# Uczenie Maszynowe: Wprowadzenie

(c) Marcin Sydow

# Plan

- Dane
- Eksploracja danych i uczenie maszynowe: motywacja
- Na czym polega uczenie z danych
- Tablice decyzyjne: atrybuty i obserwacje
- Uczenie z nadzorem i bez nadzoru
- Klasyfikacja i regresja
- Przykłady

# Dane: Motywacja dla eksploracji danych



Obserwacje:

1. Danych jest dużo, są interesujące ale trudne do analizy przez człowieka
2. są w formie elektronicznej

Ergo: zaprząc do tego algorytmy i komputery

# Zalew danych



W każdej sekundzie produkowane są ogromne ilości danych:

- odwiedzenia stron WWW
- dzienne ceny ropy
- notowania partii politycznych
- zapytania do wyszukiwarek
- kliknięcia (logi serwerów WWW)
- zamówienia towarów w sklepach internetowych
- rachunki w elektronicznych kasach sklepowych
- wyniki pomiarów astronomicznych, fizycznych, etc...

# Przykładowe zadania

- grupowanie obiektów podobnych
- rozpoznawanie istotnych **wzorców** w danych
- klasyfikacja nowo-obserwowanych przypadków
- przewidywanie przyszłości na podstawie poprzednich obserwacji
- wykrywanie trendów w danych (np. wczesne wykrycie kryzysów ekonomicznych, itp.)

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

# Podział



- Uczenie z nadzorem
- Uczenie bez nadzoru

# Typowe fazy w uczeniu maszynowym



- zbieranie danych
- czyszczenie i wstępne przetworzenie danych
- (tylko w uczeniu z nadzorem) podział na zbiór treningowy i testowy
- uczenie się na danych
- ewaluacja (iteracyjnie)
- używanie systemu do zadań

# Uczenie z nadzorem



1. podawanie systemowi "prawidłowych" rozwiązań w tzw **zbiorze danych treningowych** (sygnał uczący)
2. system "uczy się" (dane treningowe) uogólnić sposób rozwiązania zadania poprzez automatyczne wykrycie związków pomiędzy danymi a prawidłowymi rozwiązaniami (automatyczne budowanie modelu prawidłowego rozwiązania)
3. automatycznie "wyuczony" model jest stosowany do nowych przypadków (nie trenujących)

# Uczenie bez Nadzoru



- brak sygnału uczącego (surowe dane)
- cel: wykrycie pewnych związków między obiektami i atrybutami (np. grupowanie, reguły asocjacyjne)

# Tablica Decyzyjna



## Przykład - diagnostyka okulistyczna.

| wiek | presc. | astygmatyzm | łzawienie | OKULARY |
|---|---|---|---|---|
| młody | myope | nie | niskie | zbędne |
| młody | myope | nie | normalne | lekkie |
| młody | myope | yes | niskie | zbędne |
| młody | myope | tak | normalne | mocne |
| młody | hypermetrope | nie | niskie | zbędne |
| młody | hypermetrope | nie | normalne | lekkie |
| młody | hypermetrope | tak | niskie | zbędne |
| młody | hypermetrope | tak | normalne | mocne |
| pre-presbyopic | myope | nie | niskie | zbędne |
| pre-presbyopic | myope | nie | normalne | lekkie |
| pre-presbyopic | myope | tak | niskie | zbędne |
| pre-presbyopic | myope | tak | normalne | mocne |
| pre-presbyopic | hypermetrope | nie | niskie | zbędne |
| pre-presbyopic | hypermetrope | nie | normalne | lekkie |
| pre-presbyopic | hypermetrope | tak | niskie | zbędne |
| pre-presbyopic | hypermetrope | tak | normalne | zbędne |
| presbyopic | myope | nie | niskie | zbędne |
| presbyopic | myope | nie | normalne | zbędne |
| presbyopic | myope | tak | niskie | zbędne |
| presbyopic | myope | tak | normalne | mocne |
| presbyopic | hypermetrope | nie | niskie | zbędne |
| presbyopic | hypermetrope | nie | normalne | lekkie |
| presbyopic | hypermetrope | tak | niskie | zbędne |
| presbyopic | hypermetrope | tak | normalne | zbędne |

# Przykład: nieznana gra, możliwa tylko przy pewnych specyficznych warunkach atmosferycznych (nie wiemy jakich):

| pogoda | temperatura | wilgotność | wiatr | **GRAĆ?** |
|---|---|---|---|---|
| słonecznie | ciepło | wysoka | brak | nie |
| słonecznie | ciepło | wysoka | jest | nie |
| pochmurno | ciepło | wysoka | brak | tak |
| deszczowo | normalnie | wysoka | brak | tak |
| deszczowo | chłodno | normalna | brak | tak |
| deszczowo | chłodno | normalna | jest | nie |
| pochmurno | chłodno | normalna | jest | tak |
| słonecznie | normalnie | wysoka | brak | nie |
| słonecznie | chłodno | normalna | brak | tak |
| deszczowo | normalnie | normalna | brak | tak |
| słonecznie | normalnie | normalna | jest | tak |
| pochmurno | normalnie | wysoka | jest | tak |
| pochmurno | ciepło | normalna | brak | tak |
| deszczowo | normalnie | wysoka | jest | nie |

# Przykład, cd

Zadanie:
"Przewidzieć przy jakich warunkach gra się w tę grę?"

Jeśli odpowiedź nie jest znana można posłużyć się wieloma zaobserwowanymi przypadkami i sprawić aby system wychwycił ogólną regułę.

Jeśli uda się w ten automatyczny sposób pozyskać "wiedzę" o regułach gry z obserwacji znanych przypadków można ją następnie zastosować do przypadków nieznanych

# Nowy przypadek

| outlook | temperature | humidity | windy | PLAY? |
|---|---|---|---|---|
| słonecznie | ciepło | wysoka | brak | nie |
| słonecznie | ciepło | wysoka | jest | nie |
| **pochmurno** | ciepło | wysoka | brak | tak |
| deszczowo | normalnie | wysoka | brak | tak |
| deszczowo | chłodno | normalna | brak | tak |
| deszczowo | chłodno | normalna | jest | nie |
| **pochmurno** | chłodno | normalna | jest | tak |
| słonecznie | normalnie | wysoka | brak | nie |
| słonecznie | chłodno | normalna | brak | tak |
| deszczowo | normalnie | normalna | brak | tak |
| słonecznie | normalnie | normalna | jest | tak |
| **pochmurno** | normalnie | wysoka | jest | tak |
| **pochmurno** | ciepło | normalna | brak | tak |
| deszczowo | normalnie | wysoka | jest | nie |
| **pochmurno** | chłodno | wysoka | jest | ??? |

# Tablica decyzyjna: obserwacje i atrybuty



Wiedza może być budowana w oparciu o poprzednio zaobserwowane dane:

Każda obserwacja (przypadek) opisana za pomocą **atrybutów** określonego typu (nominalnego albo numerycznego)

**Tablica Decyzyjna:**

- **obserwacje** (przypadki) = wiersze
- **atrybuty** = kolumny

# Atrybuty

- numeryczne albo kategoryczne
- uporządkowane lub nie
- przeskalowanie, transformacje atrybutów
- kwantyzacja (zamiana z numerycznych na kategoryczne)

# Tabela Decyzyjna: atrybuty nominalne

| pogoda | temperatura | wilgotność | wiatr | GRAĆ? |
|---|---|---|---|---|
| słonecznie | ciepło | wysoka | brak | nie |
| słonecznie | ciepło | wysoka | jest | nie |
| pochmurno | ciepło | wysoka | brak | tak |
| deszczowo | normalnie | wysoka | brak | tak |
| deszczowo | chłodno | normalna | brak | tak |
| deszczowo | chłodno | normalna | jest | nie |
| pochmurno | chłodno | normalna | jest | tak |
| słonecznie | normalnie | wysoka | brak | nie |
| słonecznie | chłodno | normalna | brak | tak |
| deszczowo | normalnie | normalna | brak | tak |
| słonecznie | normalnie | normalna | jest | tak |
| pochmurno | normalnie | wysoka | jest | tak |
| pochmurno | ciepło | normalna | brak | tak |
| deszczowo | normalnie | wysoka | jest | nie |

# Tabela Decyzyjna: atrybuty numeryczne

| pogoda | temperatura (F) | wilgotność | wiatr | GRAĆ? |
|---|---|---|---|---|
| słonecznie | 85 | 85 | brak | nie |
| słonecznie | 80 | 90 | jest | nie |
| pochmurno | 83 | 86 | brak | tak |
| deszczowo | 70 | 96 | brak | tak |
| deszczowo | 68 | 80 | brak | tak |
| deszczowo | 65 | 70 | jest | nie |
| pochmurno | 64 | 65 | jest | tak |
| słonecznie | 72 | 95 | brak | nie |
| słonecznie | 69 | 70 | brak | tak |
| deszczowo | 75 | 80 | brak | tak |
| słonecznie | 75 | 70 | jest | tak |
| pochmurno | 72 | 90 | jest | tak |
| pochmurno | 81 | 75 | brak | tak |
| deszczowo | 71 | 91 | jest | nie |

# Inne formy danych



Dane nie muszą być w formie prostokątnej tablicy

- logi (np. serwerów)
- dane relacyjne (np. w sieciach społecznych)
- dane sekwencyjne (np. bioinformatyka)
- dane grafowe, etc.

# Uczenie Maszynowe

Zadanie: "nauczyć się" relacji pomiędzy wartościami atrybutów

Dwa główne podejścia:

1. Uczenie z nadzorem
2. Uczenie bez nadzoru

# Uczenie z nadzorem



1. **atrybut decyzyjny**: wyszczególniony atrybut w tabeli decyzyjnej (np. "GRAĆ?")
2. Zadanie: "przewidzieć" prawidłową (nieznaną) wartość atrybutu decyzyjnego na podstawie (znanych) wartości pozostałych atrybutów
3. Wykorzystać do tego **zbiór treningowy** - tj taki zbiór obserwacji (przypadków), dla których prawidłowa wartość atrybutu decyzyjnego (oraz wszystkich pozostałych atrybutów) jest znana

Uczenie z nadzorem nazywane jest:

- **klasyfikacją**, gdy przewidywany atrybut decyzyjny jest nominalny
- **regresją**, gdy przewidywany atrybut decyzyjny jest numeryczny

# Podsumowanie idei uczenia z nadzorem



Cel:
**input:** nowy przypadek (obserwacja) z nieznaną wartością atrybutu decyzyjnego

**output:** "prawidłowa" wartość atrybutu decyzyjnego

System może "uczyć się" tylko na ograniczonej liczbie znanych przypadków (zbiór treningowy) dodstarczonych przez nadzorującego

Problemy praktyczne:

- brakujące wartości (jak je wypełnić?)
- błędne wartości (jak je wykryć i poprawić?)
- dane zaszumione (jak je "odszumić"?)
- dane sprzeczne (co z tym zrobić?)

# Przykład zadania klasyfikacji

Botanika: rozpoznawanie gatunków roślin (dane "Iris")

Rozpatrzmy 3 różne podgatunki kwiatu o łac. nazwie Iris:

- Iris-setosa
- Iris-versicolor
- Iris-virginica

Task: nauczyć się rozpoznawać gatunek rośliny na podstawie rozmiarów liści i płatków (atrybuty):

- długość listka (cm)
- szerokość listka (cm)
- długość płatka (cm)
- szerokość płatka (cm)

# Rozpoznawanie roślin, cont.



Zbiór trenujący:
150 znanych przypadków (zmierzone części roślin i znana prawidłowa klasyfikacja)

System "uczy się" na zbiorze treningowym

Następnie, każdy nowy (nieznany) przypadek jest klasyfikowany na podstawie pomiarów płatków i listków
Automatycznie "wyuczona" wiedza jest stosowana do klasyfikacji nowych przypadków (dla których prawidłowa odpowiedź nie musi być znana przez nadzorującego proces)

# Zbiór danych (fragment)



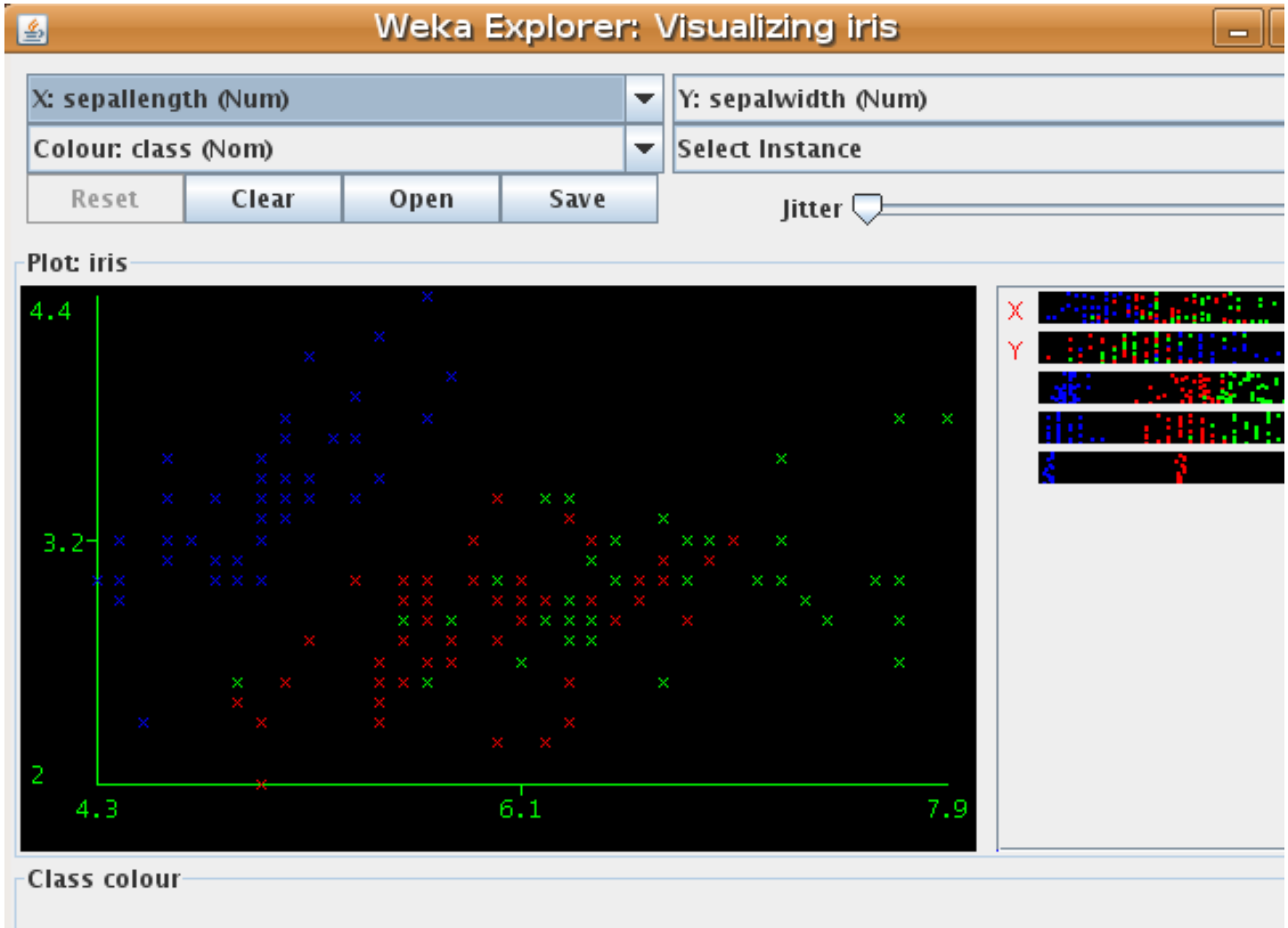
S - iris setosa, V - iris versicolor, VG - iris virginica

| sl | sw | pl | pw | ? | sl | sw | pl | pw | ? | sl | sw | pl | pw | ? |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 5.1 | 3.5 | 1.4 | 0.2 | S | 7.0 | 3.2 | 4.7 | 1.4 | V | 6.3 | 3.3 | 6.0 | 2.5 | VG |
| 4.9 | 3.0 | 1.4 | 0.2 | S | 6.4 | 3.2 | 4.5 | 1.5 | V | 5.8 | 2.7 | 5.1 | 1.9 | VG |
| 4.7 | 3.2 | 1.3 | 0.2 | S | 6.9 | 3.1 | 4.9 | 1.5 | V | 7.1 | 3.0 | 5.9 | 2.1 | VG |
| 4.6 | 3.1 | 1.5 | 0.2 | S | 5.5 | 2.3 | 4.0 | 1.3 | V | 6.3 | 2.9 | 5.6 | 1.8 | VG |
| 5.0 | 3.6 | 1.4 | 0.2 | S | 6.5 | 2.8 | 4.6 | 1.5 | V | 6.5 | 3.0 | 5.8 | 2.2 | VG |
| 5.4 | 3.9 | 1.7 | 0.4 | S | 5.7 | 2.8 | 4.5 | 1.3 | V | 7.6 | 3.0 | 6.6 | 2.1 | VG |
| 4.6 | 3.4 | 1.4 | 0.3 | S | 6.3 | 3.3 | 4.7 | 1.6 | V | 4.9 | 2.5 | 4.5 | 1.7 | VG |
| 5.0 | 3.4 | 1.5 | 0.2 | S | 4.9 | 2.4 | 3.3 | 1.0 | V | 7.3 | 2.9 | 6.3 | 1.8 | VG |
| 4.4 | 2.9 | 1.4 | 0.2 | S | 6.6 | 2.9 | 4.6 | 1.3 | V | 6.7 | 2.5 | 5.8 | 1.8 | VG |
| 4.9 | 3.1 | 1.5 | 0.1 | S | 5.2 | 2.7 | 3.9 | 1.4 | V | 7.2 | 3.6 | 6.1 | 2.5 | VG |
| 5.4 | 3.7 | 1.5 | 0.2 | S | 5.0 | 2.0 | 3.5 | 1.0 | V | 6.5 | 3.2 | 5.1 | 2.0 | VG |
| 4.8 | 3.4 | 1.6 | 0.2 | S | 5.9 | 3.0 | 4.2 | 1.5 | V | 6.4 | 2.7 | 5.3 | 1.9 | VG |
| 4.8 | 3.0 | 1.4 | 0.1 | S | 6.0 | 2.2 | 4.0 | 1.0 | V | 6.8 | 3.0 | 5.5 | 2.1 | VG |
| 4.3 | 3.0 | 1.1 | 0.1 | S | 6.1 | 2.9 | 4.7 | 1.4 | V | 5.7 | 2.5 | 5.0 | 2.0 | VG |
| 5.8 | 4.0 | 1.2 | 0.2 | S | 5.6 | 2.9 | 3.6 | 1.3 | V | 5.8 | 2.8 | 5.1 | 2.4 | VG |
| 5.7 | 4.4 | 1.5 | 0.4 | S | 6.7 | 3.1 | 4.4 | 1.4 | V | 6.4 | 3.2 | 5.3 | 2.3 | VG |
| 5.4 | 3.9 | 1.3 | 0.4 | S | 5.6 | 3.0 | 4.5 | 1.5 | V | 6.5 | 3.0 | 5.5 | 1.8 | VG |
| 5.1 | 3.5 | 1.4 | 0.3 | S | 5.8 | 2.7 | 4.1 | 1.0 | V | 7.7 | 3.8 | 6.7 | 2.2 | VG |
| 5.7 | 3.8 | 1.7 | 0.3 | S | 6.2 | 2.2 | 4.5 | 1.5 | V | 7.7 | 2.6 | 6.9 | 2.3 | VG |
| 5.1 | 3.8 | 1.5 | 0.3 | S | 5.6 | 2.5 | 3.9 | 1.1 | V | 6.0 | 2.2 | 5.0 | 1.5 | VG |
| 5.4 | 3.4 | 1.7 | 0.2 | S | 5.9 | 3.2 | 4.8 | 1.8 | V | 6.9 | 3.2 | 5.7 | 2.3 | VG |
| 5.1 | 3.7 | 1.5 | 0.4 | S | 6.1 | 2.8 | 4.0 | 1.3 | V | 5.6 | 2.8 | 4.9 | 2.0 | VG |
| 5.0 | 3.0 | 1.6 | 0.2 | S | 6.6 | 3.0 | 4.4 | 1.4 | V | 7.2 | 3.2 | 6.0 | 1.8 | VG |
| 5.0 | 3.4 | 1.6 | 0.4 | S | 6.8 | 2.8 | 4.8 | 1.4 | V | 6.2 | 2.8 | 4.8 | 1.8 | VG |

# Wizualizacja zbioru danych: rzut na płaszczyznę 2-wym.



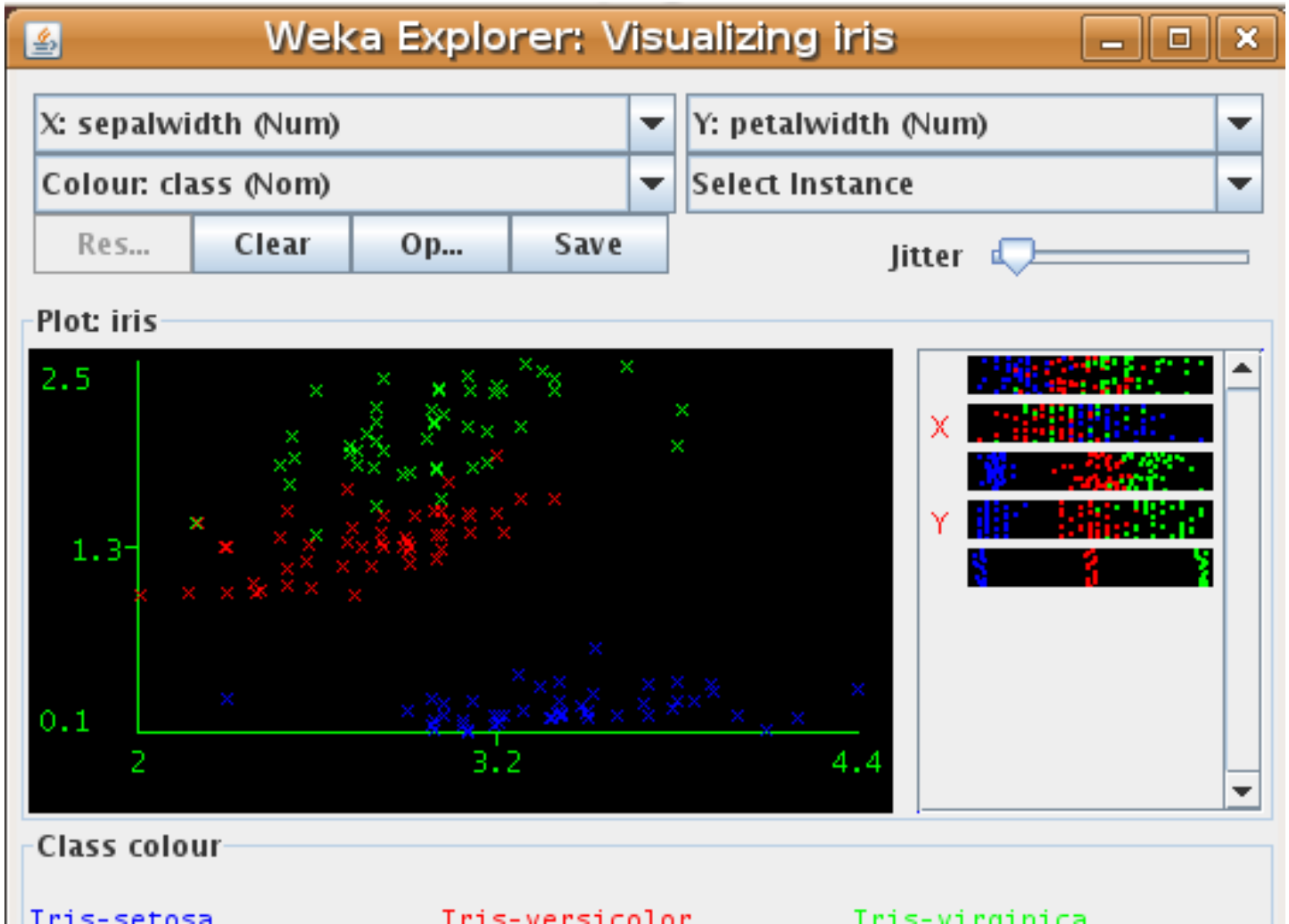
(zbiór jest 4-wymiarowy) np.: szerokość/długość listka - nie jest to wystarczająca informacja



# Inna wizualizacja rzutu na płaszczyznę 2-wym.



szerokość listka/długość płatka - niesie dużo “wiedzy” (tzw. dobry dyskriminant)



# W jaki sposób system sam "uczy się" zależności?



Istnieje wiele podejść/modeli, przykłady:

- metoda k najbliższych sąsiadów (kNN)
- Oparte na regułach decyzyjnych
- Drzewa decyzyjne
- Podejście Bayesowskie
- Regresja liniowa
- Sztuczne Sieci Neuronowe (Perceptron, sieci wielo-wartswowe)
- SVM (support vector machines)
- wiele innych...

# Inne przykłady problemu klasyfikacji



- Maszynowe rozpoznawanie ręcznie pisanych cyfr na formularzach
- Klasyfikacja zdolności kredytowej klienta banku
- Identyfikacja chłamu pocztowego (ang. e-mail spam)
- Automatyczne rozróżnianie wycieków oleju z tankowców od ciepłych prądów na podstawie zdjęć satelitarnych
- Maszynowa identyfikacja języka w dokumentach tekstowych (np. portugalski czy hiszpański, itp.)
- Automatyczna klasyfikacja tematu dokumentu elektronicznego (do jednej z kilku kategorii)
- Identyfikacja tzw. chłamu wyszukiwarkowego (ang. Search Engine Spam)

# Zadanie Regresji



W zadaniu klasyfikacji system "przewidywał" wartość atrybutu decyzyjnego typu nominalnego.
Jeśli natomiast przewidujemy atrybut numerycznego mówimy o **regresji**
Przykłady zadania regresji:

- przewidzieć wartość (cenę) papieru wartościowego na podstawie poprzednich notowań i innych czynników (ekonomicznych, politycznych, etc.)
- oszacować ilościowe zapotrzebowanie na dany towar (np. woda mineralna) w przyszłym tygodniu w supermarkecie na podstawie bieżącej sprzedaży, pory roku, pogody, etc.
- przewidzieć temperaturę powietrza w następnym dniu

# Przykład zadania regresji



Przewidywana skuteczność procesora na podstawie jego parametrów technicznych

Przykładowe atrybuty:

- MYCT cycle time (ns)
- MMIN main memory min
- MMAX main memory max
- CACH cache
- CHMIN channels min
- CHMAX channels max

# Example: regression



| MYCT | MMIN | MMAX | CACH | CHMIN | CHMAX | performance |
|---|---|---|---|---|---|---|
| 125 | 256 | 6000 | 256 | 16 | 128 | 199 |
| 29 | 8000 | 32000 | 32 | 8 | 32 | 253 |
| 29 | 8000 | 16000 | 32 | 8 | 16 | 132 |
| 26 | 8000 | 32000 | 64 | 8 | 32 | 290 |
| 23 | 16000 | 32000 | 64 | 16 | 32 | 381 |
| 23 | 16000 | 32000 | 64 | 16 | 32 | 381 |
| 23 | 16000 | 64000 | 64 | 16 | 32 | 749 |
| 23 | 32000 | 64000 | 128 | 32 | 64 | 1238 |
| 400 | 1000 | 3000 | 0 | 1 | 2 | 23 |
| 400 | 512 | 3500 | 4 | 1 | 6 | 24 |
| 60 | 2000 | 8000 | 65 | 1 | 8 | 70 |
| 50 | 4000 | 16000 | 65 | 1 | 8 | 117 |
| 167 | 524 | 2000 | 8 | 4 | 15 | 23 |
| 143 | 512 | 5000 | 0 | 7 | 32 | 29 |
| 143 | 1000 | 2000 | 0 | 5 | 16 | 22 |
| 110 | 5000 | 5000 | 142 | 8 | 64 | 124 |
| 143 | 1500 | 6300 | 0 | 5 | 32 | 35 |
| 143 | 3100 | 6200 | 0 | 5 | 20 | 39 |
| 143 | 2300 | 6200 | 0 | 6 | 64 | 40 |

# Uczenie bez Nadzoru



Nie dajemy systemowi przykładów (nie dysponujemy). System musi automatycznie "odkryć" zależności pomiędzy danymi. Podstawowe zadania uczenia bez nadzoru:

- grupowanie (ang. clustering)
- wykrywanie przypadków nietypowych (ang. outliers)
- odkrywanie reguł asocjacyjnych

# Grupowanie (ang. clustering)



Należy podzielić wszystkie badane przypadki na grupy obiektów podobnych do siebie (wewnątrz każdej grupy), przy czym obiekty z różnych grup powinny się jak najbardziej różnić między sobą.
Nie wiemy jaka jest faktyczna kategoria odpowiadająca każdej grupie - nie mamy przykładów.
Jest to często wstępny etap analizy danych.
Najprostszy algorytm grupowania: **K-means**

# Wykrywanie przypadków nietypowych (ang. outliers)



Należy automatycznie wykryć obiekty, które z jakichś powodów **odstają** od pozostałych elementów. Mamy tu tylko do dyspozycji same wartości atrybutów. Obiekty wyraźnie odstające od ogółu są w pewnym sensie "podejrzane".
Zastosowania:

- automatyczne wykrywanie włamań do systemów komputerowych
- wykrywanie nadużyć (ang. fraud) w handlu elektronicznym
- wykrywanie "prania brudnych pieniędzy" na podstawie analizy transferów bankowych
- wykrywanie błędów w danych i błędów urządzeń pomiarowych
- czyszczenie danych

# Minimum z tego wykładu:

- Reprezentacja danych w Uczeniu Maszynowym
- Schemat Uczenia Maszynowego (w krokach)
- Na czym polega podział: “z nadzorem” i “bez nadzoru”
- Co to jest klasyfikacja a co to jest regresja
- Przykłady zadań klasyfikacji i regresji (po 3)
- Przykłady zadań uczenia bez nadzoru
- Na czym polega zadanie grupowania (ang. clustering)?
- Przykłady technik uczenia z nadzorem

# Dziękuję za uwagę

Dziękuję za uwagę.