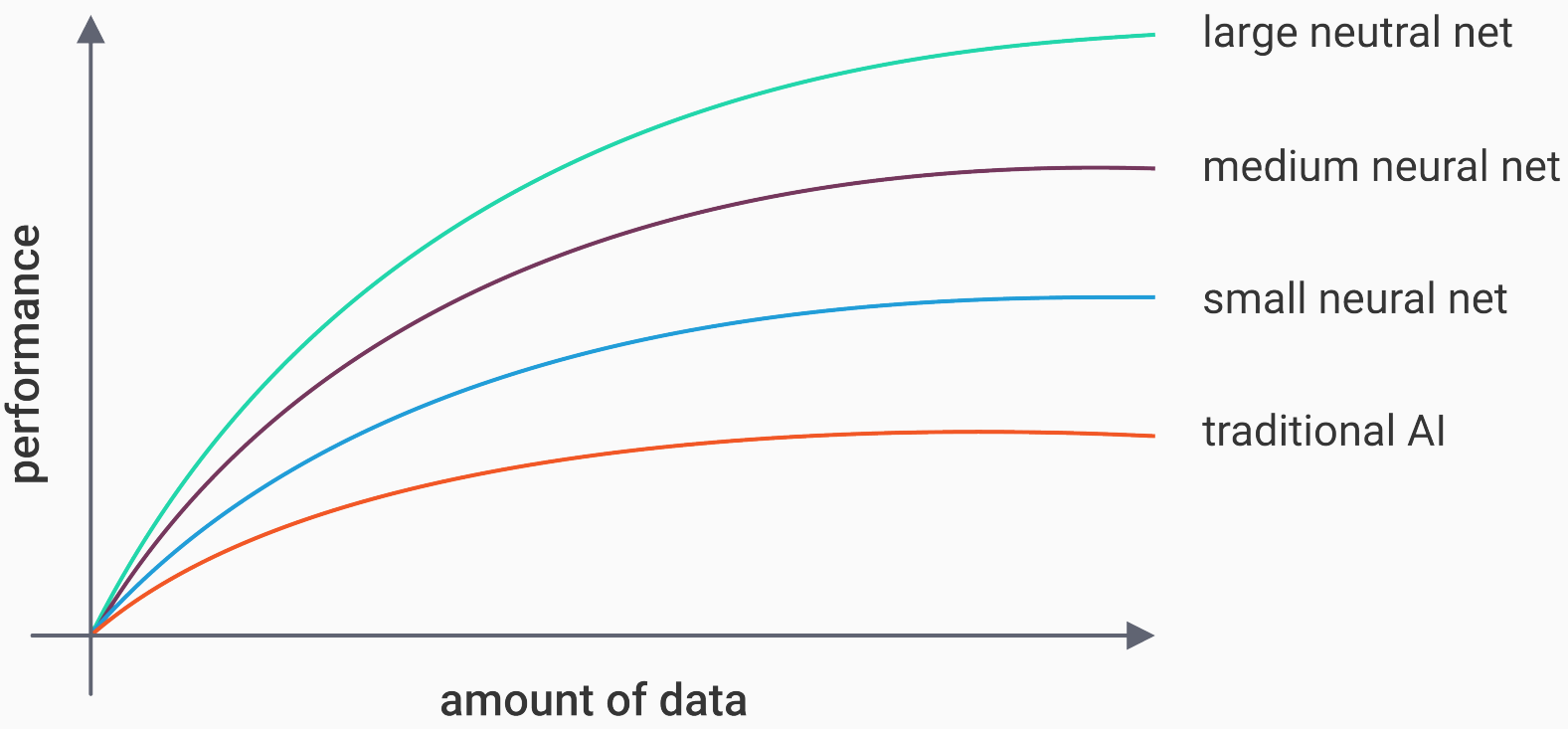

machine learning methods taxonomy
shallow
deep
nearest neighbour-hood
linear discrimina-tive analysis
quadratic dis-criminative anal-ysis
support vector machine
decision trees
linear regression
logistic regression
supervised
recursive neural network
deep belief network
resonance neural network
perceptron
restricted boltzman machine
convolution neural network
neural networks
autoencoders
kohonen network
generative adversatial network
unsupervised
k-means
hierarchical clustering
density clustering


START
get more data
NO
>50 samples
YES
predicting a category
YES
do you have labeled data
YES
<100K samples
NO
Naive Bayes
YES
Text Data
NO
KNeighbors Classifier
NOT WORKING
SVC
Ensemble Classifiers
NOT WORKING
kernel approximation
NOT WORKING
SGD Classifier
NO
NOT WORKING
Linear SVC
YES
classification
NO
number of categories known
YES
<10K samples
YES
KMeans
NOT WORKING
Spectral Clustering
GMM
NO
MiniBatch KMeans
NO
<10K samples
YES
MeanShift
VBGMM
NO
clustering
NO
predicting a quantity
YES
<100K samples
YES
few features should be important
YES
Lasso
ElasticNet
NO
SGD Regressor
NOT WORKING
SVR(kernel='rbf')
EnsembleRegressors
NO
RidgeRegression
SVR(kernel='linear')
regression
NO
just looking
YES
Randomized PCA
NOT WORKING
<10K samples
YES
Isomap
Spectral Embedding
NOT WORKING
LLE
NO
kernel approximation
dimensionality reduction
NO
predicting structure
tough luck

AI taxonomy


Artificial General Intelligence
Artificial Intelligence
Artificial Narrow Intelligence

Machine and deep learning



Taxonomy summary

Supervised – labeled data set; method is learning based on labeled data and assign a label when classifying

Reinforcement learning – the method is learning on unlabeled data sets, but got penalties or reward - depends on the classification result

Unsupervised – unlabeled data set; also known as clustering; method is trained and tested on unlabeled data sets

Deep learning – a group of methods that are based on deep neural networks

Quality metrics

Features importance

During the training we can face two commonly known issues. The first one is related to the number of features. It is hard to say what is the best number of features to use as it depends on the problem that needs to be solved. There are several methods that can measure the importance of each feature, so we can choose only the most important.

Important means here how a feature affects the accuracy.

Features importance

e have several common used feature selection methods that can be divided into four major groups:

- ranking,
- wrapper,
- hybrid,
- embedded.

Features importance

Ranking methods are also known as filter methods. In this type of feature selection methods we do not use classification method. The goal is to set a ranking of features. Pearson correlation method is one of such a method.

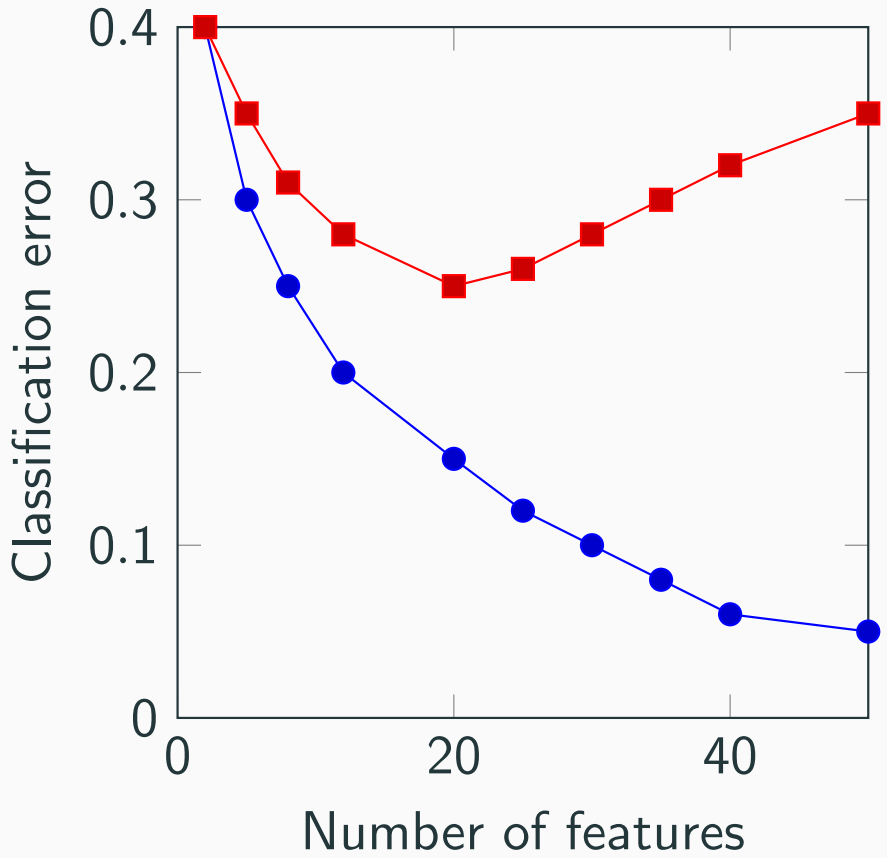
The name of the second group of methods is such because the feature selection method **wraps** the classification method and based on the accuracy selects the features. It takes a set features and do the classification, gets the accuracy and compare it to the accuracy got from a data set of different feature set.

Features importance

Hybrid methods are a mix of ranking and wrapper methods. In this type of methods we use ranking part in the first place and next the wrapper part. It makes a huge difference if we have a lot of features.

The last type is **embedded** into a classification method. It means that it is not used outside the machine learning method, but it is a part of it. An often used approach are genetic algorithms used within the classifier.

Features importance



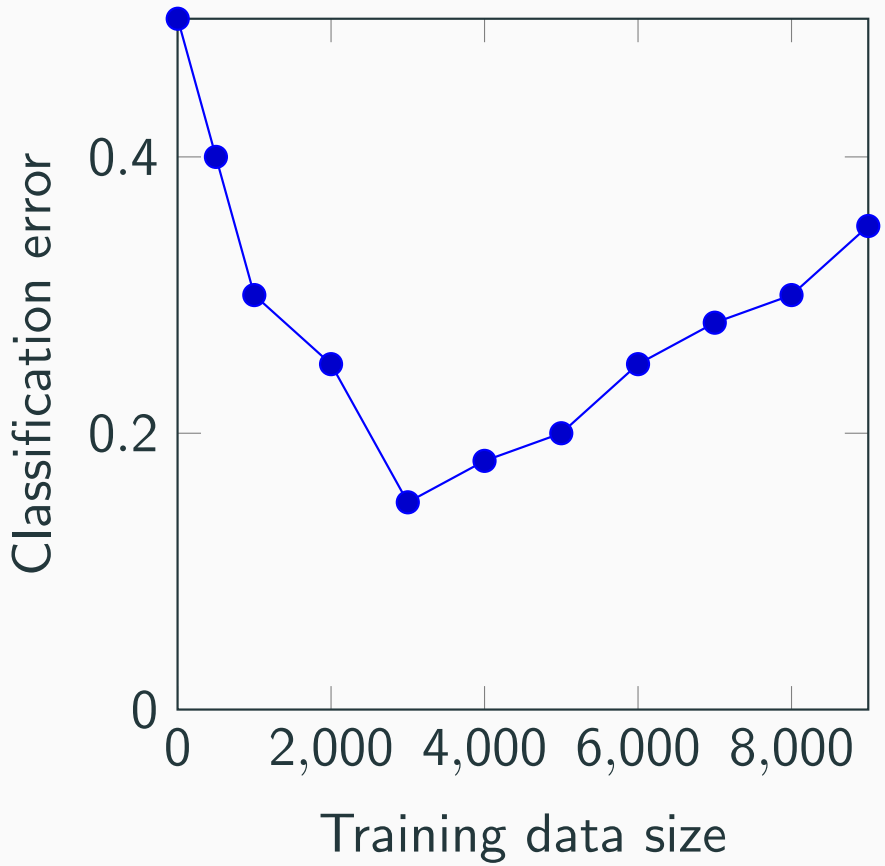
Overfitting

Overfitting is another common problem that can happen while training a classification method. It is about under-training and over-training. The goal of the training part is to generalize. It means that we would like to have a method that gives high accuracy for any data of a given problem.

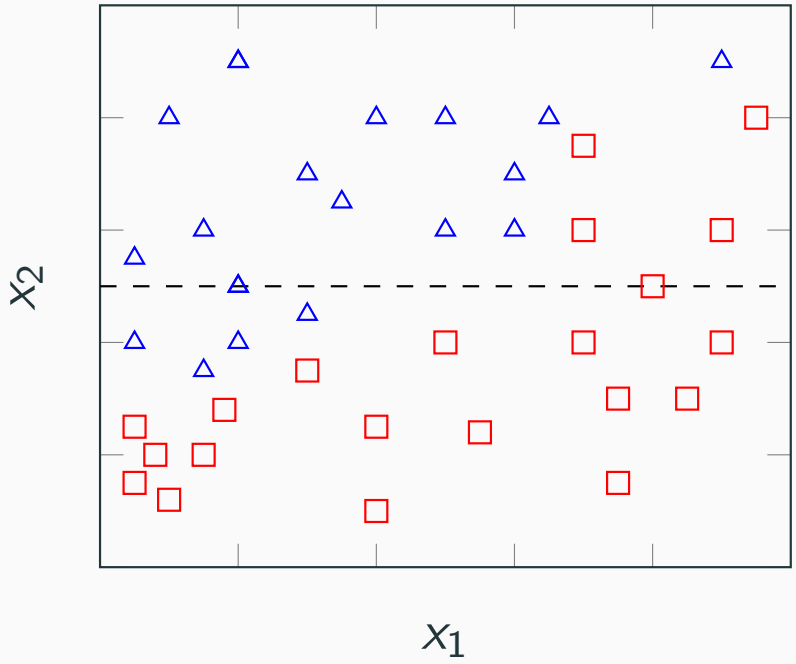
Under-training happens when we do not train the method enough. We have not prepared enough training data and the method does not have the data to train on. Therefore, the method gives lower accuracy, because it assigns labels wrongly.

The same result we get with **over-training**, but the reason is slightly different. We train the method with too many data. Especially when we have a lot of features values of the same label that are close to each other.

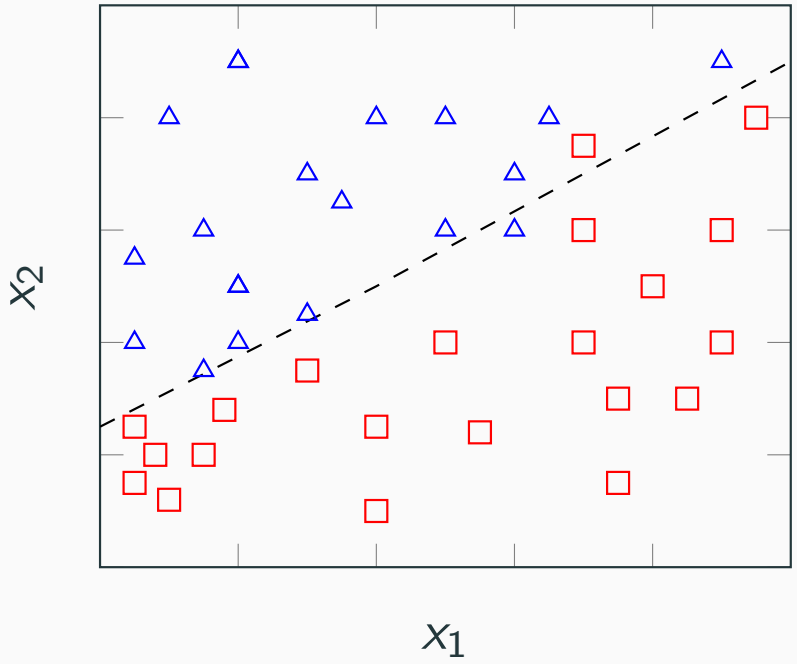
Overfitting



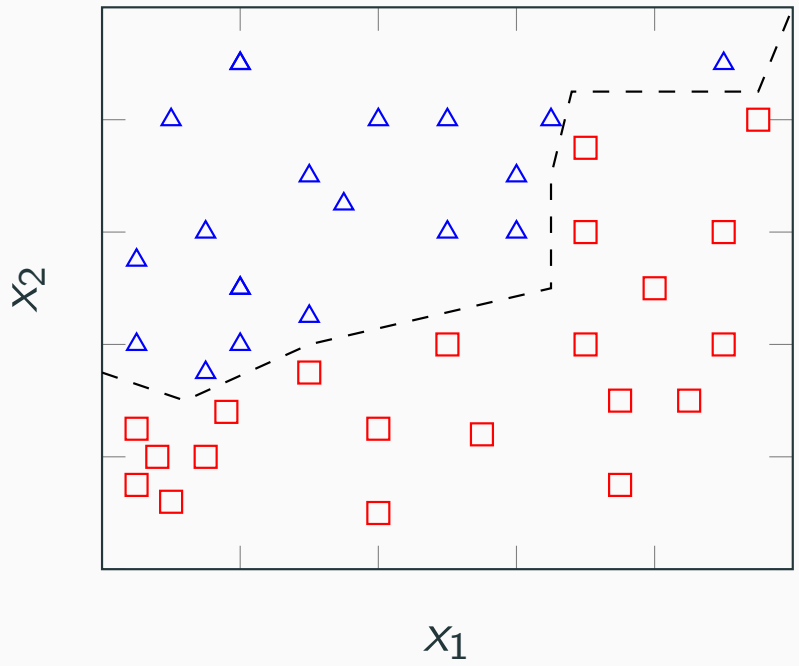
Overfitting – underfit



Overfitting – good fit



Overfitting – overfit



Data set preparation – error rate

One of the common problem that each data scientist has is about how to divide the data set into training and testing data sets. To understand the following equations we need to introduce new designations. Let $\mathcal{L}_n$ be our training data set of size n, T_m our testing data set of size m, M_e the number of misclassified cases, $\mathcal{I}$ a function that return 1 if there is a match between predicted and label value and $e(d)$ the error rate of classifier d.

We can write the error rate like following:

$$e(d) = \frac{M_e}{m}. \tag{1}$$

Data set strategies

It is the opposite to accuracy that is described later in this section. Error rate can be calculated differently depending on which method of data set preparation is used. There are few commonly used approached of how we can handle the training, testing and validation data sets:

- resubstitution – R-method,
- hold-out – H-method,
- cross-validation – π-method,
- bootstrap,
- jackknife.

Data set strategies

The first method is a very simple one. We have the same data set for training and testing. It is not the best solution if we consider to have a solid classifier. The error rate can be written as following:

$$e_R(d) = \frac{1}{n} \sum_{j=1}^{n} \mathcal{I}(d(X_j; \mathcal{L}_n) \neq Y_j). \tag{2}$$

The second method is about dividing a data set into two data sets. It can be divided by half or other proportions. One set is our training data set and the second training data set. We can swap those sets and calculate the average of both sets. The error rate can be calculated as following:

$$e_\tau(\hat{d}) = \frac{1}{m} \sum_{j=1}^{m} I(\hat{d}(X_j^t; \mathcal{L}_n \neq Y_j^t). \tag{3}$$

Compared to resubstitution method it uses the testing data set only.

Data set strategies

Cross-validation is the most common approach. It can be also called as rotation method. We need to divide the data set into k subsets. The elements in each set are randomly chosen. One of those sets are taken as a testing set where the other sets are merged into a training set. It should be repeated k times for each k subset. The error rate can be calculated like following:

$$e_{CV}(d) = \frac{1}{n} \sum_{j=1}^{n} I(\hat{d}(X_j; \mathcal{L}_n^{(-j)} \neq Y_j). \tag{4}$$

Bootstrap method can be considered as an extension of resubstitution. The goal is to generate multiple sets from the main set by randomly selection. We use resubstitution method on each set and calculate an average error at the end:

$$e_B(d) = \frac{1}{B} \sum_{b=1}^{B} \frac{\sum_{j=1}^{n} \mathcal{I}(Z_j \notin \mathcal{L}_n^{\star b}) \mathcal{I}(d(X_j; \mathcal{L}_n^{\star b}) \not= Y_j)}{\sum_{j=1}^{n} (Z_j \notin \mathcal{L}_n^{\star b})}. \tag{5}$$

Metrics

There are several metrics to show the quality of our classification model:

- ROC that stands for Receiver Operating Characteristic curve,
- AUC – Area Under Curve,
- F_1 score,
- Precision,
- Recall.

Output/prediction matrix

		True condition	
		condition positive	**condition negative**
Predicted	**positive**	True Positive (TP)	False Positive (FP)
	negative	False Negative (FN)	True Negative (TN)

Metrics

It can be calculated like following:

$$ACC = \frac{\#TP + \#TN}{\#TP + \#TN + \#FP + \#FN}. \tag{6}$$

First one that we describe is called True Positive Rate (TPR). It can be calculated like following:

$$TPR = \frac{\#TP}{\#TP + \#FN}. \tag{7}$$

TPR is also called sensitivity or recall and is a measure of good predictions within a set of cases.

Metrics

By $\#TP, \#FP$ we mean the number of True Positive and False Positive cases. An opposite to it is specificity. It is also called TNR what stands for True Negative Rate. It can be calculated as following:

$$TNR = \frac{\#TN}{\#TN + \#FP}. \tag{8}$$

Another important metric is precision that is also known as Positive Predictive Value (PPV):

$$PPV = \frac{\#TP}{\#TP + \#FP}. \tag{9}$$

The opposite to it is the Negative Predictive Value:

$$NPV = \frac{TN}{TN + FN}. \tag{10}$$

Metrics

It is about how bad we are on predicting positive cases:

$$FPR = 1 - TNR. \tag{11}$$

The opposite to FPR is False Negative Rate:

$$FNR = 1 - TPR. \tag{12}$$

Another popular metric is called F_1 score and it is a weighted accuracy measure. It takes PPV and TPR to calculate the score:

$$F_1 = 2\frac{PPV \cdot TPR}{TPR + PPV}. \tag{13}$$

The F_1 value as in case of all previous metrics between 1 and 0, where 1 is the best.

Metrics

A interesting measure is Matthews Correlation Coefficient measure that is about the correlation between observed and predicted values. The value of MCC is between -1 and 1. If we have a perfect classifier we get MCC=1. A random classifier is when we have MCC=0 and a totally bad classifier if have MCC=-1. This measure can be calculated as following:

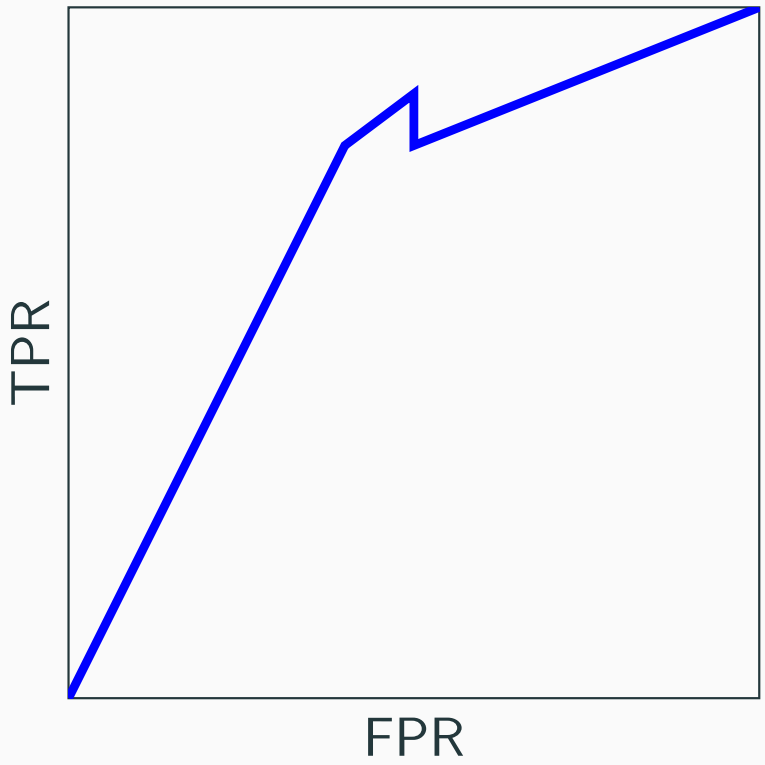
$$MCC = \frac{\#TP \cdot \#TN - \#FP \cdot \#FN}{\sqrt{(\#TP + \#FP)(\#TP + \#FN)(\#TN + \#FP)(\#TN + \#FN)}}. \tag{14}$$

ROC

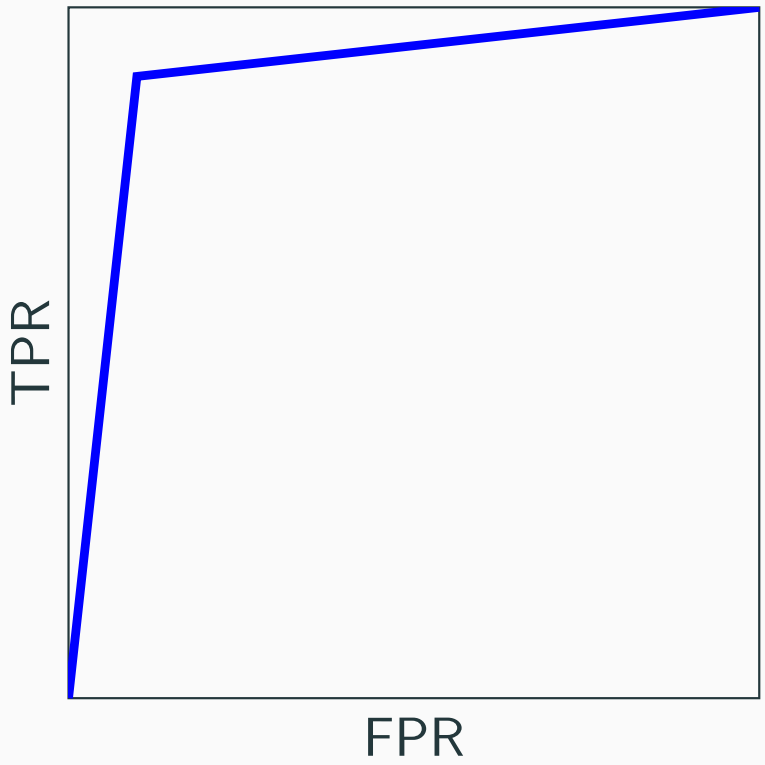
We can take the previous example to explain the AUC metric. To calculate it we need TPR and FPR metrics for each cut point. Area under curve is a part of ROC curve and is just the surface area under the curve.

value	**classifier quality**
1.0	perfect
$0.99 - 0.9$	excellent
$0.89 - 0.8$	very good
$0.79 - 0.7$	good
$0.69 - 0.51$	poor
0.5	worthless

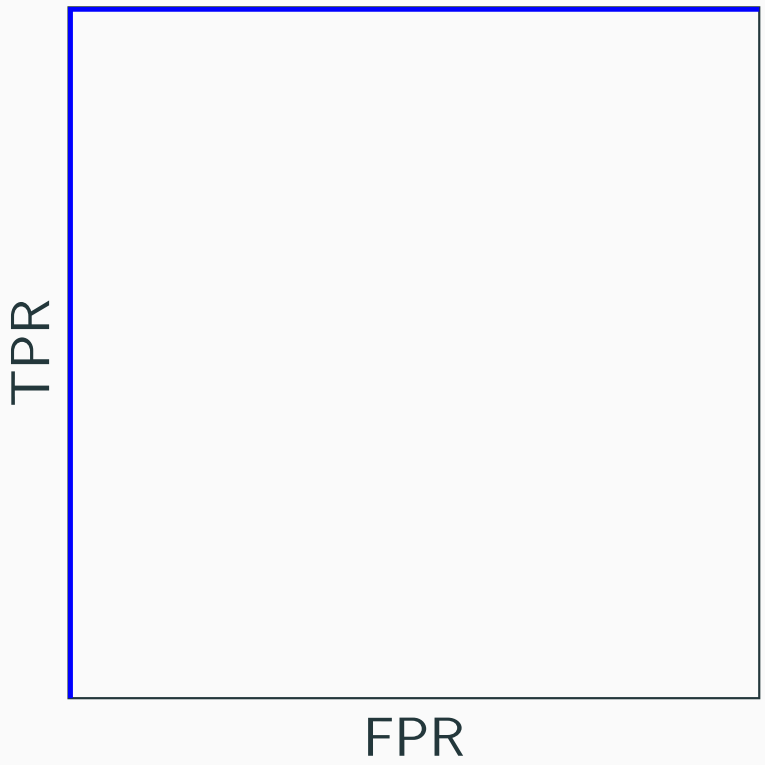
ROC



ROC



ROC



Good reads

Recommended books to read:

1. Metody klasyfikacji obiektow w wizji komputerowej, K. Stapor, PWN 2011,
2. Bayesian Probability Theory, U. von Toussaint, Volker Dose, W. von der Linden, Cambridge University Press 2014,
3. Metody i techniki sztucznej inteligencji, L. Rutkowski, PWN 2009,
4. Mathematical classification and clustering, B. Mirkin, Kluwer 1996,
5. Data Mining and Knowledge Discovery Handbook (chapter 9), L. Rokach, O. Maimon, Springer 2010,
6. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods, N. Cristianini, J. Shawe-Taylor, Cambridge University Press 2000,
7. Sieci neuronowe, R. Tadeusiewicz, Akademicka Oficyna Wydawnicza RM 1993,
8. Machine learning, S. Marsland, CRC 2015.

Questions?