Zrównoleglanie algorytmów uczenia maszynowego

Sztuczna inteligencja w środowisku dużych zbiorów danych

Filip Wójcik
Senior Data Scientist
http://maddatascientist.eu
filip.wojcik@outlook.com
filip.wojcik@ue.wroc.pl

Uniwersytet Ekonomiczny
we Wrocławiu

Agenda

1. Problemy ze zrównoleglaniem algorytmów uczenia maszynowego
2. Typy zrównolegleń algorytmów
   - Zrównoleglanie danych
   - Zrównoleglanie modeli
3. Analiza algorytmów
   - Drzewa decyzyjne
   - Random Forest / lasy losowe
   - XGBoost – eXtreme Gradient Boosting
   - Sieci neuronowe

Problemy ze zrównoleglaniem uczenia maszynowego

Dlaczego tak trudno zrównoleglić uczenie statystyczne

Problemy ze zrównoleglaniem uczenia maszynowego

1. Większość algorytmów wymaga pełnego zbioru danych do policzenia określonych własności (entropia, korelacje)
2. Bardzo często kolejne kroki algorytmu są zależne od poprzednich – liniowa zależność operacji
3. Duża złożoność obliczeniowa – eksplozja kombinatoryczna
4. Zależność czasowa danych – niektóre obserwacje w zbiorze są zależne od innych i nie można ich zaburzać

Problemy ze zrównoleglaniem uczenia maszynowego

Brak „najlepszej metody” – ilość algorytmów jest tak duża, że nie ma jednego sposobu

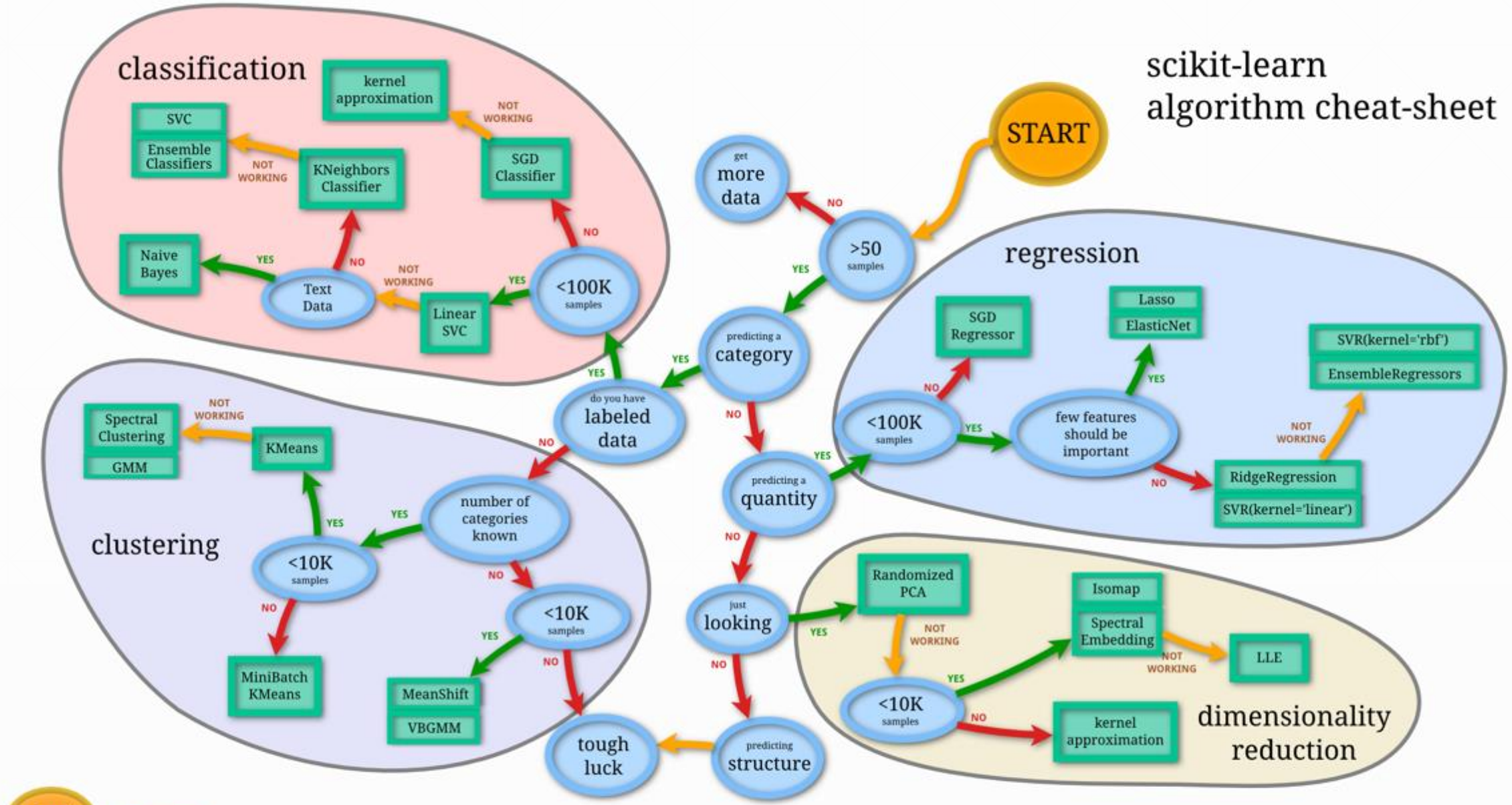
Zmienna \ Typ uczenia	Nienadzorowane	Nadzorowane
Ciągła	1. Klastrowanie • K-means • Hierarchiczne 2. Redukcja wymiarów: • SVD • PCA • ICA • NNMF	1. Regresja • Liniowa • Wielomianowa • Nieliniowa 2. Regresja drzewami • Drzewa CART • Random Forest 3. Sieci neuronowe
Dyskretna	1. Analiza asocjacyjna 2. Modele Markova	1. Klasyfikacja • KNN • Drzewa • Regresja logistyczna • Naiwny klasyfikator Bayesowski • SVM

Typy zrównoleglеń algorytmów

Czyli w jaki sposób można algorytmy wykonywać szybciej

Typy zrównoległeń algorytmów

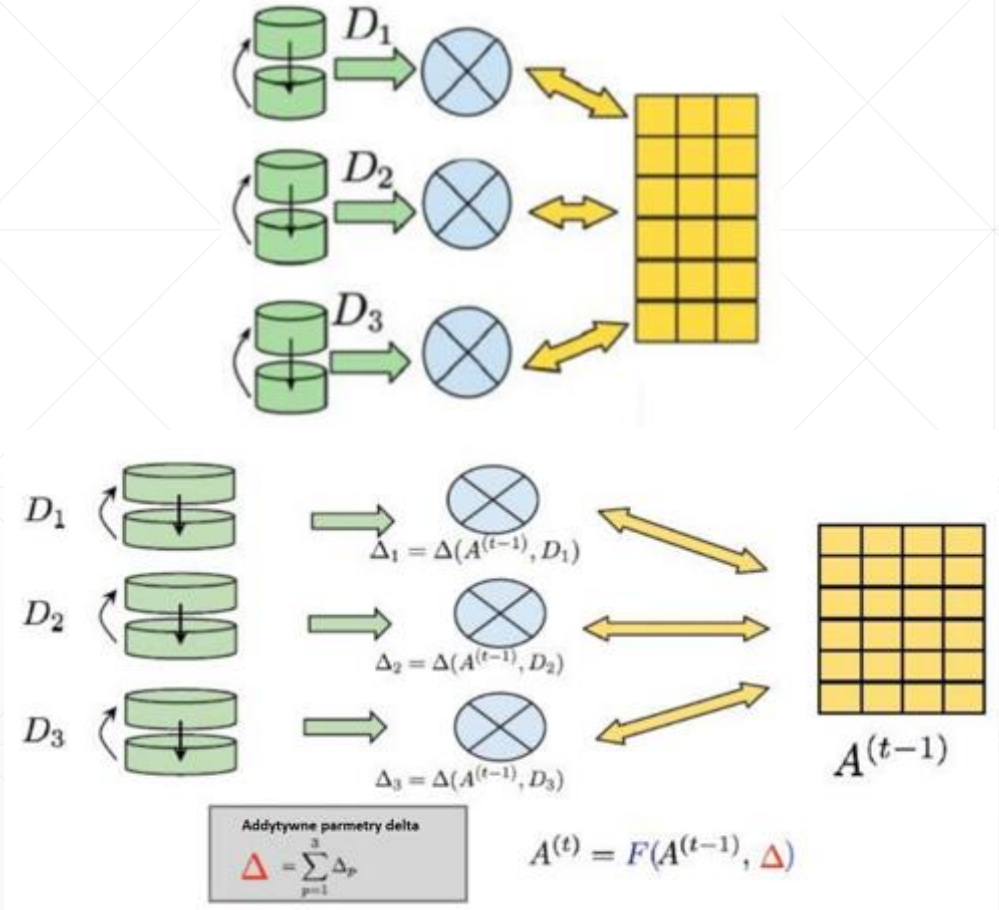
1. Istnieją dwie główne metody zrównoleglania algorytmów:
   - Zrównoleglanie danych/danymi (ang. *Data parallelization*)
   - Zrównoleglanie modeli (ang. *Model parallelization*)
2. Nie każdą z metod można zastosować w każdej sytuacji, ale nierzadko można znaleźć kompromisowe wyjście
3. Najlepiej do zrównoleglania nadają się modele złożone (ang. *ensemble model*), gdzie każda instancja jest niezależna od pozostałych, a wybór obserwacji (przykładów) jest mocno zrandomizowany
4. Najgorzej zrównolegla się algorytmy gradientowe, gdzie liczone są pochodne błędu poprzednich instancji (np. boosting) albo stan w czasie *t+1* zależny jest od stanu w czasie *t*
   - Np. sieci neuronowe
   - Wszystkie algorytmy typu *Gradient Boosting*


scikit-learn
algorithm cheat-sheet
START
>50 samples
NO
get more data
YES
predicting a category
YES
do you have labeled data
NO
predicting a quantity
YES
<100K samples
NO
SGD Regressor
YES
few features should be important
NO
Lasso
ElasticNet
YES
NOT WORKING
RidgeRegression
SVR(kernel='linear')
NOT WORKING
SVR(kernel='rbf')
EnsembleRegressors
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
NO
number of categories known
YES
<10K samples
YES
KMeans
NOT WORKING
Spectral Clustering
GMM
NO
MiniBatch KMeans
YES
<10K samples
NO
MeanShift
VBGMM
clustering
<100K samples
NO
SGD Classifier
NOT WORKING
kernel approximation
NOT WORKING
KNeighbors Classifier
NOT WORKING
SVC
Ensemble Classifiers
YES
<100K samples
NO
Linear SVC
NOT WORKING
Text Data
NO
Naive Bayes
YES
classification
Back
scikit learn

Typy zrównoleglеń algorytmów

Zrównoleglenie danych

1. Dane dzielone są na rozłączne części
2. Fragmenty modelu (węzły drzewa/pojedyncze parametry?) szkolone są na tych fragmentach
3. Potem następuje integracja parametrów/stanów modelu
4. Takie szkolenie jest możliwe tylko dla niektórych algorytmów – takich, gdzie można kombinować stany

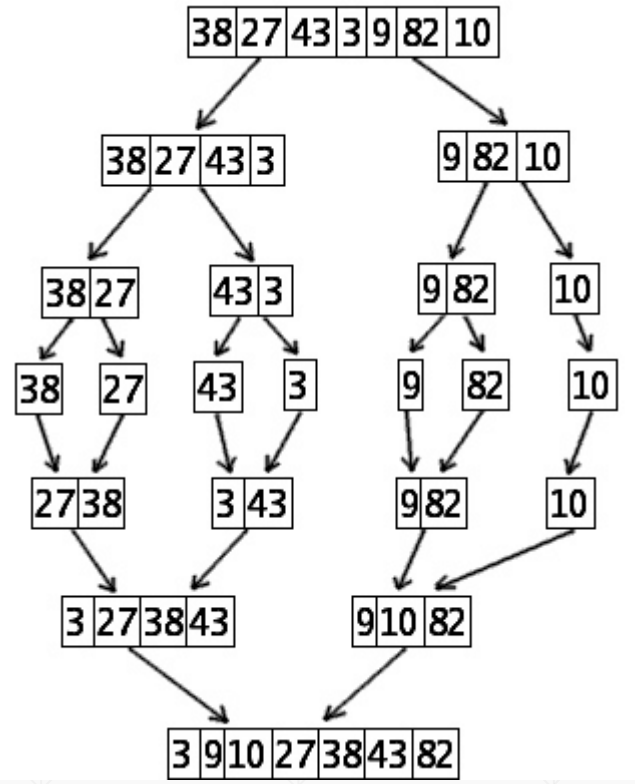


Typy zrównoleglеń algorytmów

Zrównoleglenie danych

Przykład sortowania – Merge sort

1. Zbiór danych dzielony jest rekurencyjnie na partycje
2. Każda para jest następnie sortowana lokalnie
3. Wyniki są scalane i łączone ze sobą – również rekurencyjnie, aż do rekonstrukcji całości

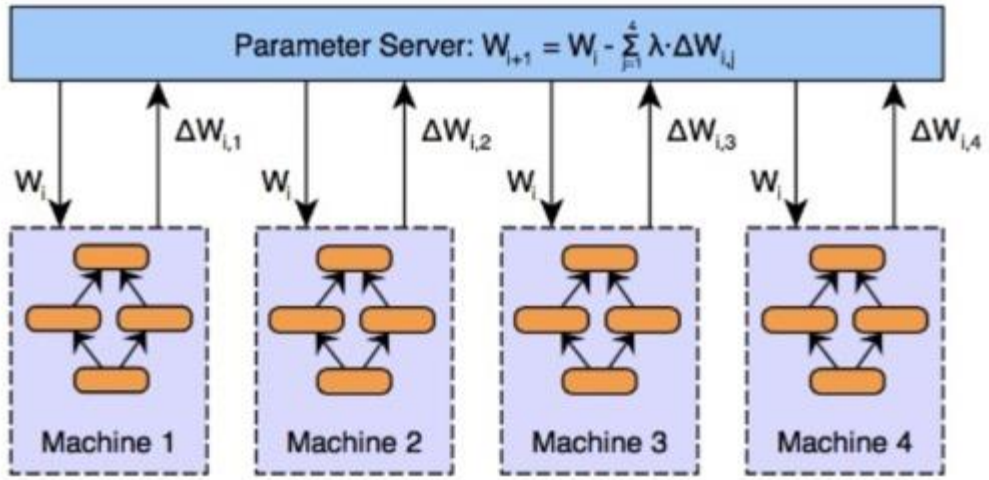


Typy zrównoleglеń algorytmów

Zrównoleglenie danych

Przykład Sparka

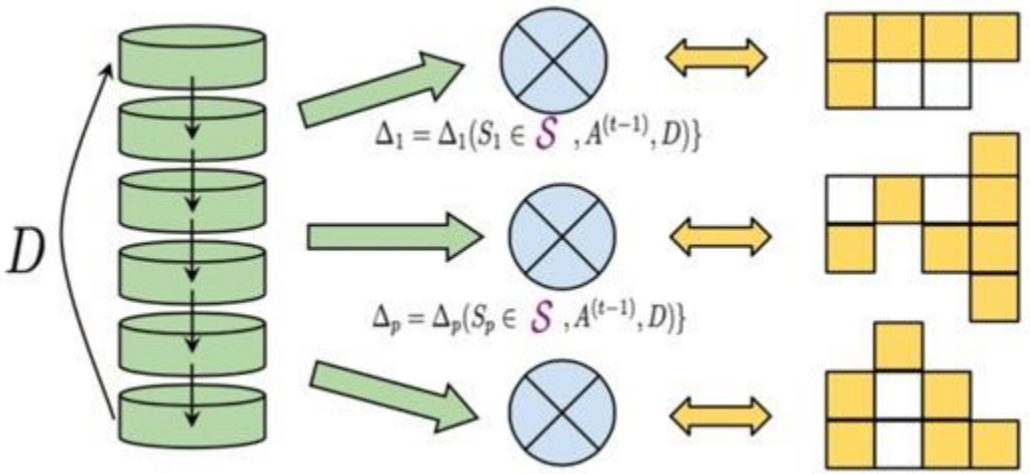
1. Regresja liniowa wykonywana na kilku maszynach na fragmentach danych
2. Parametry regresji liniowej są następnie centralnie zbierane
3. Następuje ich uśrednienie – w ten sposób jest budowany model złożony z kilku modeli



Typy zrównoleglеń algorytmów

Zrównoleglenie modeli

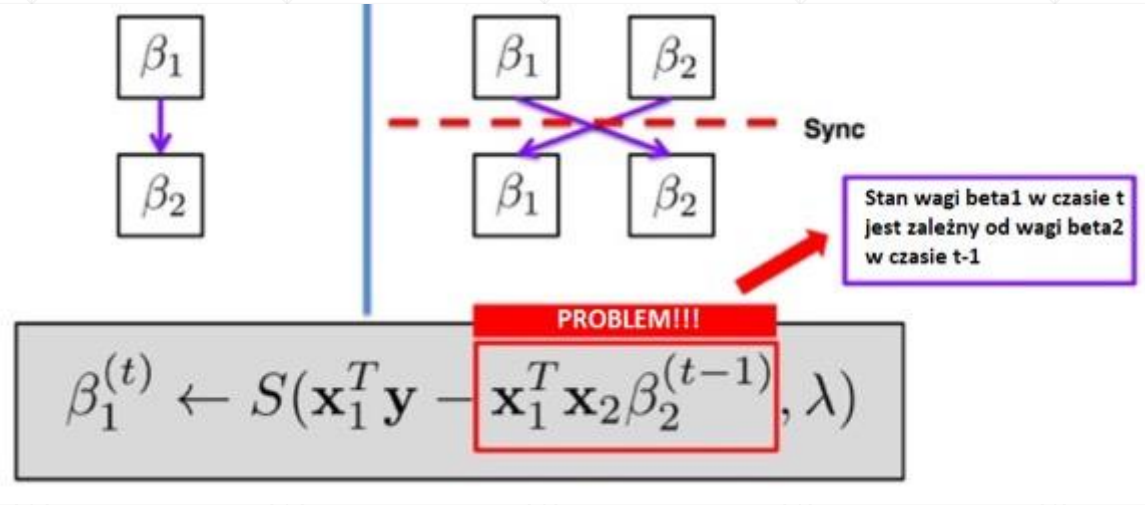
1. Wszystkie dane / ich podzbiór przekazywane są niezależnym instancjom algorytmu
2. De facto szkolonych jest wiele instancji tego samego algorytmu
3. Ich predykcje / wyniki działania są następnie kombinowane ze sobą:
   - Kombinowanie przez głosowanie
   - Kombinowanie za pomocą funkcji/analitycznie
4. Ten sposób jest dużo łatwiejszy i praktycznie uniwersalny – można go stosować z (niemal) każdym algorytmem



Typy zrównoleglеń algorytmów

Zrównoleglenie modeli

1. To podejście nie jest jednak wolne od problemów i wad
2. Może się zdarzyć, że modele są od siebie zależne w czasie – nie można utworzyć modelu w czasie *t+1* nie znając stanu modelu w czasie *t*
3. Jest to bardzo częste w przypadku tzw. *boostingu* oraz *sieci neuronowych*

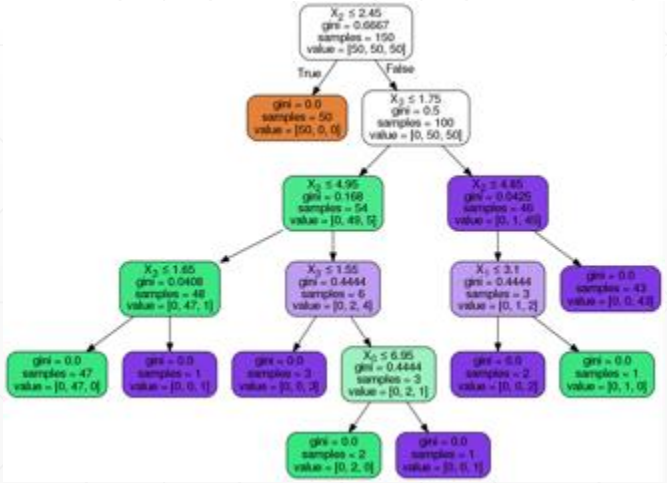


Analiza algorytmów

Jak można zrównoleglać poszczególne rodzaje algorytmów

Analiza algorytmów – drzewa decyzyjne

1. Pojedyncze drzewa poddają się **bardzo dobremu zrównoleglaniu**
2. Oparte są na zasadzie dziel-i-rządź czyli rekurencyjnemu partycjonowaniu danych
3. Każda partycja jest niezależna od pozostałych dzięki czemu możliwe jest zrównoleglenie
4. Zrównoleglenie „hybrydowe” – kolejne partycje zrównoleglane jako dane
   jednocześnie – są fragmentami modelu



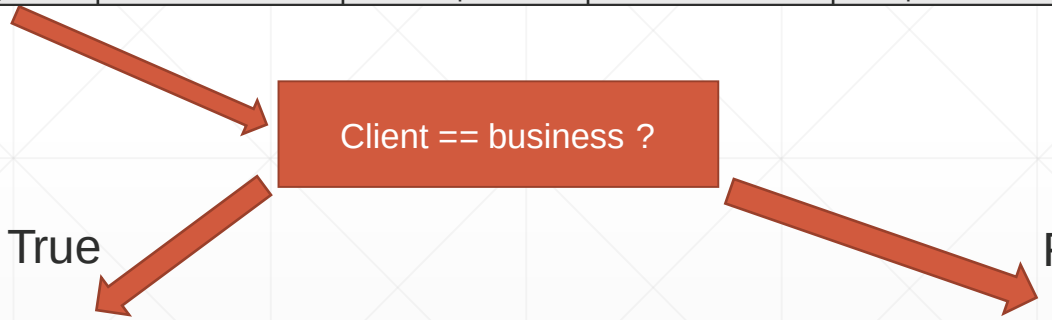
Analiza algorytmów – drzewa decyzyjne

client	hotel	addons	money_spent	offer
business	Hilton	trip	40000	deluxe
business	Hilton	full board	38000	deluxe
business	Hilton	trip	40000	deluxe
middle class	Meta	none	800	basic
middle class	Meta	meal	900	basic
manager	Meta	spa	1500	premium

Value	Count	%
Deluxe	3	0.5
Basic	2	0.333
Premium	1	0.16666

Analiza algorytmów – drzewa decyzyjne

client	hotel	addons	money_spent	offer
business	Hilton	trip	40000	deluxe
business	Hilton	full board	38000	deluxe
business	Hilton	trip	40000	deluxe
middle class	Meta	none	800	basic
middle class	Meta	meal	900	basic
manager	Meta	spa	1500	premium



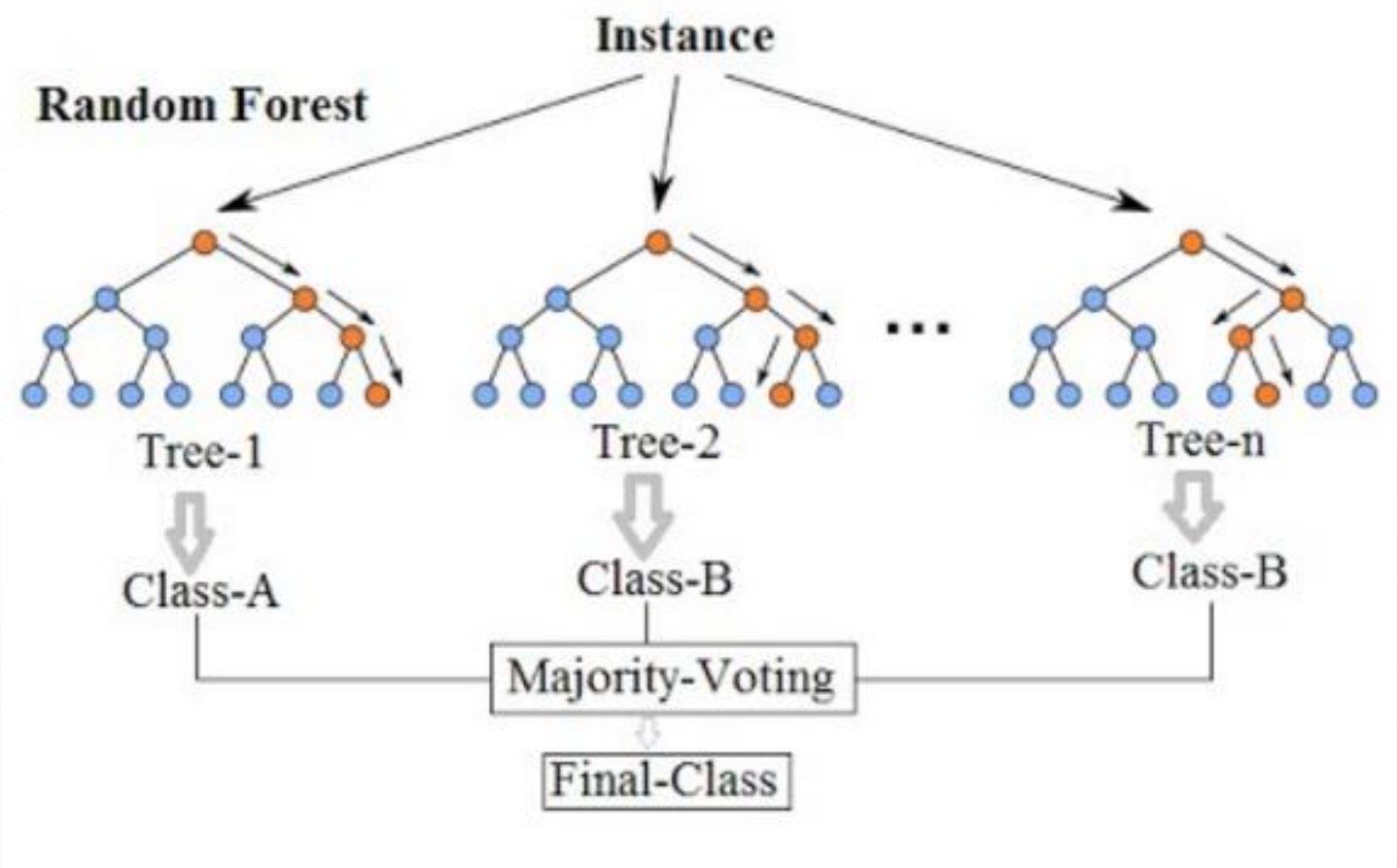
hotel	addons	money_spent	offer
Hilton	trip	40000	deluxe
Hilton	full board	38000	deluxe
Hilton	trip	40000	deluxe

hotel	addons	money_spent	offer
Meta	none	800	basic
Meta	meal	900	basic
Meta	spa	1500	premium

Analiza algorytmów – random forest

1. Random Forest („lasy losowe”) jeden z najpopularniejszych algorytmów klasyfikacji, głównie ze względu na wysoce zrównoleglony charakter
2. Należy do kategorii tzw. algorytmów *oczywiście równoległych* (ang. *embarrassingly paralel*) – ich zrównoleglenie narzuca się samo ☺
3. Seria drzew decyzyjnych
   - Każde niezależne od pozstałych – nic praktycznie ich nie łączy
   - Każde drzewo dostaje losową próbkę danych
   - Ostateczna predykcja jest wynikiem głosowania (ważonego lub nie) wszystkich drzew
4. Możliwość zrównoleglenia:
   - **Modelu –** każde drzewo szkolone osobno
   - **Danych** – każde drzewo dostaje podzbiór danych

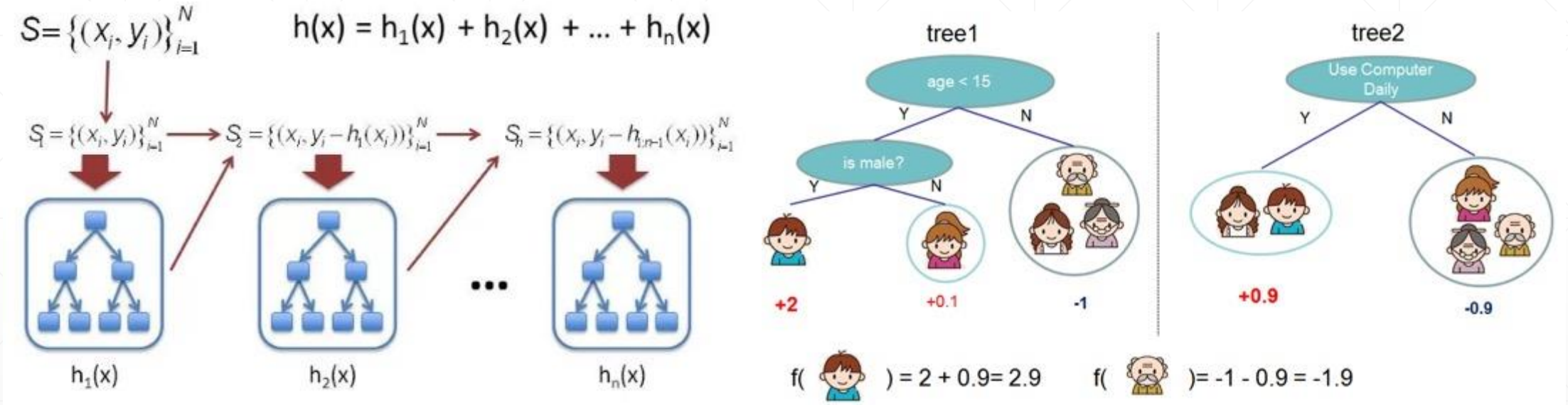
Analiza algorytmów – random forest



Analiza algorytmów – Gradient Tree Boosting

1. XGBoost – jeden z najbardziej trafnych algorytmów regresji/klasyfikacji, wygrywający większość konkursów na platformie Kaggle.com
2. Oparty na podwójnej zasadzie:
   - Randomizacja wielu drzew, jak w algorytmie Random Forest
   - *Gradient Boosting* czyli uczenie się na błędach poprzednich instancji, w oparciu o gradient (pochodną) funkcji błędu
3. Algorytm sekwencyjny – kolejne drzewa zależą od poprzednich
4. Nadaje się do **zrównoleglania tylko w ramach pojedynczego drzewa**
5. Nowe implementacje wprowadzają **zrównoleglone licznie gradientów, następnie scalanych pomiędzy maszynami**

Analiza algorytmów – Gradient Tree Boosting



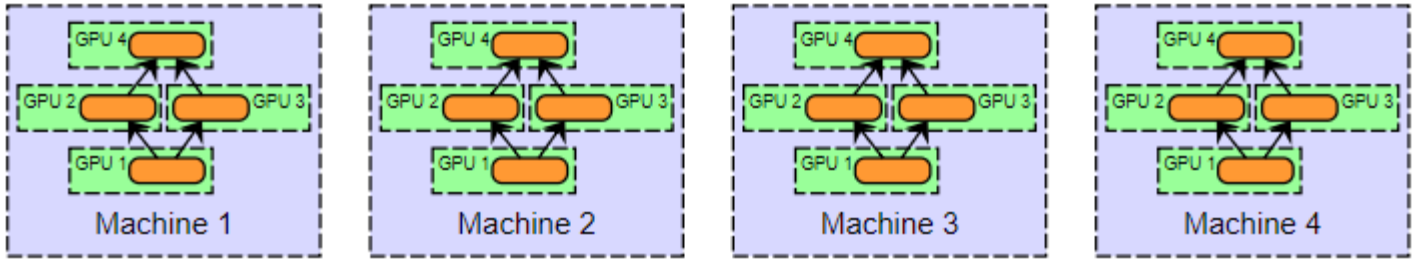
Źródło: official Xgboost documentation. http://xgboost.readthedocs.io/en/latest/model.html

Analiza algorytmów – Sieci neuronowe

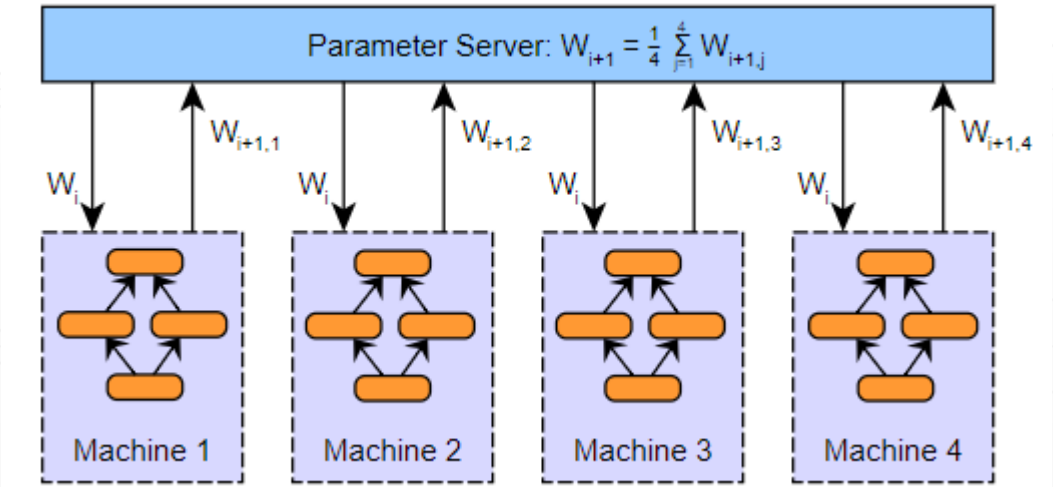
1. Sieci neuronowe są bardzo zróżnicowanym zestawem algorytmów
    1. MLP – wielowarstwowe (zazwyczaj dwu) perceptrony
    2. Sieci głębokie
    3. Sieci splątane i rekurencyjne
2. Wszystkie są jednak szkolone metodami gradientowymi – wymagają więc pełnego przebiegu i liczenie pochodnych funkcji błędu z poprzednich iteracji
3. Dodatkowo – sieci neuronowe potrzebują bardzo dużych zbiorów danych uczących
4. Stosowane są różne techniki przyspieszania sieci:
    1. Szkolenie na podzbiorach danych (zrównoleglenie danych)
    2. Integracja gradientów z poszczególnych (zrównoleglenie modeli)

Analiza algorytmów – Sieci neuronowe

Zrównoleglenie modelu



Integracja parametrów



Źródło:https://blog.skymind.ai/distributed-deep-learning-part-1-an-introduction-to-distributed-training-of-neural-networks/

Analiza algorytmów – Sieci neuronowe

Obliczenia w ramach sieci neuronowych nadzwyczaj dobrze wpisują się w architekturę **procesorów graficznych (GPU)**. Obecnie produkowane są dedykowane modele, mające za zadanie wspierać tworzenie sieci neuronowych.

Do najpopularniejszych bibliotek, wspomagających ten proces należą:

theano

PYTORCH

TensorFlow

Keras