# Wprowadzenie do Maszynowego Uczenia

## Podstawowe pojęcia i przykłady

dr hab. Andrzej Jankowski, prof. UWM

3 lutego 2025

# Agenda

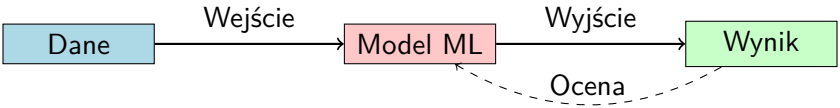
# Definicja uczenia maszynowego

## Co to jest ML?

Uczenie maszynowe (Machine Learning) to dziedzina informatyki, która pozwala komputerom **uczyć się** na podstawie danych i **wyciągać wnioski** bez wyraźnego zaprogramowania każdego kroku.

- Komputer **sam** dostosowuje swoje zachowanie w oparciu o doświadczenie (dane).
- Przykład: Rozpoznawanie zdjęć kotów i psów, przewidywanie pogody, rekomendacje filmów.

# Wizualizacja procesu uczenia maszynowego



# Rodzaje uczenia maszynowego

- **Uczenie nadzorowane (z nauczycielem)**
- **Uczenie nienadzorowane**
- **Uczenie przez wzmacnianie**

## Intuicja

- *Nadzorowane*: Mamy przykłady z prawidłowymi odpowiedziami.
- *Nienadzorowane*: Nie wiemy, jakie są prawidłowe odpowiedzi, odkrywamy ukryte struktury w danych.
- *Przez wzmacnianie*: System uczy się przez **nagrody** i **kary** za swoje decyzje.

# Uczenie nadzorowane - szczegóły

**W uczeniu nadzorowanym mamy "czarną skrzynkę"(model ML), która:**

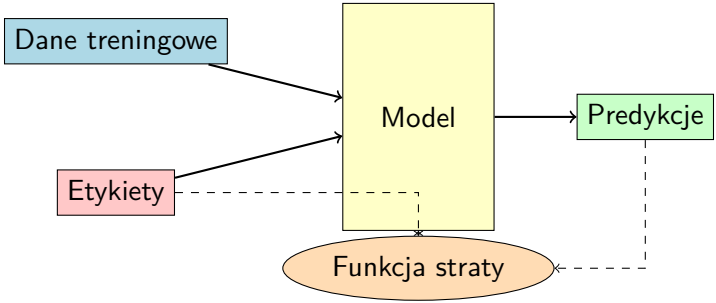
- Otrzymuje na wejściu dane treningowe (np. zdjęcia z opisem, że to jabłko lub gruszka),
- Wyjściem są przewidywane wyniki (np. informacja, czy na obrazku jest jabłko czy gruszka),
- Posiada **funkcję straty**, która mierzy jakość przewidywań.

**Proces uczenia krok po kroku:**

1. Model otrzymuje dane wejściowe.
2. Generuje przewidywania.
3. Porównujemy przewidywania z rzeczywistymi wartościami.
4. Obliczamy funkcję straty (różnicę między przewidywaniem a prawdą).
5. Dostosowujemy **parametry modelu**, aby zmniejszyć stratę.
6. Powtarzamy proces w pętli wiele razy.

**Cel:** *Minimalizować funkcję straty, np. za pomocą* ***metod gradientowych.***

# Wizualizacja uczenia nadzorowanego



# Parametry modelu

- Model (intuicyjnie tzw. "czarna skrzynka") posiada **wiele parametrów** (intuicyjnie tzw. "śrubek"). Bez zrozumienia modelu, w tym roli najważniejszych parametrów "sróbek"praktycznie niemożliwe jest prawidłowe maszynowe uczenie z nietrywialnych danych.
- Liczba parametrów: od **kilkunastu** do **setek miliardów** (np. w dużych modelach językowych – LLM).
- Dostosowanie parametrów modelu podczas treningu to kluczowy element **uczenia się**.

## Przykład intuicyjny

Wyobraźmy sobie, że **model** to robot kuchenny z mnóstwem pokręteł do regulacji. Aby idealnie ubić śmietanę, musimy ustawić pokrętła w *odpowiednich pozycjach*. Podczas próby ubijania (treningu) sprawdzamy, czy śmietana jest idealnie ubita (wynik). Jeśli nie, poprawiamy ustawienia pokręteł (parametrów) i próbujemy ponownie.

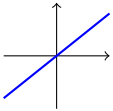
# Najczęściej spotykane modele – intuicje

1. **Regresja logistyczna**
   - Jak *narysowanie linii*, która najlepiej rozdziela dwie grupy (np. jabłka vs. gruszki).
   - Prosty model do klasyfikacji binarnej (tak/nie).
2. **Drzewo decyzyjne**
   - Jak *gra w "20 pytań"*: zadaje pytania, a odpowiedzi prowadzą do decyzji.
   - Przykład: Czy jest okrągłe? Czy ma kolor zielony? Itp.
3. **Las losowy (Random Forest)**
   - *Grupa wielu drzew decyzyjnych*, które głosują nad najlepszą odpowiedzią.
   - Zazwyczaj dokładniejszy niż pojedyncze drzewo.
4. **Naive Bayes**
   - Jak *detektyw*, który sprawdza częstotliwość występowania różnych wskazówek razem.
   - Szybka metoda do kategoryzacji tekstu (np. filtry antyspamowe).

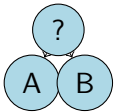
# Najczęściej spotykane modele – ciąg dalszy

5. **Support Vector Machines (SVM)**
   - Szuka *najlepszej "granicy"* między różnymi grupami.
   - Chce zostawić jak najwięcej miejsca po obu stronach granicy.
6. **K-Nearest Neighbors (KNN)**
   - Patrzy na *najbliższych sąsiadów* w zbiorze danych.
   - Decyzja zależy od tego, do jakiej grupy należy większość sąsiadów.
7. **Gradient Boosting**
   - *Zespół wielu prostych modeli*, gdzie każdy kolejny uczy się z błędów poprzedników.
   - Często wygrywa w konkursach data science (np. XGBoost, LightGBM).
8. **Sieci neuronowe**
   - Inspirowane działaniem *mózgu* – wiele **neuronom** (punktów) przekazuje sobie sygnały.
   - Stosowane w rozpoznawaniu obrazów, głosu, języka.

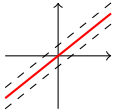
# Wizualizacja modeli ML



Regresja




Drzewo



SVM

# Walidacja krzyżowa (k-fold cross-validation)

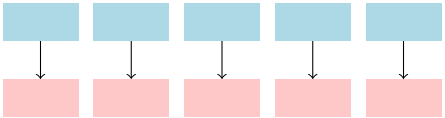
## Dlaczego testujemy model?

Chcemy sprawdzić, czy nasz model **rzeczywiście** się nauczył uogólniać, a nie tylko zapamiętał przykłady.

## k-fold cross-validation

1. Dzielimy dane na **k** równych części (foldów).
2. **k-1** części używamy do trenowania modelu.
3. Ostatnią (**1**) część wykorzystujemy do testowania.
4. Powtarzamy proces **k** razy, za każdym razem zmieniając tę część testową.
5. Wyniki **uśredniamy**, aby uzyskać ocenę jakości modelu.

# Walidacja krzyżowa



Walidacja 5-krotna

# Przeuczenie - co to jest?

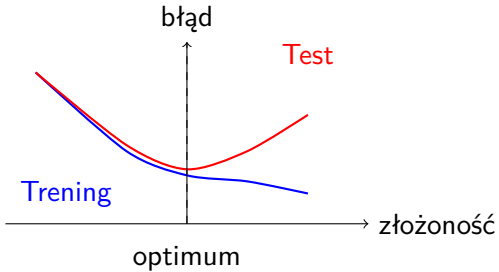
## Definicja

**Przeuczenie** (*overfitting*) oznacza, że model *zbyt dokładnie* "zapamiętuje"dane treningowe i nie potrafi uogólniać. To jak uczeń, który **wykuł** odpowiedzi na pamięć, ale **nie** rozumie materiału.

**Objawy:**

- Bardzo niska strata na zbiorze treningowym.
- Duża strata/błędy na nowym, nieznanym zbiorze testowym.

# Problem przeuczenia



# Jak unikać przeuczenia?

- **Więcej danych treningowych** – im więcej przykładów, tym trudniej "wykuć"je na pamięć.
- **Uproszczenie modelu** – mniej parametrów to mniejsze ryzyko przeuczenia.
- **Regularyzacja** – specjalne techniki, które *karzą* model za zbyt duże wagi.
- **Wczesne zatrzymywanie (early stopping)** – kończymy trening, gdy model zaczyna przeuczać się.
- **Walidacja krzyżowa** – pozwala obiektywnie ocenić, jak model radzi sobie na nieznanych danych.
- **Dropout (w sieciach neuronowych)** – losowe "wyłączanie"neuronów w trakcie treningu, aby uniknąć przeuczenia.

## Przykład intuicyjny

Jeśli uczeń za bardzo **wykuwa** wypracowanie słowo w słowo, może nie zrozumieć istoty tematu. Podobnie model, który jest zbyt "dopasowany"do treningowych przykładów, nie radzi sobie z **nowymi** pytaniami.

# Podsumowanie

- Maszynowe Uczenie to nauka o budowaniu modeli, które uczą się z danych.
- Wyróżniamy **uczenie nadzorowane**, **nienadzorowane** i **przez wzmacnianie**.
- **Uczenie nadzorowane** opiera się na danych z prawidłowymi odpowiedziami.
- Istnieje wiele modeli ML (regresja logistyczna, drzewa, sieci neuronowe i inne).
- Ważne jest **testowanie modelu** i unikanie **przeuczenia**.

## Do czego możemy wykorzystać ML?

- Rozpoznawanie twarzy, mowy, pisma.
- Personalizowane reklamy i rekomendacje filmów.
- Systemy wspomagające diagnozę medyczną.
- Automatyczne tłumaczenia, czatboty, itp.

Dziękuję za uwagę
&
zapraszam do dyskusji!