

Wprowadzenie do uczenia maszynowego

Joanna Jaworek-Korjakowska

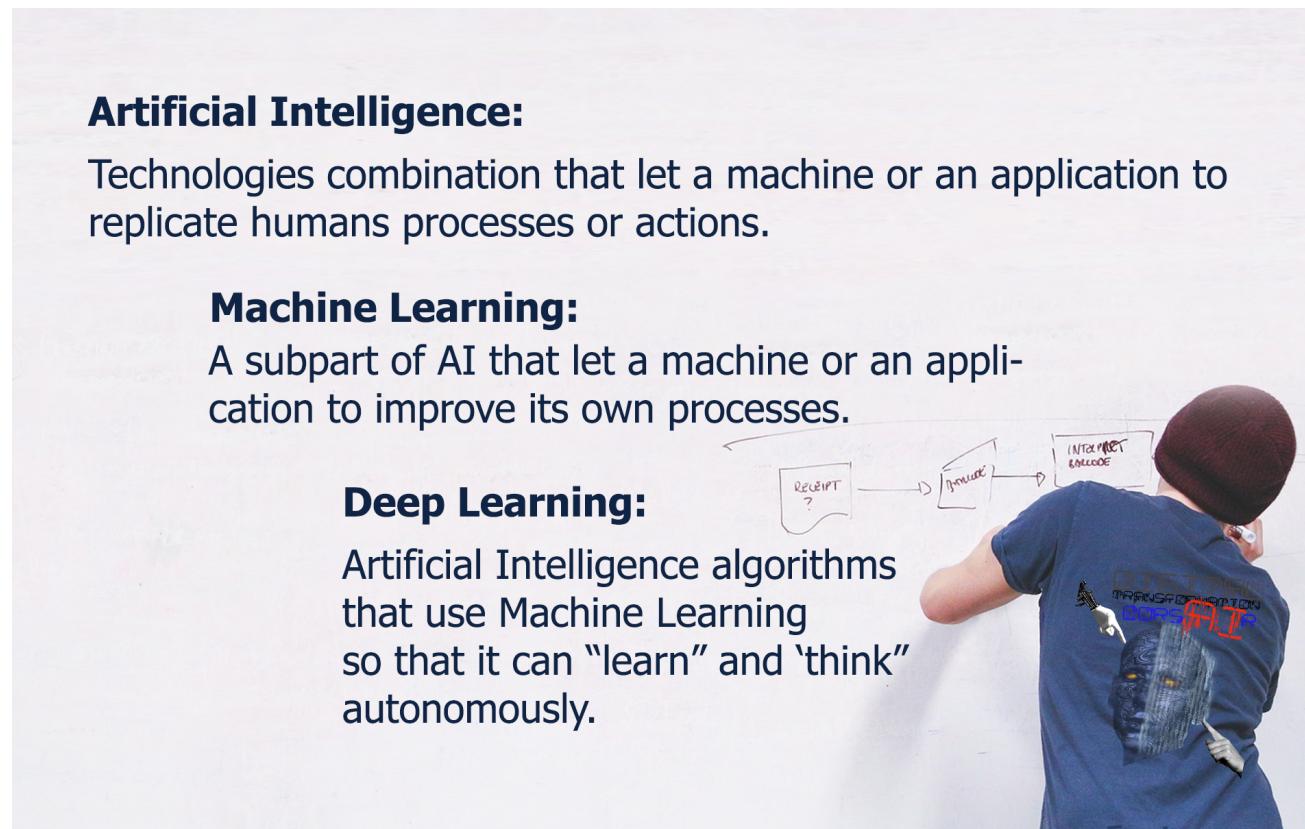
WEAIIB, Katedra Automatyki i Robotyki, ISS

2019

Definicja procesu uczenia maszynowego

Maszynowe uczenie się (ang. **Machine Learning**) jest analizą procesów uczenia się oraz tworzeniem systemów, które doskonala swoje działanie na podstawie doświadczeń z przeszłości.

Maszynowe uczenie się – część sztucznej inteligencji (AI) lub inteligencji obliczeniowej (Computational Intelligence - CI).

A photograph of a person from behind, wearing a blue t-shirt with text on it, writing on a whiteboard. The whiteboard contains a hand-drawn diagram illustrating a machine learning process. The diagram shows three main components: a box labeled "RECEIPT?", an arrow pointing to a box labeled "Analyze", and another arrow pointing from "Analyze" to a box labeled "INTERPRET & ACT".

Artificial Intelligence:
Technologies combination that let a machine or an application to replicate humans processes or actions.

Machine Learning:
A subpart of AI that let a machine or an application to improve its own processes.

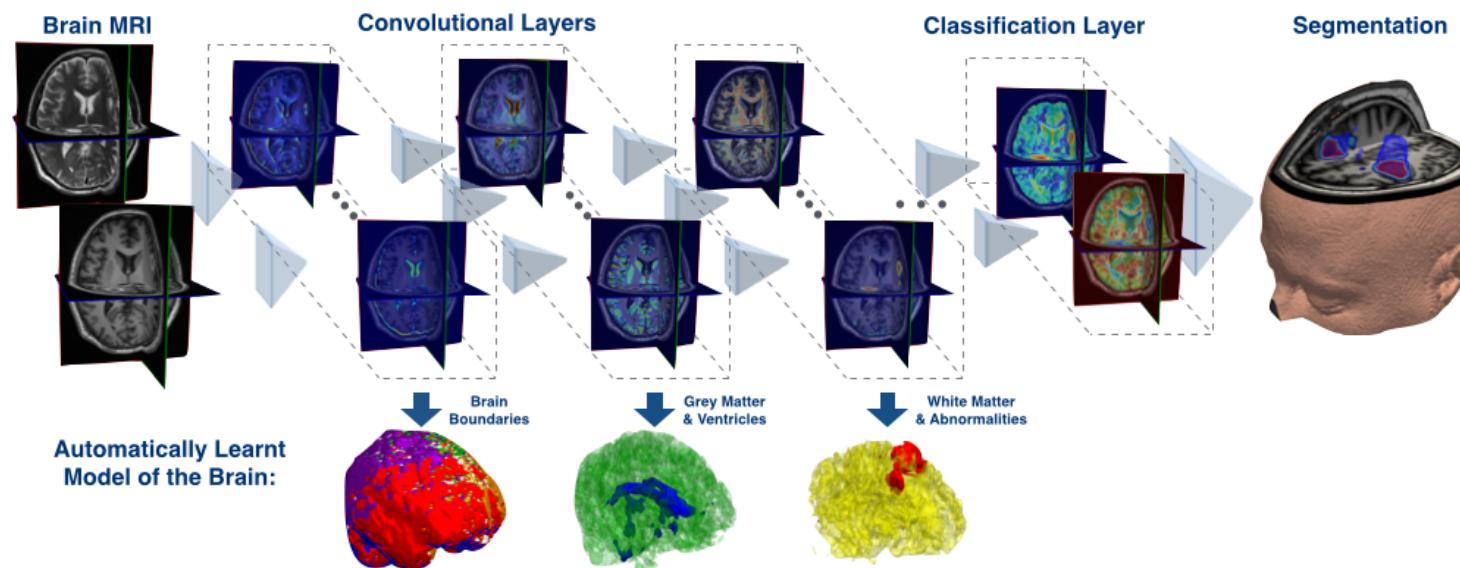
Deep Learning:
Artificial Intelligence algorithms that use Machine Learning so that it can “learn” and ‘think’ autonomously.

Definicja procesu uczenia maszynowego

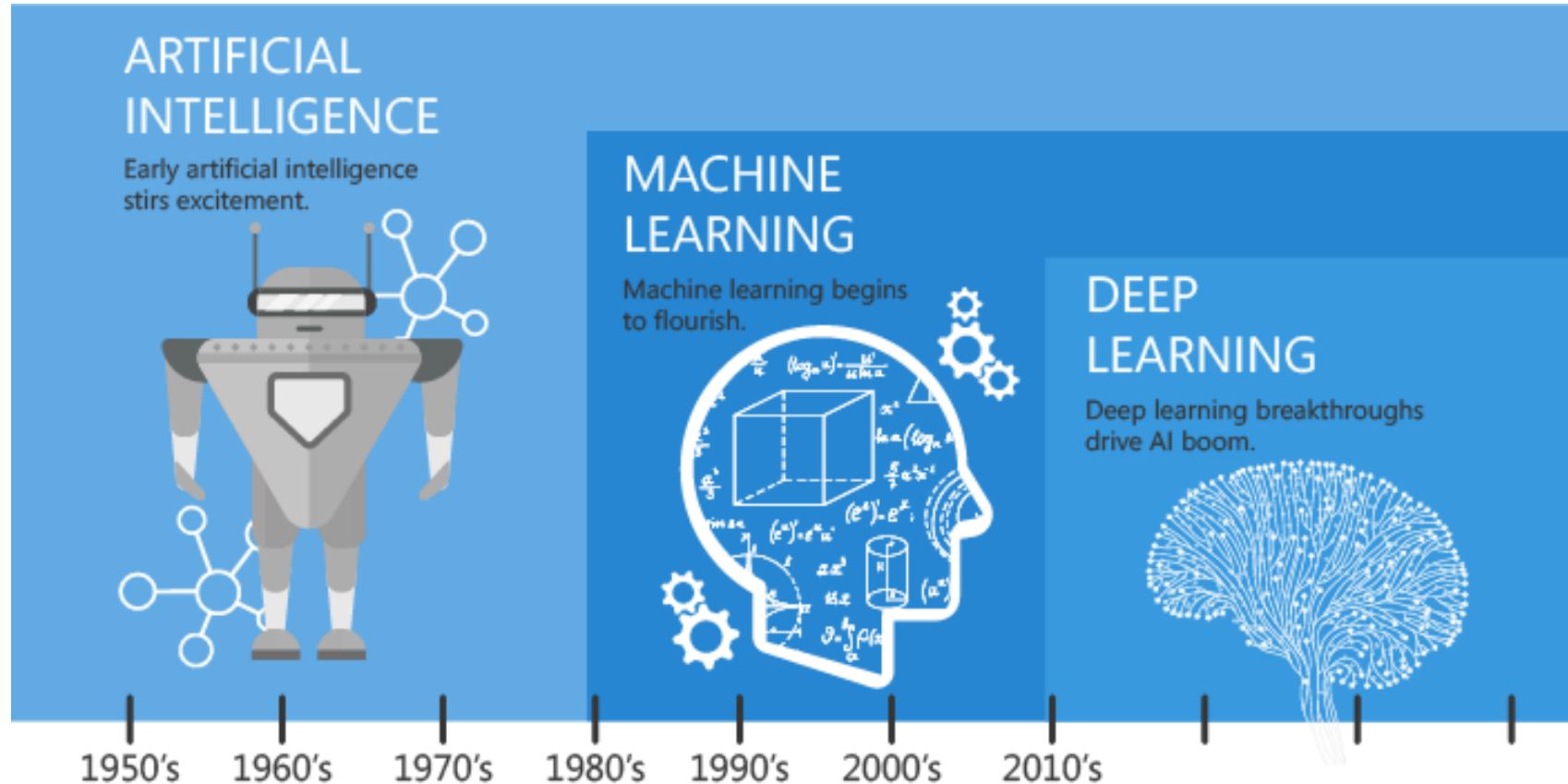
Systemy uczące się – systemy adaptujące się, zmieniające swoje wewnętrzne parametry tak, aby rozpoznać charakter danych.

ML umożliwia pozyskiwanie wiedzy na podstawie analizy zachowań ekspertów lub danych doświadczalnych, tj. przykładów uczących.

Wiedza otrzymana metodami ML może być lepsza niż wiedza bezpośrednio wydedukowana przez ludzi.



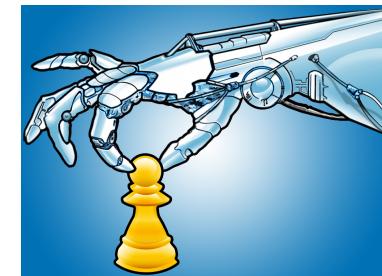
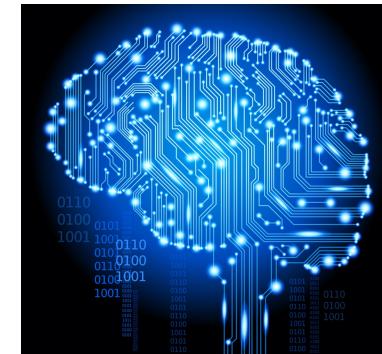
Definicja procesu uczenia maszynowego



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

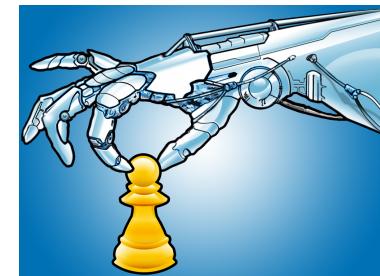
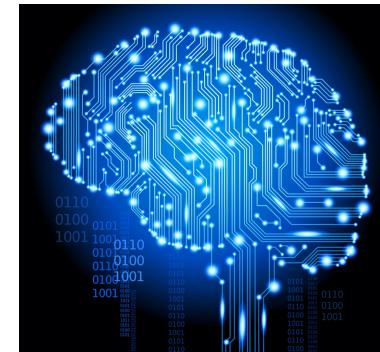
Pojęcie wiedzy w sztucznej inteligencji

- Pojęcie **wiedzy w sztucznej inteligencji** odnosi się do struktur modeli reprezentujących pewne **procesy podejmowania decyzji**.
- W zależności od **procesu podejmowania decyzji** wiedza może być **reprezentowana** w postaci rozmaitych struktur, takich jak **funkcje, drzewa, grafy, reguły, bądź zbiory**.
- Wiedza może mieć charakter:
 - **zrozumiały (interpretowalny)**;
 - **niejawny (nieinterpretowalny)**.
- Źródła wiedzy:
 - wiedza eksperta;
 - wiedza pozyskana z danych.



Pojęcie wiedzy w sztucznej inteligencji

- Pojęcie **wiedzy w sztucznej inteligencji** odnosi się do struktur modeli reprezentujących pewne **procesy podejmowania decyzji**.
- W zależności od **procesu podejmowania decyzji** wiedza może być **reprezentowana** w postaci rozmaitych struktur, takich jak **funkcje, drzewa, grafy, reguły, bądź zbiory**.
- Wiedza może mieć charakter:
 - **zrozumiały (interpretowalny)**;
 - **niejawny (nieinterpretowalny)**.
- Źródła wiedzy:
 - wiedza eksperta;
 - **wiedza pozyskana z danych**.



Uczenie maszynowe i eksploracja danych (1)

- **Uczenie maszynowe** (*ang. machine learning*) to proces pozyskiwania wiedzy do rozwiązania pewnego **zadania** w oparciu o **doświadczenie** i z wykorzystaniem pewnej **miary jakości**.
- Wraz ze **wzrostem doświadczenia**, następuje **przyrost wiedzy** potrzebnej do realizacji **zadania** mierzony z wykorzystaniem **miary jakości**.
- **Eksploracja (ekstrakcja) danych** (*ang. data mining*) to proces **pozyskiwania wiedzy z danych** reprezentowanej przez pewne **wzorce**.

ZADANIE
Jaka to litera ?

4

DOŚWIADCZENIE

0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4
5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5
6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6
7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7
8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8
9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9

MIARA JAKOŚCI

4 → OK → 4

Definicja procesu uczenia maszynowego

Mówimy, że maszyna uczy się zadania T w oparciu o doświadczenie E i miarę jakości P, jeśli wraz z przyrostem doświadczenia E poprawia się jakość wykonywanego zadania T mierzona przez miarę P. (Tom Mitchell, *Machine Learning*, 1997)

- T: Gra w szachy
- P : Liczba wygranych partii w turnieju z człowiekiem
- E: Rozgrywanie partii przeciw sobie

Uczenie się oznacza zmiany w systemie, które mają charakter adaptacyjny w tym sensie, że pozwalają systemowi wykonać za następnym razem takie same zadanie lub zadania podobne bardziej efektywnie.

Wspólne pojęcia w definicjach:

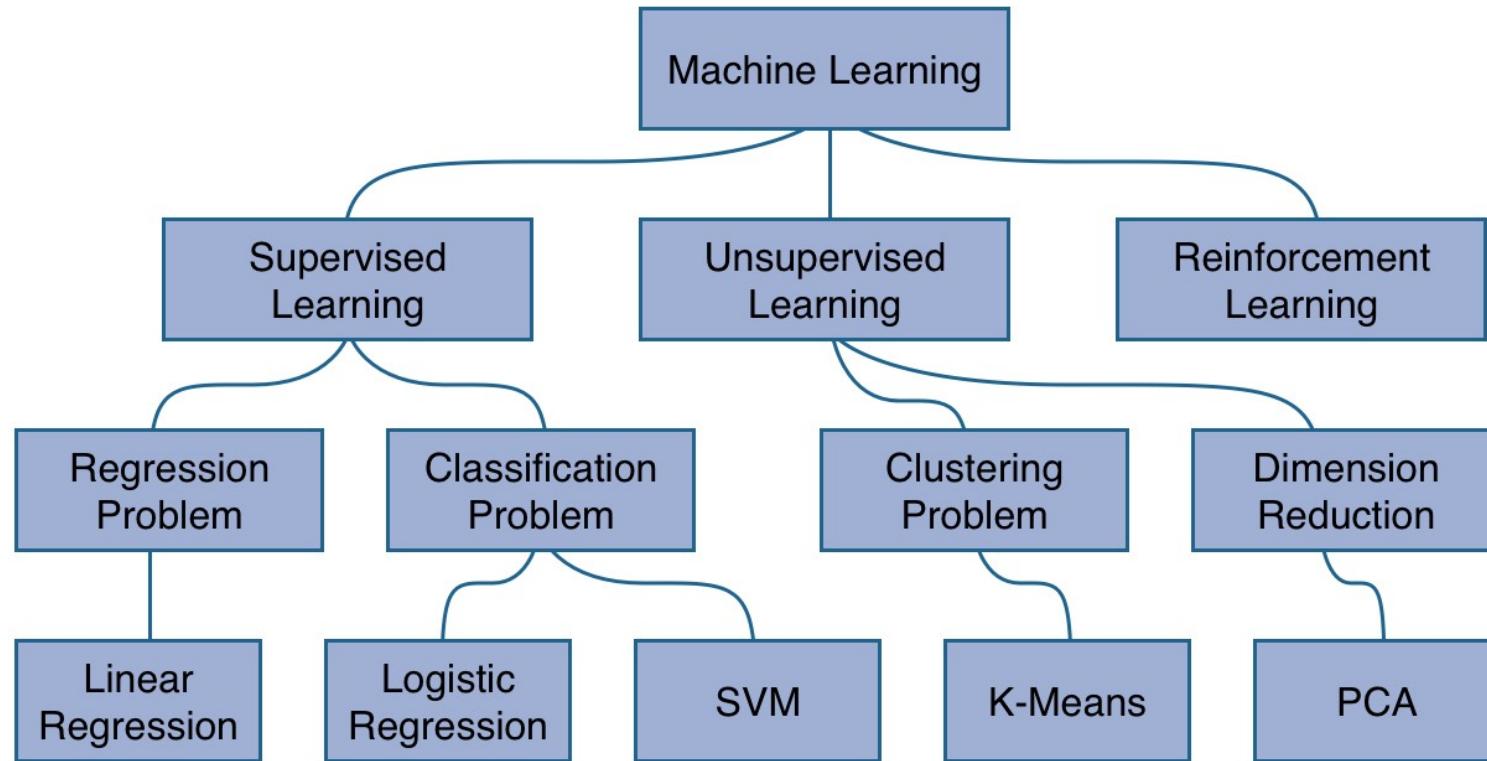
- Wejście: dane empiryczne
- Zmiany/poprawa działania (miara oceny)
- Postulat zdobywania wiedzy, reprezentowania jej wewnątrz systemu i stosowania do wykonywania zadania

Przykładowe zadania ML

- **grupowanie** obiektów podobnych
- **rozpoznawanie** istotnych wzorców w danych
- **klasyfikacja** obserwowanych przypadków
- **przewidywanie** przyszłości na podstawie obserwacji
- **wykrywanie** trendów w danych

W uczeniu maszynowym powyższe cele realizowane są automatycznie lub przy niewielkim wsparciu człowieka

Metody uczenia maszynowego



Uczenie maszynowe

Etapy uczenia maszynowego:

- zbieranie danych
- przetwarzanie i czyszczenie danych
- podział na zbiór treningowy i uczący (tylko w uczeniu z nadzorem)
- uczenie systemu na podstawie danych treningowych
- ewaluacja (iteracja)
- wykorzystanie systemu do zadań

Notation and Simple Matrix Algebra

Dane w uczeniu maszynowym

- Jeżeli rozważamy problem **uczenia nadzorowanego (predykcji)**, to interesuje nas znalezienie **mapowania** wartości wejściowych x na wartości wyjściowe y .
- Mapowanie to odbywa się na podstawie tzn. **zbioru uczącego (treningowego)**, który zawiera pary wejście-wyjście nazywane **przykładami**:

$$\mathcal{D} = \{(\mathbf{x}_n, y_n)\}_{n=1}^N.$$

- Każdy element wejściowy \mathbf{x}_i zawiera zestaw wartości **nominalnych i liczbowych**, które nazywane są **cechami**, bądź **atrybutami**.
- Każdy element wyjściowy y_i reprezentowany jest przez wartość liczbową (**regresja**), bądź też nominalną (**klasyfikacja**).
- Jeżeli rozważamy problem **uczenia nienadzorowanego (deskrypcji)** to interesuje nas znalezienie "**ciekawych wzorców**" w danych:

$$\mathcal{D} = \{\mathbf{x}_n\}_{n=1}^N.$$

Notation

Notation for this course:

- n represents the **number** of distinct data points, or observations, in our sample
- p denotes the **number of variables** that are available for use in making predictions.

For example, the Skin-cancer data set consists of 12 variables for 3,000 people, so we have $n = 3,000$ observations and $p = 12$ variables (such as skin type, age, wage, and more)

Notation

In general, we will let x_{ij} represent the value of the j th variable for the i th observation, where $i = 1, 2, \dots, n$ and $j = 1, 2, \dots, p$.

i will be used to index the samples or observations (from 1 to n)

j will be used to index the variables (from 1 to p). We let \mathbf{X} denote a $n \times p$ matrix whose (i,j) th element is x_{ij} . That is,

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}$$

Notation

At times we will be interested in the rows of \mathbf{X} , which we write as $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$.

Here \mathbf{x}_i is a vector of length p , containing the p measurements for the i th observation. That is,

$$\mathbf{x}_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{pmatrix}$$

For example, for the Skin-cancer dataset, \mathbf{x}_1 contains the $n = 3,000$ values for Skin Type.

Notation

Using this notation, the matrix \mathbf{X} can be written as :

$$\mathbf{x}_i = (\mathbf{x}_{i1} \ \mathbf{x}_{i2} \ \dots \ \mathbf{x}_{ip})$$

or

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix}$$

Notation

Let y_i represent the response variable for the i th observation. We write the set of all n observations in vector form:

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

Then our observed data consists of $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, where each x_i is a vector of length p . (If $p = 1$, then x_i is simply a scalar.)

Źródła danych

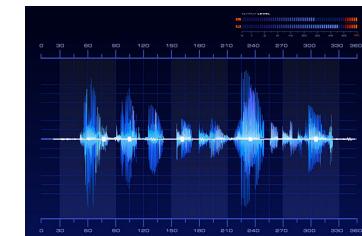
DANE BANKOWE



DANE MEDYCZNE



DANE DŹWIĘKOWE



OBRAZY



DANE MAILOWE



PORTALE SPOŁECZNOŚCIOWE



DANE O Klientach



DANE Z CZUJNIKÓW



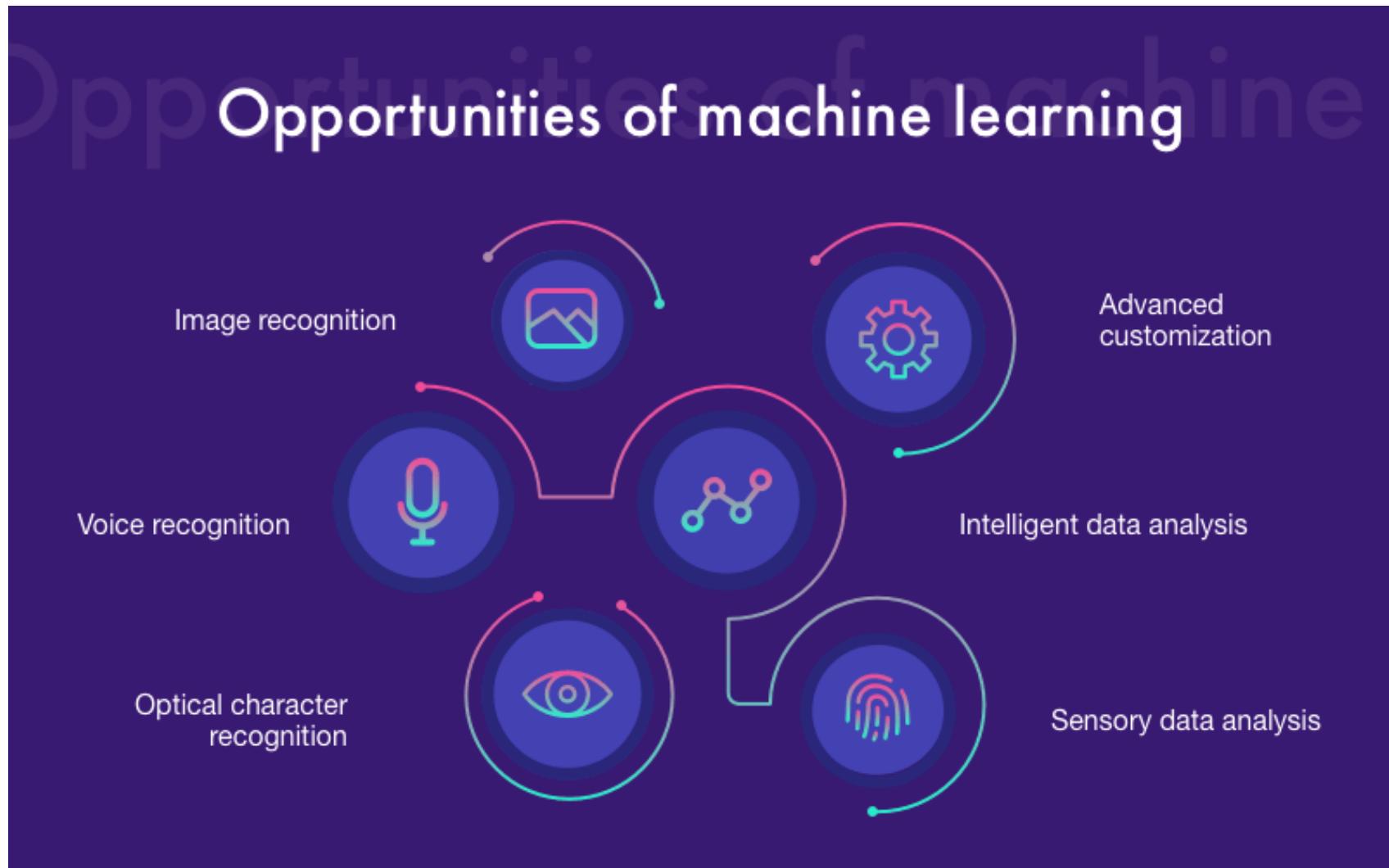
DANE GIEŁDOWE



Problemy uczenia maszynowego

- Uczenie z nadzorem (*ang. supervised learning*):
 - klasyfikacja (*ang. classification*);
 - regresja (*ang. regression*);
- Uczenie bez nadzoru (*ang. unsupervised learning*):
 - grupowanie (klasteryzacja, analiza skupień) (*ang. clustering*);
 - redukcja wymiarów (*ang. dimensionality reduction*);
 - uzupełnianie wartości (*ang. matrix completion*).
- Uczenie ze wzmacnieniem (*ang. reinforcement learning*).

Wykorzystanie ML



Wykorzystanie ML

Top 10 Use Cases for Data Science & Machine Learning



HEALTHCARE:
Patient Diagnosis



FINANCE:
Fraud Detection



MANUFACTURING:
Anomaly Detection



RETAIL:
Inventory Optimization



GOVERNMENT:
Smarter Services



TRANSPORTATION:
Demand Forecasting



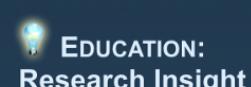
NETWORKS:
Intrusion Detection



E-COMMERCE:
Recommender Systems



MEDIA:
Interaction & Speed



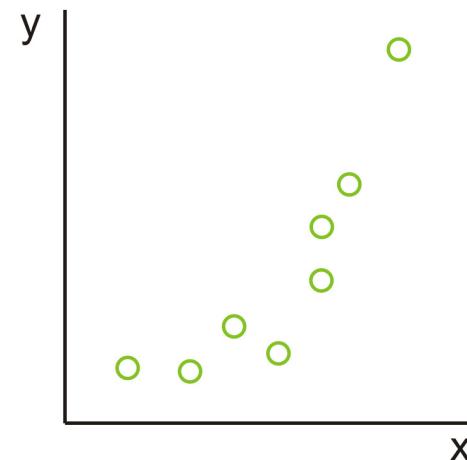
EDUCATION:
Research Insight

Podstawowe problemy

Uczenie z nadzorem: Regresja

- **Regresja** (ang. **Regression**):

- Dysponujemy obserwacjami z odpowiadającymi im **wartościami ciągłyimi**.
- Celem uczenia jest skonstruowanie **modelu regresji** na podstawie danych.
- Model konstruowany jest tak, aby możliwe było przewidywanie **nowych** obserwacji.

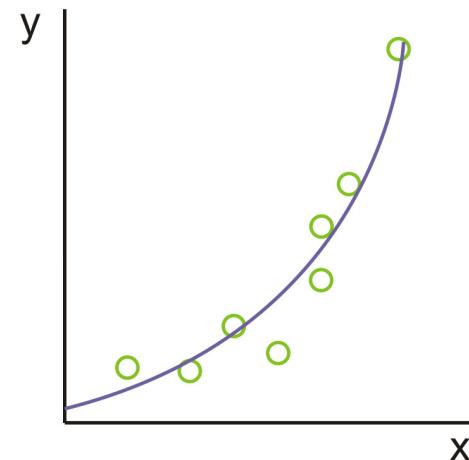


Podstawowe problemy

Uczenie z nadzorem: Regresja

- **Regresja** (ang. **Regression**):

- Dysponujemy obserwacjami z odpowiadającymi im **wartościami ciągłyimi**.
- Celem uczenia jest skonstruowanie **modelu regresji** na podstawie danych.
- Model konstruowany jest tak, aby możliwe było przewidywanie **nowych** obserwacji.

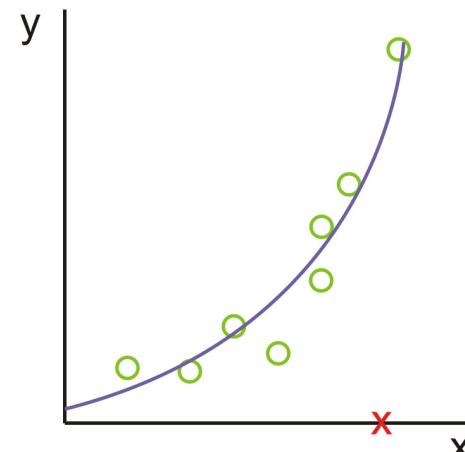


Podstawowe problemy

Uczenie z nadzorem: Regresja

- **Regresja** (ang. **Regression**):

- Dysponujemy obserwacjami z odpowiadającymi im **wartościami ciągłyimi**.
- Celem uczenia jest skonstruowanie **modelu regresji** na podstawie danych.
- Model konstruowany jest tak, aby możliwe było przewidywanie **nowych** obserwacji.

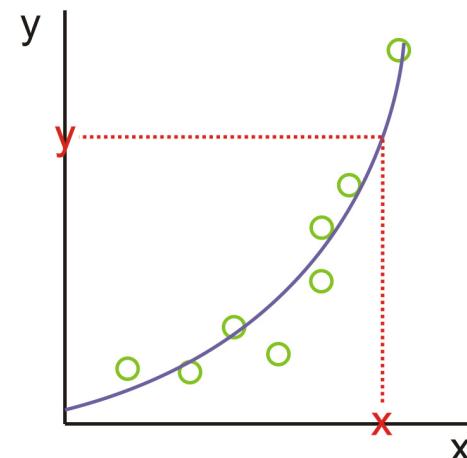


Podstawowe problemy

Uczenie z nadzorem: Regresja

- **Regresja** (ang. **Regression**):

- Dysponujemy obserwacjami z odpowiadającymi im **wartościami ciągłyimi**.
- Celem uczenia jest skonstruowanie **modelu regresji** na podstawie danych.
- Model konstruowany jest tak, aby możliwe było przewidywanie **nowych** obserwacji.



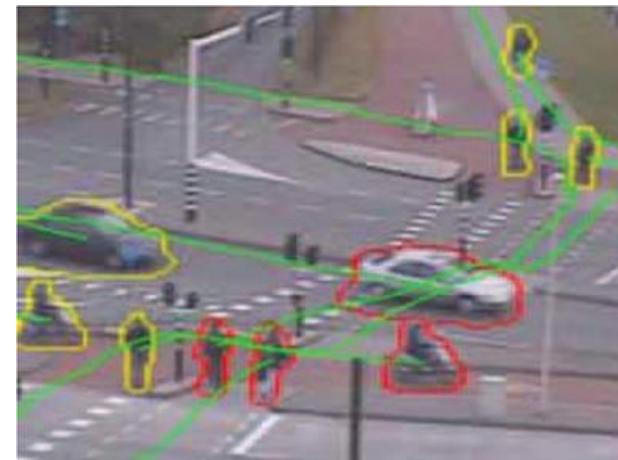
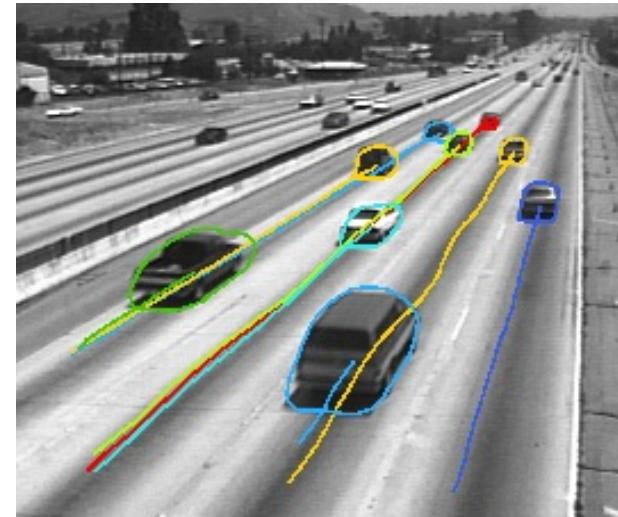
Podstawowe problemy

Regresja: Śledzenie ruchu

Cel: Wyznaczenie następnego położenia obiektu.

Dane: Sekwencja obrazów z poruszającymi się obiektami.

- Na podstawie dotychczas zarejestrowanej sekwencji obrazów wyznaczane jest położenie obiektu.



Podstawowe problemy

Regresja: Predykcja notowań giełdowych

Cel: Wycena akcji.

Dane: Notowania akcji z poprzednich okresów oraz inne czynniki wpływające na cenę akcji.

- Na podstawie notowań historycznych i innych czynników mających wpływ na cenę akcji budowany jest model predykcyjny.
- Model aktualizowany jest z wykorzystaniem bieżących notowań.



Podstawowe problemy

Regresja: Predykcja przeżywalności pooperacyjnej

Cel: Określenie jaki okres czasu pacjent przeżyje po operacji.

Dane: Wyniki badań pacjenta przeprowadzonych przed i po operacji, ogólna charakterystyka zdrowia pacjenta.

- Na podstawie danych o pacjencie należy określić jaki okres czasu przeżyje on po operacji.

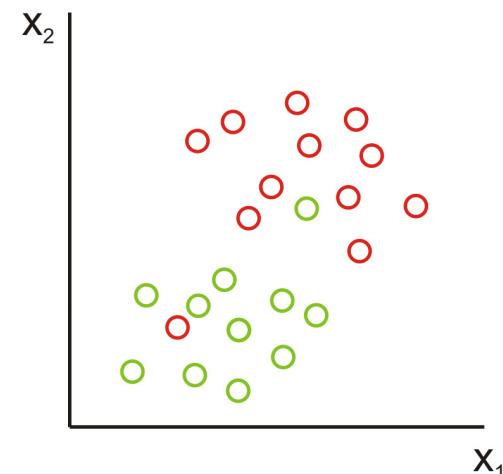


Podstawowe problemy

Uczenie z nadzorem: Klasyfikacja

- **Klasyfikacja** (ang. **Classification**):

- Dysponujemy obserwacjami z **etykietami** (klasami), które przyjmują wartości nominalne.
- Celem uczenia jest skonstruowanie **klasyfikatora** separującego obiekty należące do różnych klas.
- Klasyfikator konstruowany jest tak, aby możliwe było przewidywanie klas nowych, **niesklasyfikowanych obserwacji**.

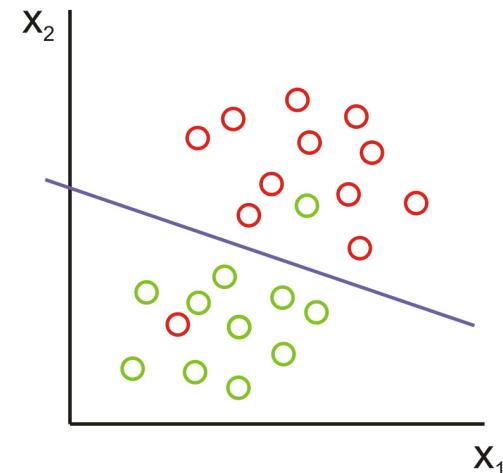


Podstawowe problemy

Uczenie z nadzorem: Klasyfikacja

- **Klasyfikacja** (ang. **Classification**):

- Dysponujemy obserwacjami z **etykietami** (klasami), które przyjmują wartości nominalne.
- Celem uczenia jest skonstruowanie **klasyfikatora** separującego obiekty należące do różnych klas.
- Klasyfikator konstruowany jest tak, aby możliwe było przewidywanie klas nowych, **niesklasyfikowanych obserwacji**.

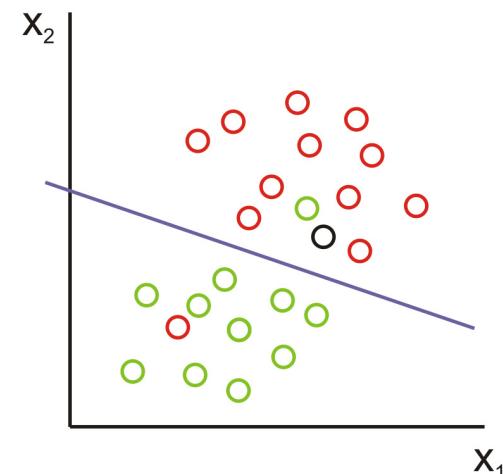


Podstawowe problemy

Uczenie z nadzorem: Klasyfikacja

- **Klasyfikacja** (ang. **Classification**):

- Dysponujemy obserwacjami z **etykietami** (klasami), które przyjmują wartości nominalne.
- Celem uczenia jest skonstruowanie **klasyfikatora** separującego obiekty należące do różnych klas.
- Klasyfikator konstruowany jest tak, aby możliwe było przewidywanie klas nowych, **niesklasyfikowanych obserwacji**.

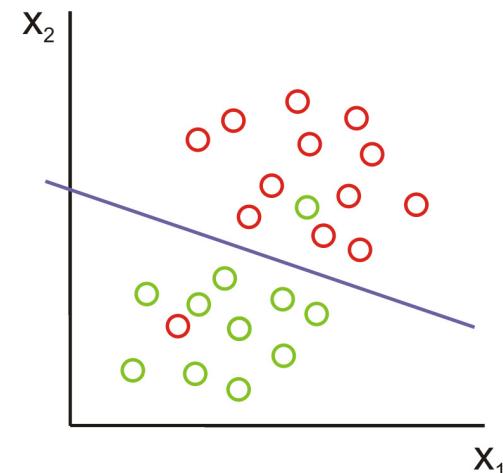


Podstawowe problemy

Uczenie z nadzorem: Klasyfikacja

- **Klasyfikacja** (ang. **Classification**):

- Dysponujemy obserwacjami z **etykietami** (klasami), które przyjmują wartości nominalne.
- Celem uczenia jest skonstruowanie **klasyfikatora** separującego obiekty należące do różnych klas.
- Klasyfikator konstruowany jest tak, aby możliwe było przewidywanie klas nowych, **niesklasyfikowanych obserwacji**.



Podstawowe problemy

Klasyfikacja: Rozpoznawanie znaków

Cel: Określenie, jaki znak (cyfra, litera) znajduje się na obrazku.

Dane: Zestaw obrazków treningowych reprezentujących różne znaki wraz z korespondującymi etykietami.

- Wydobywane są cechy obrazka różnicujące reprezentowane znaki.
- Na podstawie cech i wykorzystując dane treningowe wykonywana jest klasyfikacja obrazka do najbardziej prawdopodobnego znaku.



0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4
5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5
6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6
7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7
8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8
9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9



THE DATA SCIENCE / ANALYTICS LANDSCAPE



2,350,000

DSA job listings in 2015

By 2020, DSA job openings are projected to grow

15%

364,000

Additional job listings projected in 2020

Demand for both Data Scientists and Data Engineers is projected to grow

39%

DSA jobs remain open

5 days

longer than average

DSA jobs advertise average salaries of

\$80,265

With a premium over all BA+ jobs of

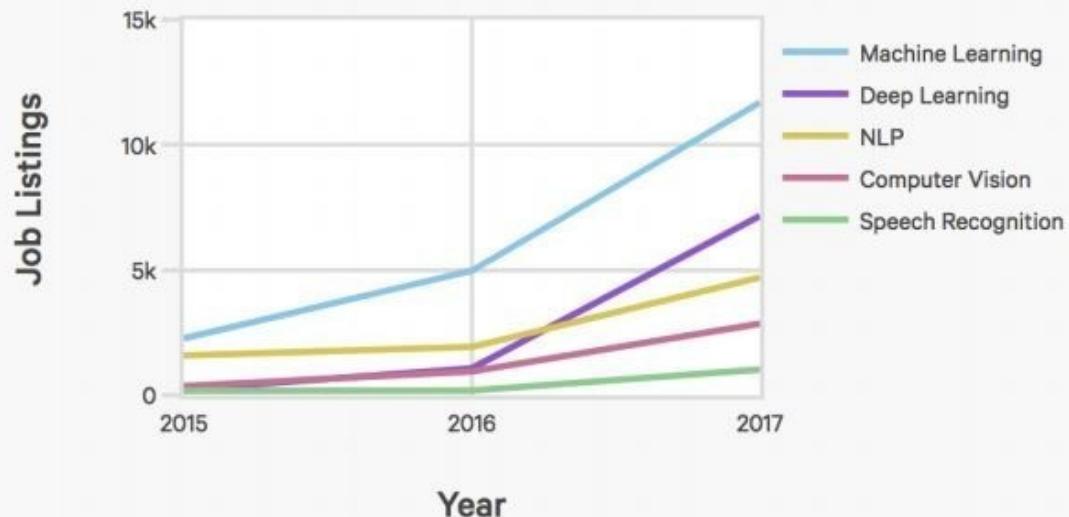
\$8,736

81%

Of DSA jobs require workers with 3-5 years of experience or more

© 2017 Burning Glass Technologies – Proprietary and Confidential

Job Openings, Skills Breakdown (Monster.com)



Source: Monster.com

AIINDEX.ORG

Note: A single AI-related job may be double counted (belong to multiple categories). For example, a job may specifically require natural language processing and computer vision skills.