

# Machine Learning

Fafer77

13 grudnia 2024

## 1 Definitions

1. Measures of central tendency (they describe a central position of your data):
  - (a) Mean
  - (b) Mode
  - (c) Median
2. Measures of spread (describe how spread out your data is - clumped together or spread far apart):
  - (a) Range
  - (b) Quartiles
  - (c) Standard deviation
  - (d) Variance
3. Training set - data used for training
4. Training example (sample) - A single instance used in machine learning to train a model. It typically consists of input features and, in supervised learning, an associated label or target value.
5. Accuracy - the ratio of correctly predicted instances to the total number of instances in the dataset.

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

6. Problem regresyjny: przewidywanie wartości na podstawie cechy wejściowej

7. Wydobywanie cech (ang. feature extraction) - polega na znalezieniu skorelowanych cech i połączenie ich w jedną
8. Classification task - involves predicting a discrete label or category for a given input based on its features e.g 'spam' or 'not spam'
9. Data mining - analyzing huge amount of data to find some patterns
10. Types of learning:
  - (a) Uczenie nadzorowane (ang. supervised learning) - dane uczące przekazywane algorytmowi zawierają dołączone rozwiązania problemu, tzw. etykiety (ang. labels).  
Typowe zadania systemu nadzorowanego:
    - i. Klasyfikacja np. filtr spamu
    - ii. Przewidywanie docelowej (ang. target) wartości numerycznej np. cena samochodu przy użyciu określonego zbioru cech. Ten typ zadania nosi nazwę regresji.
  - (b) Uczenie nienadzorowane (eng. unsupervised learning) - dane uczące są nieoznakowane. System próbuje się uczyć bez nauczyciela.
    - i. Analiza skupień
    - ii. Algorytm wizualizujący, redukcja wymiarowości - cel uproszczenie danych bez utraty nadmiernej ilości informacji.
    - iii. Wykrywanie anomalii (ang. anomaly detection) - np. nietypowe transakcje karty kredytowej w celu zapobieganiu nielegalnym operacjom, wykrywanie usterek produkcyjnych.
    - iv. Wykrywanie nowości (ang. novelty detection)
    - v. Uczenie przy użyciu reguł asocjacyjnych (ang. association rule learning) - analiza ogromnej ilości danych i wykrycie interesujących zależności pomiędzy atrybutami.
  - (c) Uczenie półnadzorowane (ang. semisupervised learning) - część danych jest oznakowana, a większość nie, bo etykietowanie jest czasochłonne i kosztowne.
  - (d) Uczenie samonadzorowane (ang. self-supervised learning) - wygenerowanie w pełni oznakowanego zestawu danych z zestawu całkowicie nieoznakowanego.
  - (e) Uczenie transferowe (ang. transfer learning) - korzysta się w głębokich sieciach neuronowych.

- (f) *Uczenie przez wzmacnianie* (ang. reinforcement learning) - system uczący tzw. agent może obserwować środowisko, dobierać i wykonywać czynności, a także odbierać nagrody lub kary. Potem uczy się samodzielnie najlepszej strategii (ang. policy), aby uzyskiwać jak największą nagrodę. Polityka definiuje rodzaj działania, jakie agent powinien wybrać w danej sytuacji.
- (g) *Uczenie wsadowe* (ang. batch learning) - system nie jest w stanie trenować przyrostowo - do jego anuki muszą wystarczyć wszystkie dostępne dane (zużywa zwykle dużo czasu i zasobów, dlatego zwykle w trybie offline). System najpierw jest uczony, a potem wdrożony do cyklu produkcyjnego i już więcej nie jest trenowany; korzysta jedynie z dotychczas zdobytych informacji. Tzw. uczenie offline (ang. offline learning). Następuje rozkład modelu (ang. model rot) albo dryf danych (ang. data drift). Rozwiązanie: systematyczne trenowanie modelu, zależne od problemu. System trenuje się od podstaw na starym i nowym zestawie za każdym razem.
- (h) *Uczenie przyrostowe* (ang. online learning) - trenowany jest na bieżąco poprzez sekwencyjne dostarczanie danych, które mogą być pojedyncze lub przejoyować postać tzw. minipakietów (mini-batches). Każdy krok uczący jest szybki i niezbyt kosztowny.