Klasyfikacje nadzorowane

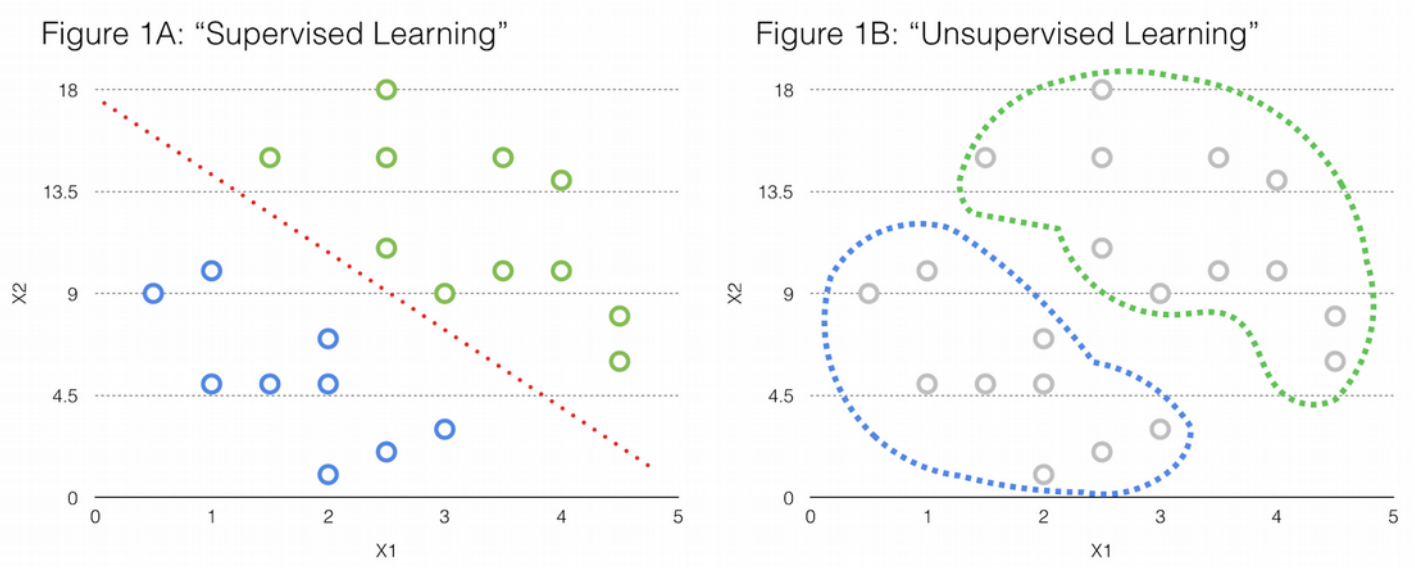
Jarosław Jasiewicz

Eksploracja danych i Uczenie maszynowe

Geoinformacja program magisterski
Specjalność Geoinformatyka

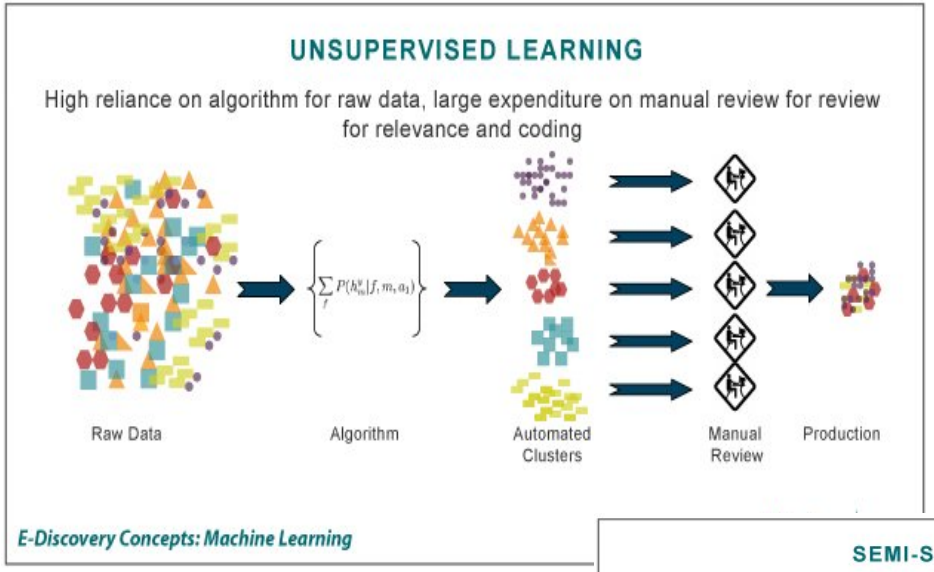
Metody nienadzorowane i nadzorowane

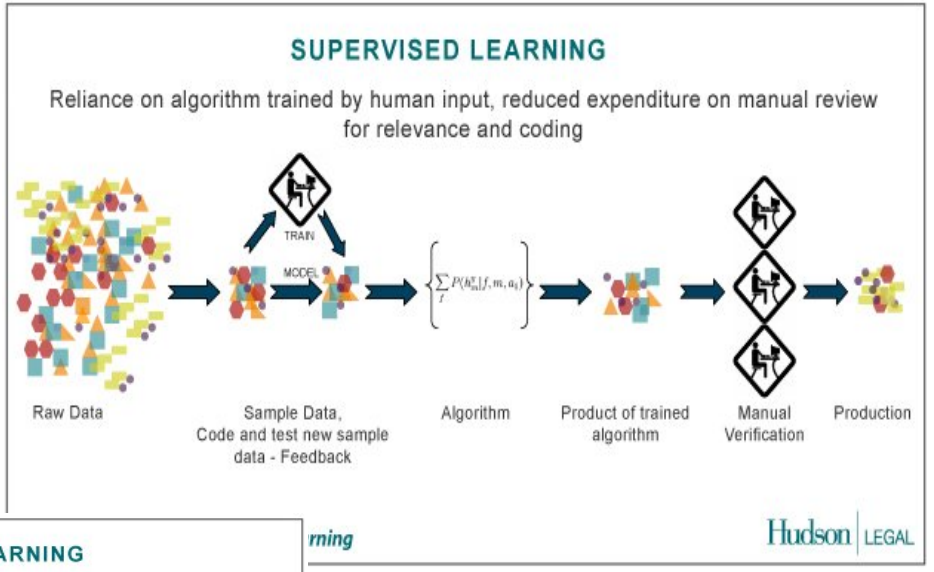
- **Metody nienadzorowane** – brakuje klasy wynikowej (zmiennej zależnej) a celem analizy jest poznanie wewnętrznej struktury danych
- **Metody nadzorowane** - znalezienie relacji pomiędzy atrybutami (zmiennymi niezależnymi) a klasą wynikową (zmienną zależną). Relacja jest następnie używana do predykcji (przewidzenia) klasy wynikowej dla obiektów, gdzie nie jest ona znana

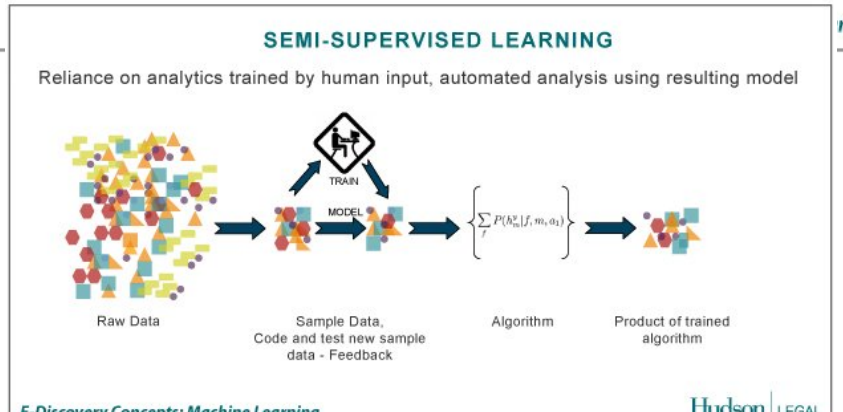


Rola człowieka

- Metody nienadzorowane: interpretacja a posteriori
- Metody nadzorowane: klasy a priori







Rodzaje klasyfikacji

- Klasyfikacja dwuklasowa (binary)
- Klasyfikacja wieloklasowa (multiclass)
- Klasyfikacja jednoklasowa (one-class)
- Wykrywanie nowości i anomalii (anomaly)
- Klasyfikacja wieloetykietowa (multi-label)

Klasyfikacja dwu- i wieloklasowa

- Klasyfikacja dwuklasowa zakłada, przynależność klasyfikowanego obiektu do jednej z dwóch klas, w których druga klasa jest dopełnieniem pierwszej
- Klasyfikacja dwuklasowa nie zakłada symetryczności grup, oraz zakłada że jedna z grup jest wyróżniona (**las** vs. **Nie-las** a nie **łąka**)
- Klasyfikacja wieloklasowa zakłada istenie wielu klas (min. 3), z których każda klasa jest jednakowo istotna
- Wiele algorytmów może pracować tylko w trybie klasyfikacji binarnej, gdzie zakłada się model wielokrotny model **1 vs. Reszta**, dla każdej z klasy z osobna

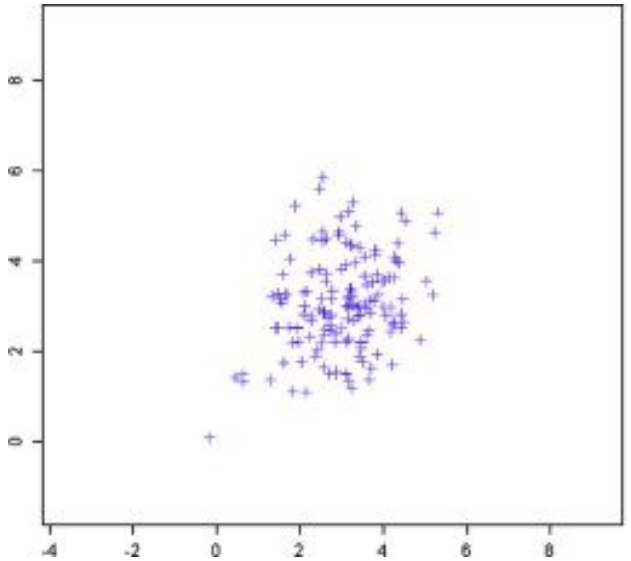
Klasyfikacja wieloetykietowa

- Problem klasyfikacyjny, gdzie jeden obiekt może być opisany więcej niż jedną etykietą (należeć do więcej niż jednej z klas)
- Mylona z klasyfikacją wieloklasową i z klasyfikacją rozmytą
- Nie ma ograniczeń co do ilości klas
- Przykłady: tematy publikacji, typ filmu, zawartość zdjęcia itp.

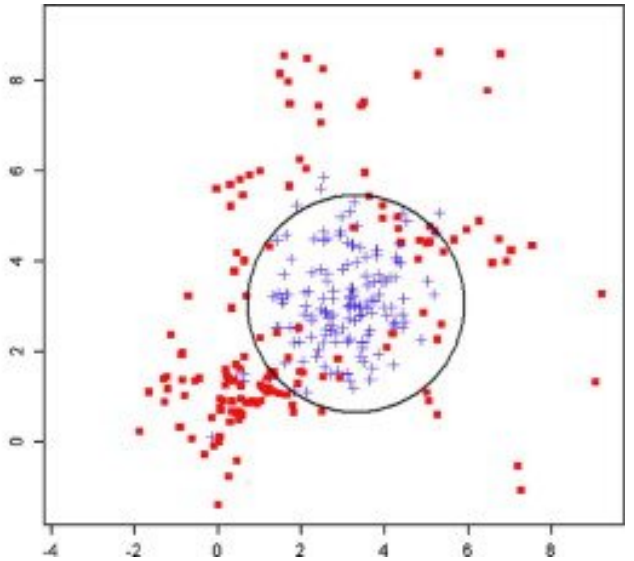
Klasyfikacja na podstawie jednej klasy

- Klasyfikacja, (one-class classification) której zbiór treningowy zawiera tylko obiekty należące do jednej klasy, nie ma klasy przeciwnej, z tego powodu jest zadanie trudniejsze niż standardowa klasyfikacja

Zbiór treningowy

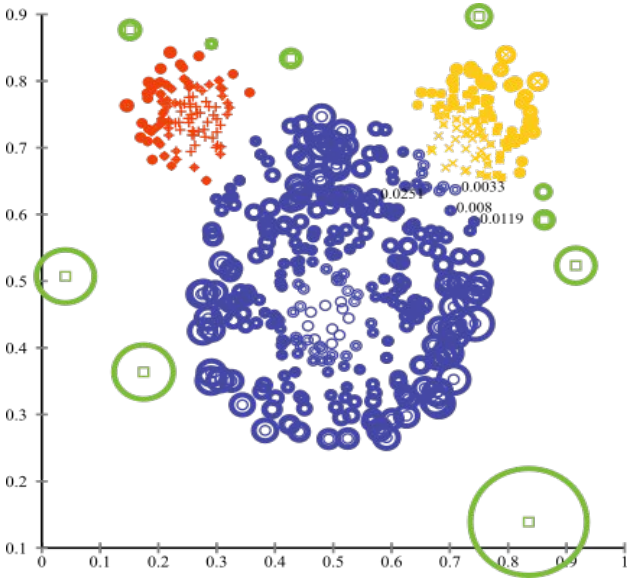


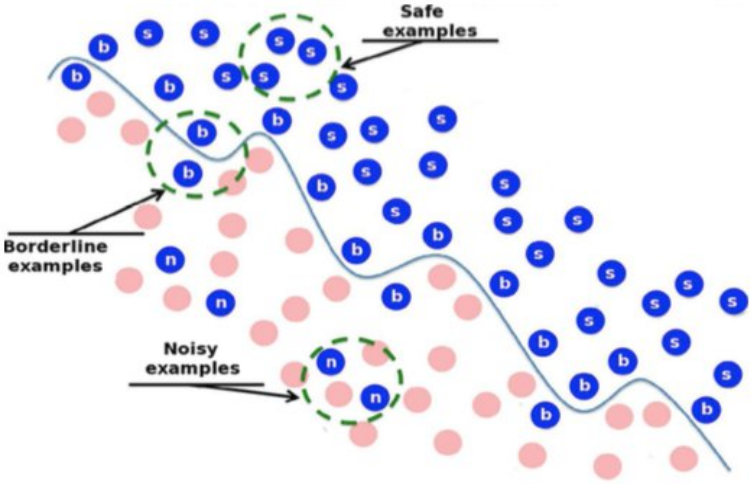
Predykcja na nowych danych



Wykrywanie nowości/anomalii

- wykrywanie rzadkich, nietypowych obserwacji, które różnią się od większości danych
  - Wykrywanie anomalii nie pasujących do żadnej z istniejących klas
  - Obiekty zakłócające, niepasujące i istniejących klasach





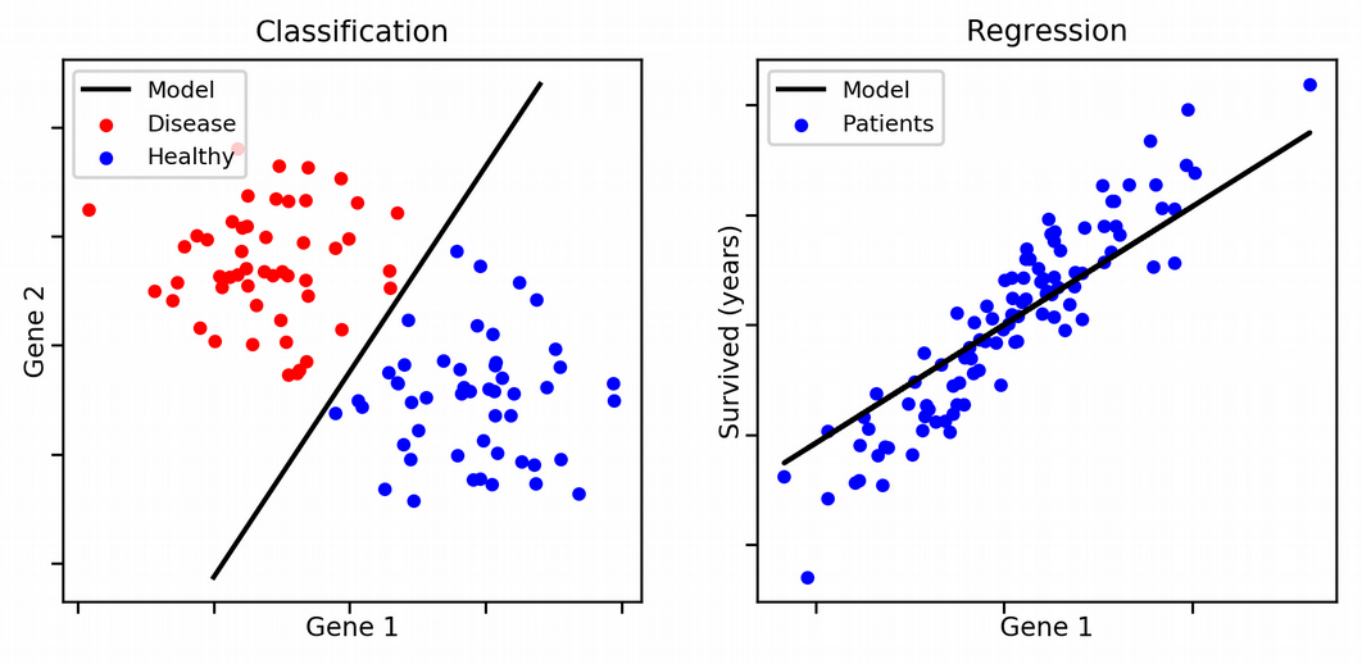
Terminologia

- **Zmienna zależna**: zmienna, której wartość/klasa znana jest tylko dla części obiektów, i której wartość/klasę chcemy modelować
- **Zmienne niezależne**: zmienne których wartości znamy dla wszystkich obiektów i które służą do modelowania zmiennej zależnej
- W przypadku, gdy celem jest zmienna dyskretna algorytm nazywamy **klasyfikatorem** (classifier) w przypadku regesji **regresorem** (regressor).
- W przypadku gdy przewidujemy etykiety dyskretnych klas mówimy o klasyfikacji, natomiast gdy przewidujemy wartość ciągłą jakiegoś zjawiska mówimy o predykcji
- Przykład:
  - kartowanie typów gleb – klasyfikacja
  - Kartowanie zawartości próchnicy w glebie - regresja

Zmienna zależna (znana)

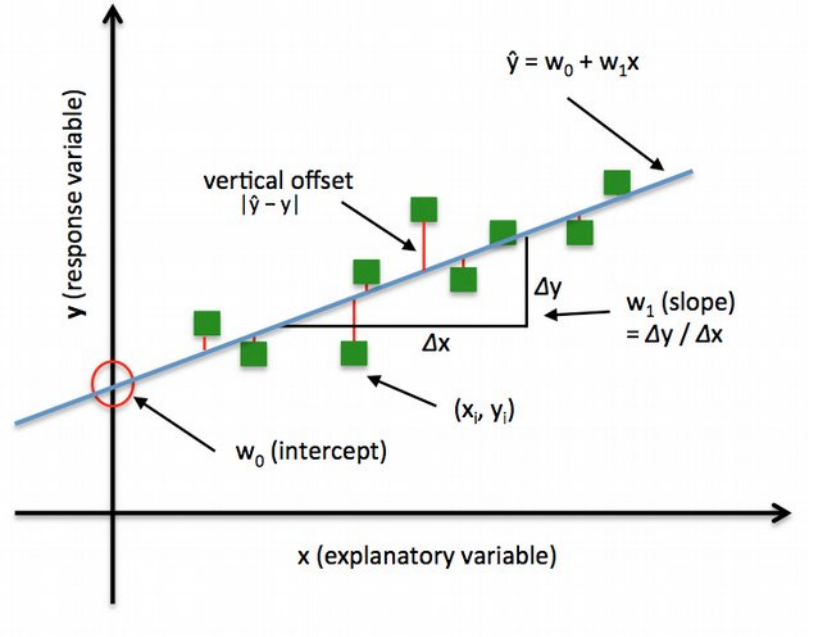
- **Kategoria** (klasa, zmienna dyskretna) – wymaga zastosowania modeli klasyfikacyjnych – przewidujących kategorię
- **Wartość** (liczba, zmienna ciągła) – wymaga zastosowania modeli regresyjnych – przewidujących wartość
- **Odsetek** – wymaga zastosowania modeli analizy przeżycia (survival) – przewidujących odsetek obiektów które przetrwają stres

Większość algorytmów jest w stanie realizować wszystkie trzy zadania



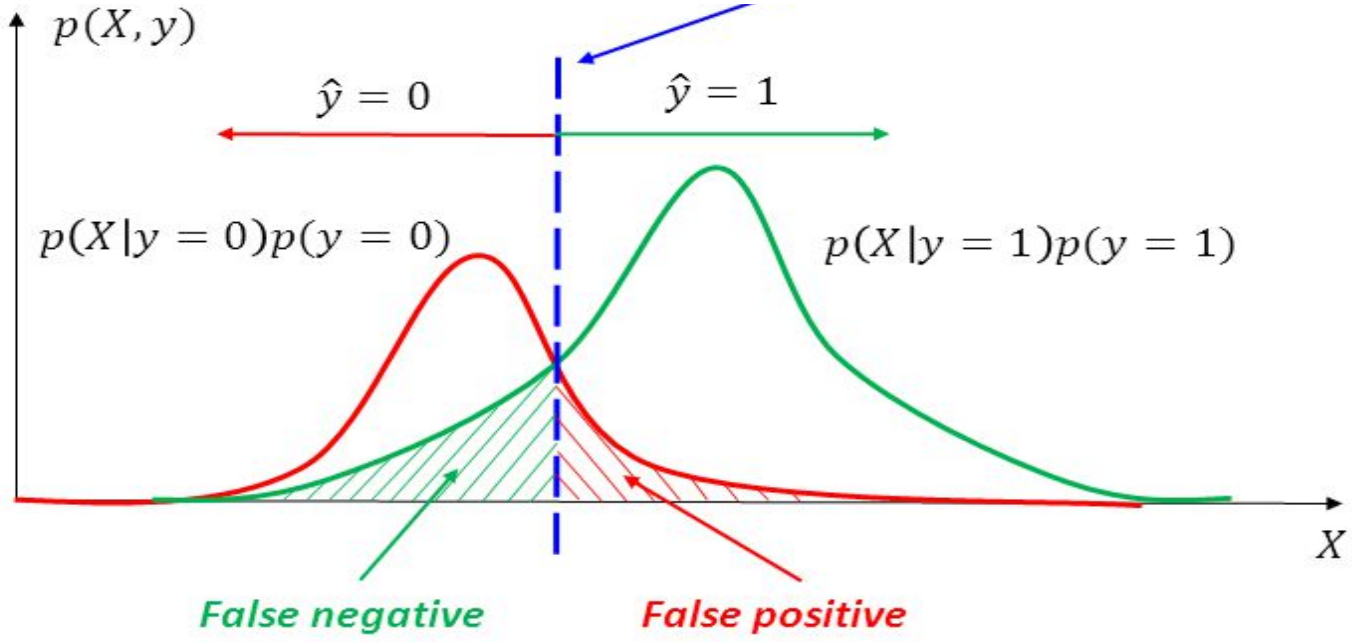
Podstawa regresji - model

- Model - funkcja (z reguły nieliniowa) przeliczająca znane wartości zmiennych wyjaśniających na zmienną modelowaną
- Model nie jest idealnie dopasowany do danych jest generalizacją
- Generalizacja powoduje że w czasie uczenia powstaje błąd dopasowania
- Algorytm uczenia minimalizuje błąd dopasowania



Podstawa klasyfikacji - Granica decyzyjna

- Granica decyzyjna pomiędzy dwoma klasami
- Granica z reguły nie jest ostra
- Powoduje to, że w trakcie procesu klasyfikacji pojawia się błąd klasyfikacji (confusion)
- Błąd powinien być minimalizowany oraz jego wartość powinna być znana



Błędy I i II typu

- Błąd I typu – błąd odrzucenia prawidłowej hipotezy 0, zwykle prowadzi do konkluzji, że istnieje relacja, której w rzeczywistości nie ma
- W uczeniu maszynowym to FALSE POSITIVE – czyli obiekty zaklasyfikowane przez klasyfikator jako pozytywne a w rzeczywistości negatywne, stwierdzenie efektu działania przy jego braku
- Błąd II typu – błąd nieodrzucenia fałszywej hipotezy 0, zwykle prowadzi do konkluzji, że nie ma relacji która w rzeczywistości jest
- W uczeniu maszynowym to FALSE NEGATIVE – czyli obiekty odrzucone przez klasyfikator jako pozytywne (i zakwalifikowane jako negatywne) a w rzeczywistości pozytywne, stwierdzenie braku efektu działania przy jego wystąpieniu

Macierz zmieszania
(confusion matrix)

n=165	**Predicted: NO**	**Predicted: YES**	
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	

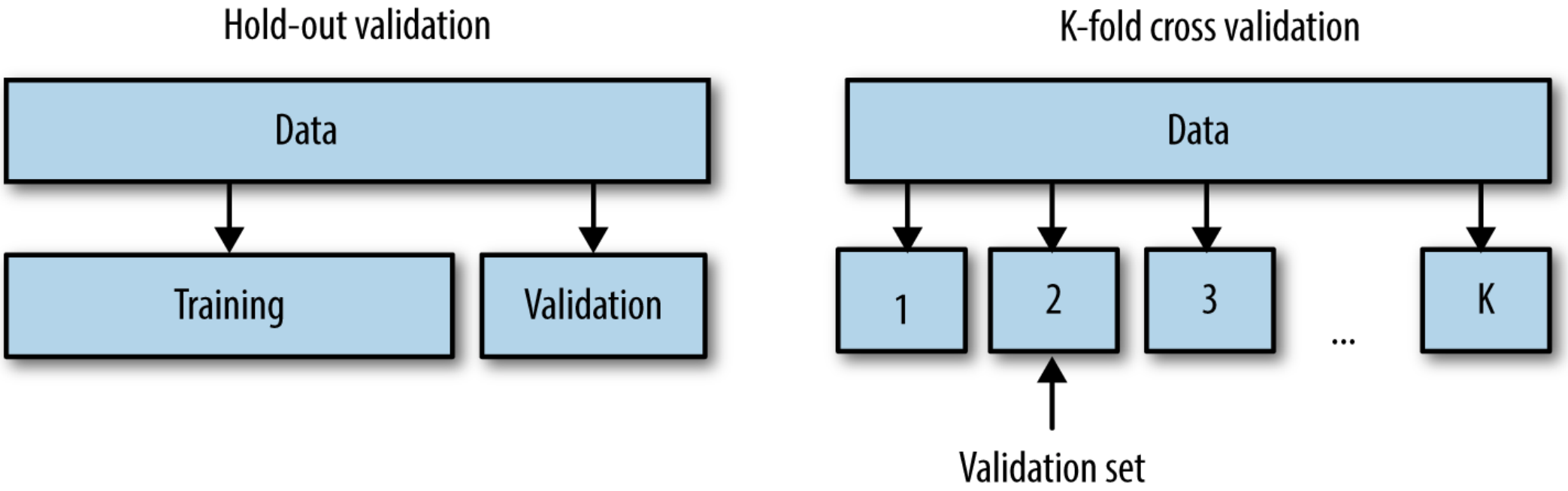
Błąd I typu:
Szukasz nie tam gdzie trzeba
(i od razu o tym wiesz)

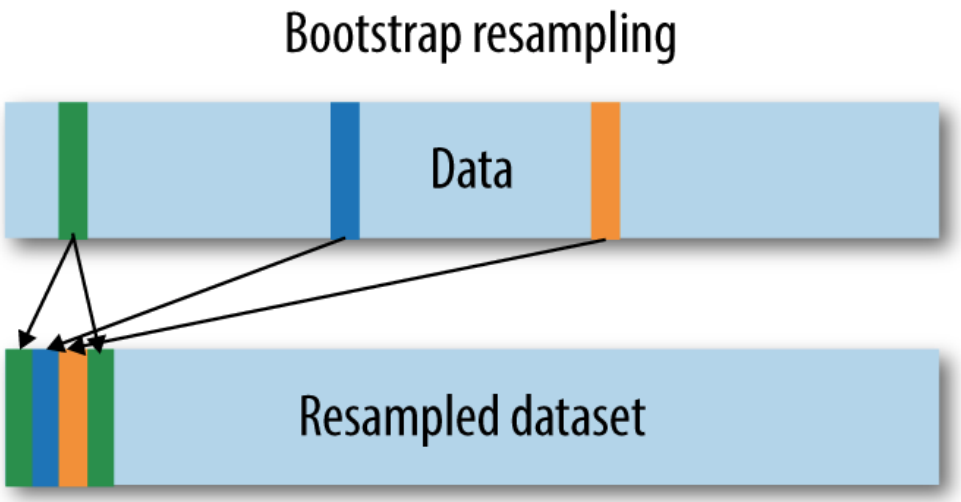
Błąd II typu:
Pominąłeś coś bardzo ważnego
(i raczej się o tym nie dowiesz)

Jak szacujemy błąd

- Poprzez sprawdzenie skuteczności uczenia na niezależnym zbiorze zwanym zbiorem testowym.
- Zbiór testowy otrzymujemy porzez wydzielenie z całości danych zawierających zmienną zależą na część treningową i testową
- Podstawowe strategie:
  - **Podział na dwa zbiory**: treningowy i testowy (holdout)
  - **Ocena krzyżowa** (cross-validation) wielokrotny: systematyczny podział na zbiór treningowy i testowy bez zwaracania
    - Leave-one-out
    - Leave-group-out (k-fold)
  - **Bootstrap**: wielokrotny losowy podział na zbiór treningowy i testowy ze zwracaniem
  - **Podział ręczny**: (holdout) najczęściej jeżeli istnieją dodatkowe kryteria podziału – np. testowanie różnic pomiędzy grupami w populacji

Różne metody próbkowania





Proces uczenia

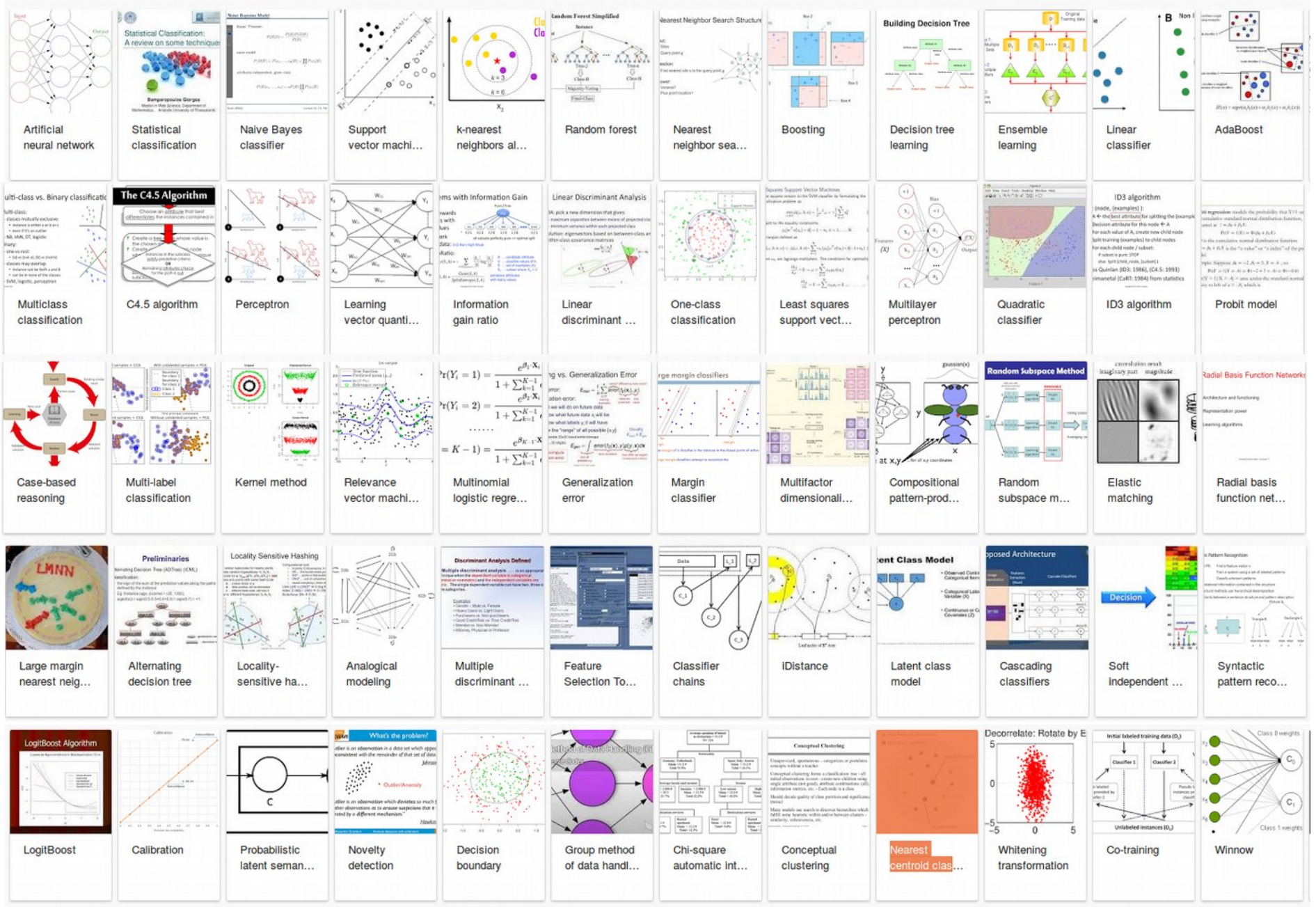
- Proces uczenia ma na celu minimalizowanie błędu klasyfikacji/regresji dla dostępnego zbioru danych uczących (takich dla których znana jest zmienna zależna)
- Proces ten obejmuje:
  - Wybór metody przepróbkowania (resampling)
- Wybór klasyfikatora/regresora
- Dostrajanie (tuning) parametrów modelu
- Ocena wydajności modelu
- Akceptacja lub odrzucenie modelu

Sama **PREDYKCJA** - zastosowanie modelu dla nowych danych

– nie jest częścią budowania modelu

Proces uczenia - schemat

60 najpopularniejszych klasyfikatorów

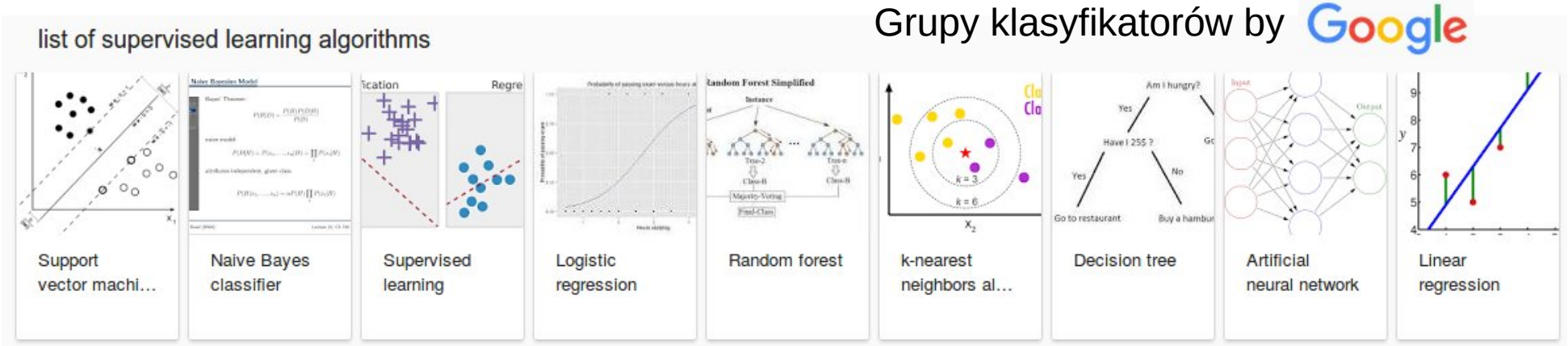


Wybrane klasyfikatory: osobny wykład

Grupy klasyfikatorów i regresorów

Grupa	Przykłady	Cechy
Liniowe i regresja	Naive Bayes, Linear Regression, Logistic Regression, MARS	Szukają dopasowania liniowego do modelu, zakładają normalny rozkład zmiennych wyjaśniających
Najbliższego sąsiada	kNN	Klasyfikacja nieparametryczna na podstawie charakterystyki sąsiedztwa
Dyskryminacyjne	LDA, QDA, PLS	Szukają kierunku największych różnic dla poszczególnych klas
Maszyny wektorów wsparcia	SVM	Konstruowanie hiperpłaszczyzn rozdzielających klasyfikowane grupy
Decyzyjne i regułowe	CART, rule systems	Niemetryczne klasyfikatory oparte o reguły podziałów w postaci struktur drzewiastych lub zwykłe zestawy reguł
Wzmacniane i łączone	RandomForest, boosted Trees, AdaBoost	Wykorzystują liczbe słabe klasyfikatory do tworzenia jednego silnego klasyfikatora
Sieci neuronowe	Neural network, deep learning	Warstwy neuronów i dobór wag w celu aktywacji

Grupy klasyfikatorów by Google

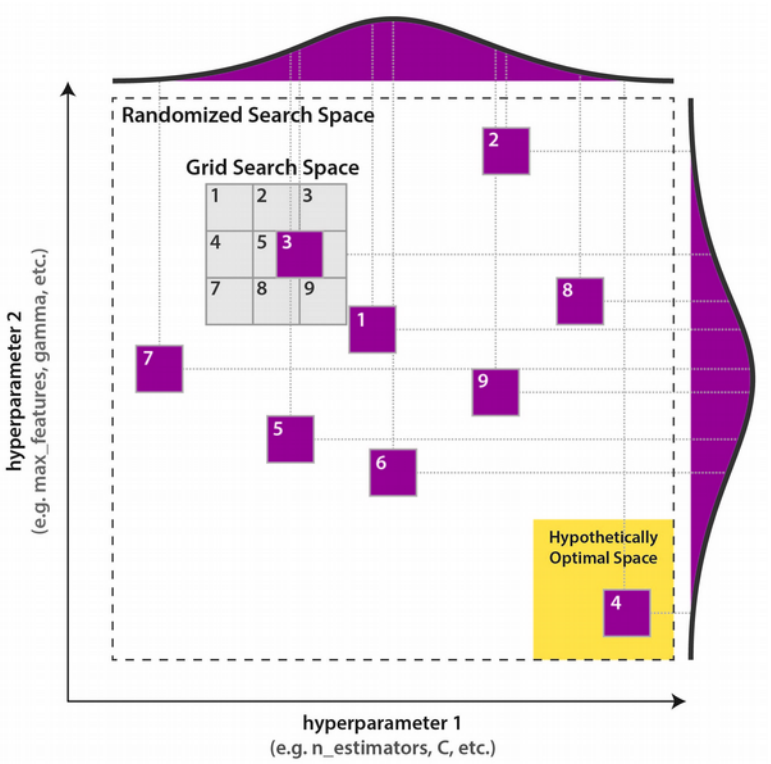


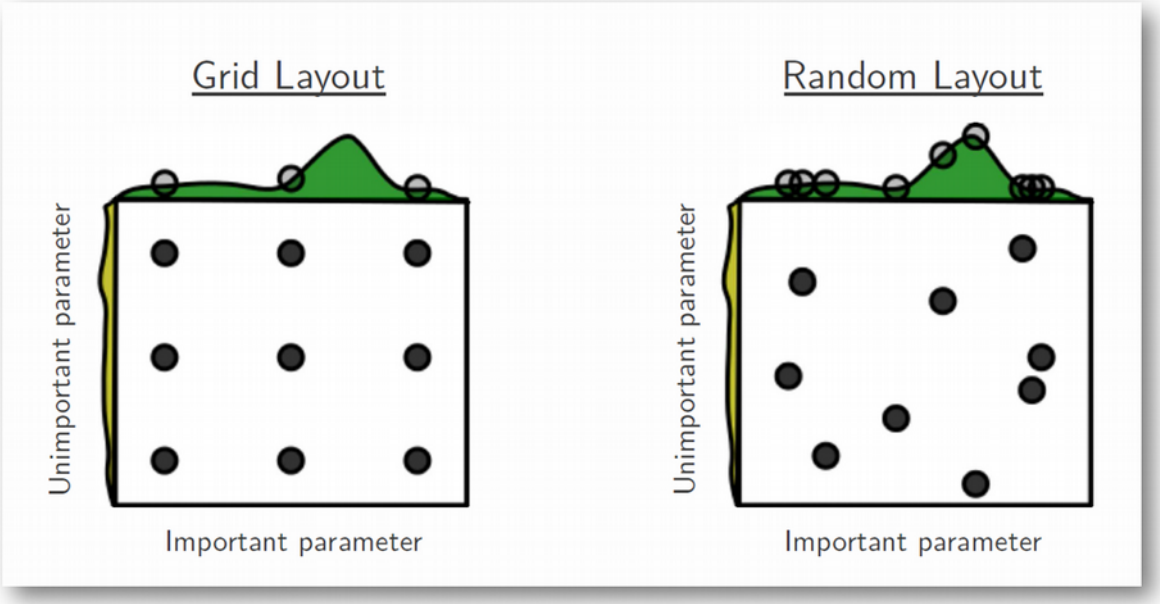
Optymalizacja parametrów modelu

- Większość modeli uczących posiada od jednego do kilku parametrów swobodnych (free parameters), które wpływają bezpośrednio na wynik klasyfikacji
- Nie ma ścisłych reguł doboru parametrów modelu, zestaw najlepszych parametrów zależy od danych
- Proces doboru parametrów odbywa się najczęściej metodą przeszukiwania zestawów parametrów w celu doboru optymalnego zestawu (tuning)
- Jest to proces kosztowny i długotrwały

Szukanie w siatce vs. losowe

- Szukanie odbywa się w regularnej siatce, losowo, lub w siatkach o zmiennej gęstości (zagęszczenia przy optimum)





Funkcja kosztu (straty)

- Loss Function/Cost Function
- Pewna funkcja która reprezentuje „koszt” intuicyjnie wiązany z zagadnieniem
- W statystyce i uczeniu maszynowym to błąd klasyfikatora/regresora:
  - Dla przypadków regresji – miara wyrażająca różnicę pomiędzy wartościami znanymi a estymowanymi
  - Dla przypadków klasyfikacji – miara wyrażająca różnicę w wartościach prowadzących do zakwalifikowania obiektów do różnych klas (niekoniecznie ilość)

Wybrane przykłady funkcji kosztu

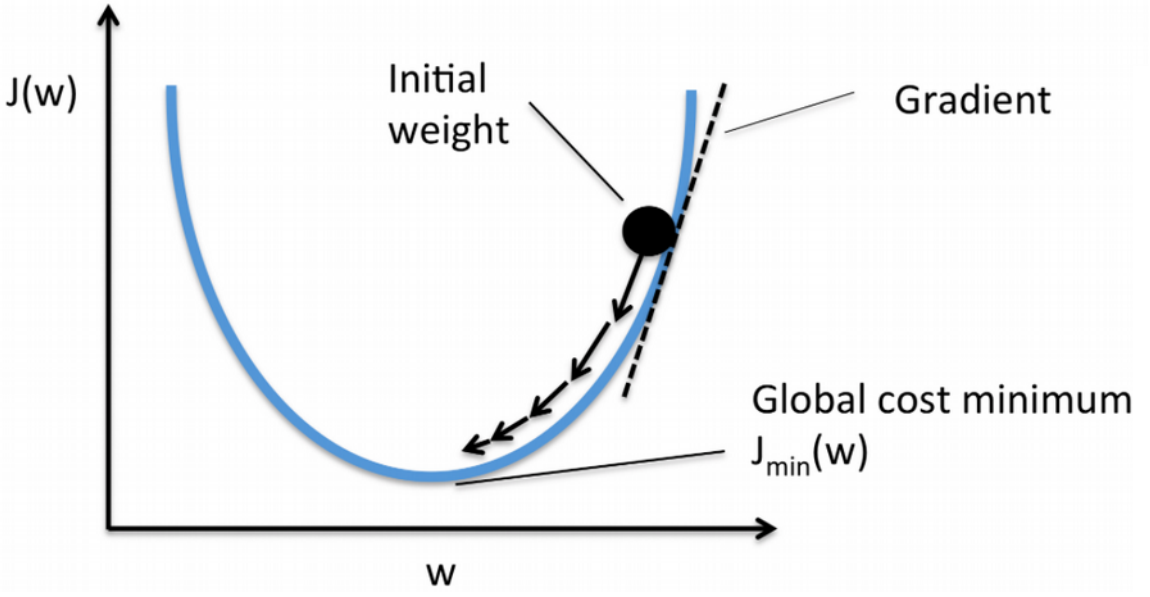
- **Cross-entropia** – miara klasyfikacji pokazująca jak bardzo prawdopodobieństwo zakwalifikowania się do danej klasy odbiega od danej klasy: np. jeżeli estymowane prawdopodobieństwo przynależności do klasy X wynosi jest 0.1 to różnica wynosi 0.9 (duże). Cross – entropia to suma tych różnic dla wszystkich obiektów
- **Funkcja zawieszone (hinge)** – miara klasyfikacji pokazująca stopień niezaklasyfikowania
  L(y) = max(0, 1-t*y), jeżeli y > 1 (jednoznaczne zaklasyfikowanie do danej klasy) to wartość kosztu wynosi 0, w przypadku niejednoznaczności pojawia się koszt
- **MAE - Średni błąd bezwzględny (L1)** – miara regresji, suma wszystkich błędów (natężenie błędu) bez uwzględnienia ich zwrotu
- **MSE i RMSE - średni błąd kwadratowy (L2)** – miara regresji, różnica do MAE polega na podniesieniu do kwadratu błędu przed sumowaniem, co oznacza że duże błędy będą mocniej wpływały na łączną miarę. Przy równomiernie rozłożonym błędzie nie ma różnic pomiędzy MAE i MSE

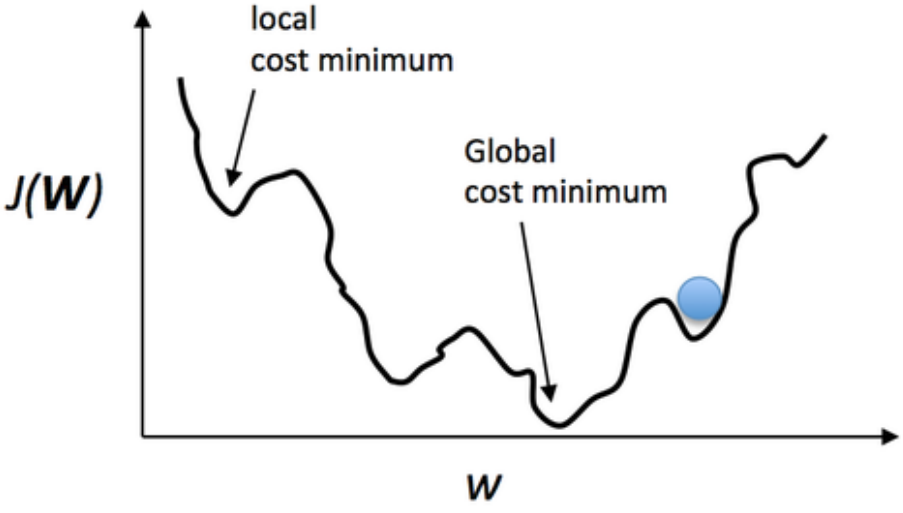
$$\mathrm{MAE} = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

Szukanie globalnego minimum

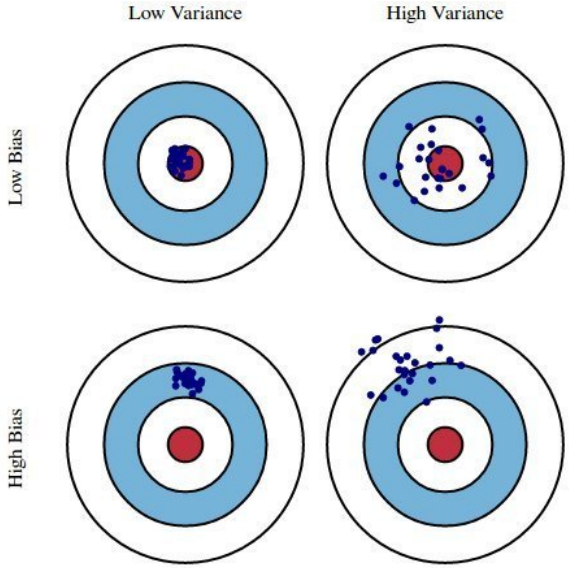
- Proces przeszukiwania trwa aż do znalezienia optymalnego zestawu (globalnego optimum)
- Bardzo często proces szukania kończy się w optimum lokalnym
- Nie zawsze znalezienie optimum jest kosztowo opłacalne, wiele klasyfikacji funkcjonuje przy tzw rozwiązaniach suboptymalnych (bliskich optymalnemu)

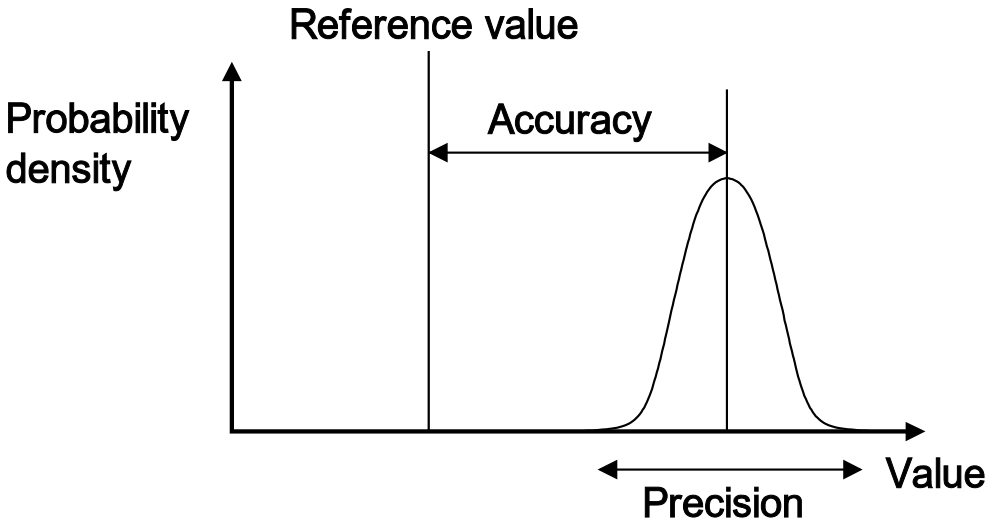




Precyzja i dokładność

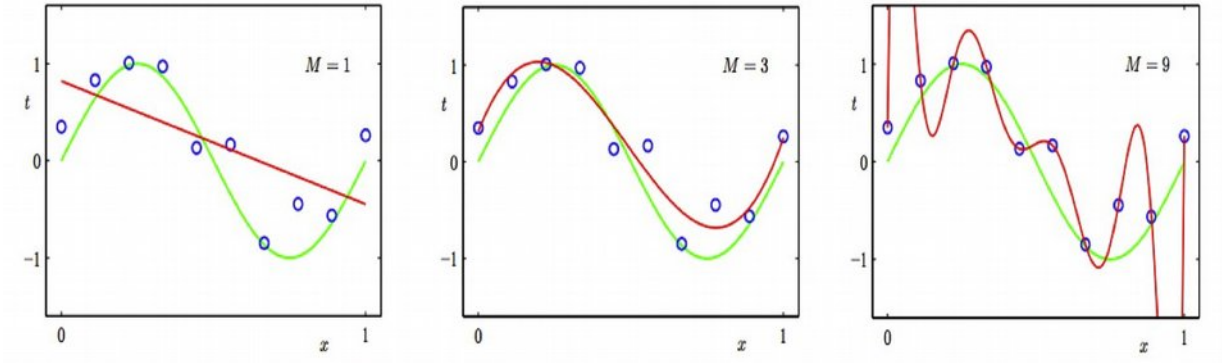
- **Dokładność** – odległość od rzeczywistej wartości. Wysoka dokładność oznacza że model pracuje wydajnie, ale jest to niedokładna miara w przypadku zbiorów niezbilansowanych (gdy istotna klasa jest w mniejszości), wartość dokładności może być zawyżana przez wysoką wartość TRUE NEGATIVES. Miarą (negatywną) dokładności jest błąd systematyczny (odstawanie, **bias**)
- **Precyzja** – stopień powtarzalności w tych samych warunkach. Wysoka wartość wskazuje, że uzyskujemy mało FALSE POSITIVES (ale może to być kosztem wysokiej wartości odcięcia). Miarą (negatywną) precyzji jest błąd losowy (wariancja)
- Zbilansowana dokładność (SENS + SPEC) /2



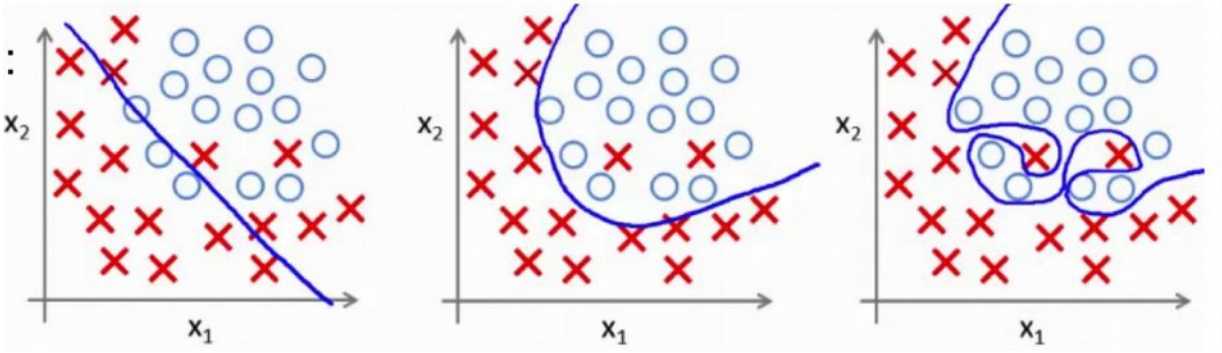


Błąd systematyczny i losowy

- Błąd systematyczny to efekt błędnych założeń w modelu (model często zbyt prosty, underfitting)
- Błąd losowy to efekt niewielkich fluktuacji modelu wynikający z nadmiernego dopasowania modelu do danych treningowych (model zbyt skomplikowany – overfitting, przeuczenie)

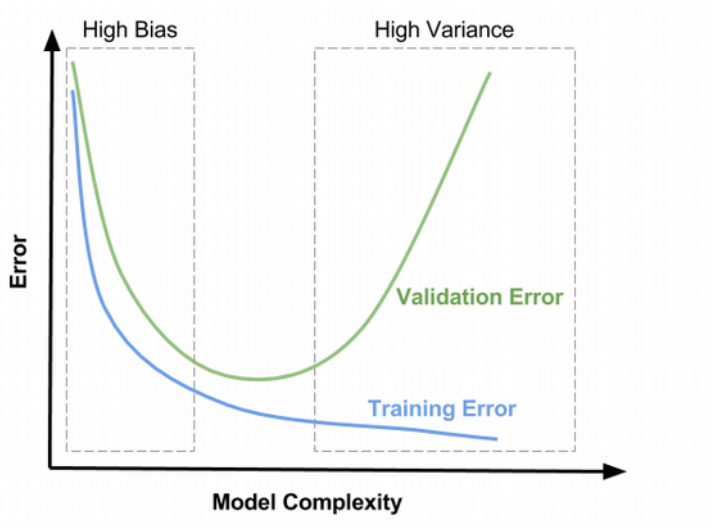


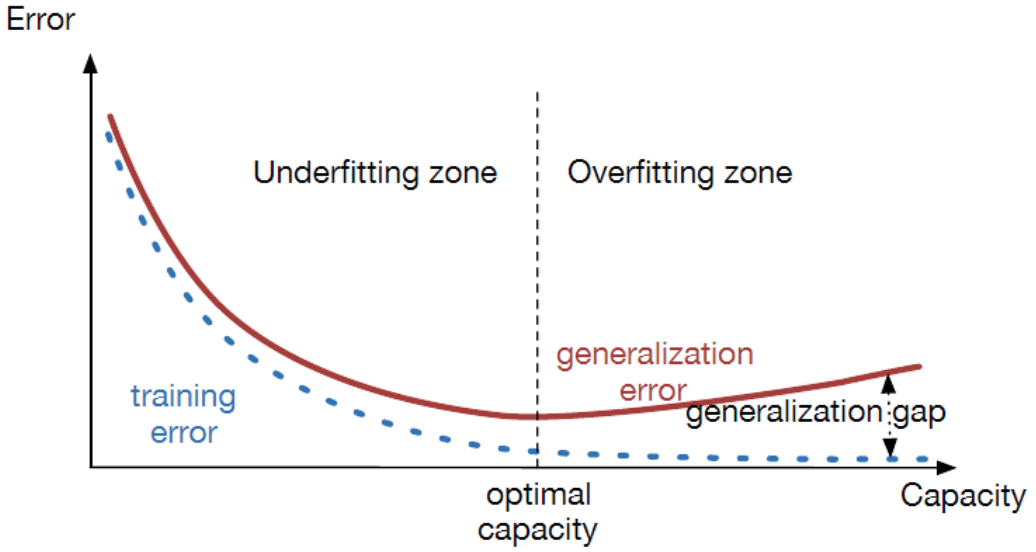
underfitting optimal overfitting



Optymalizacja vs. generalizacja

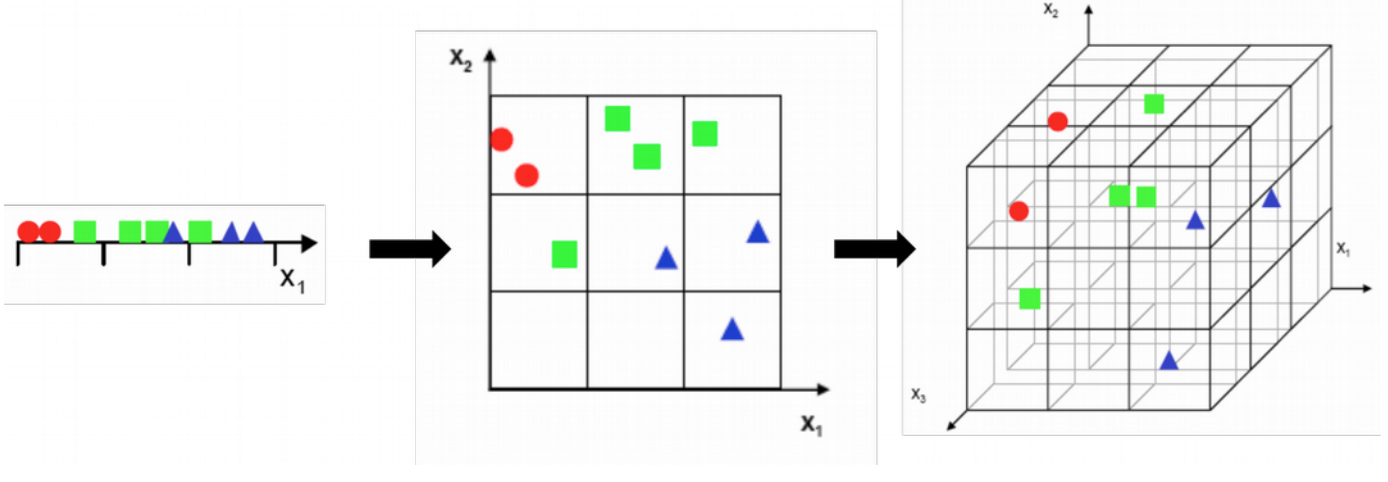
- Przy tworzeniu modelu ważny jest bilans pomiędzy niedopasowaniem a nadmiernym dopasowaniem
- Nadmierne dopasowanie skutkuje wzrostem wariancji w wyniku nadmiernego dopasowania do danych uczących. Model nie jest uniwersalny i skutkuje pogorszeniem jakości predykcji na nowych danych





Przeuczenie a liczba zmiennych

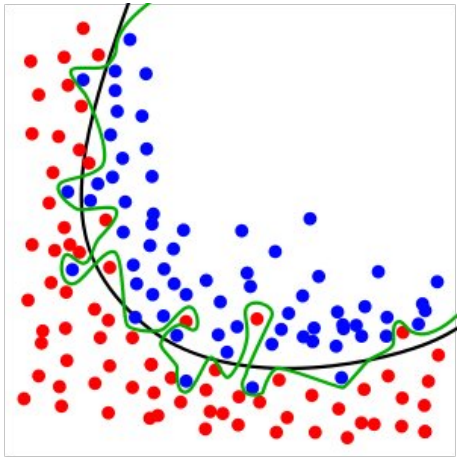
- W procesie klasyfikacji wydaje się że zwiększanie ilości zmiennych wyjaśniających prowadzi do lepszego modelu. Jest to prawdziwe dla małej liczby zmiennych
- Im więcej zmiennych (wymiarów) tym więcej przy tej samej liczbie przypadków pustek powstaje w wielowymiarowej przestrzeni

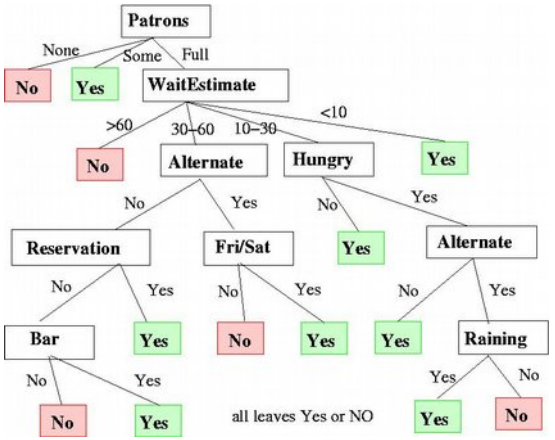


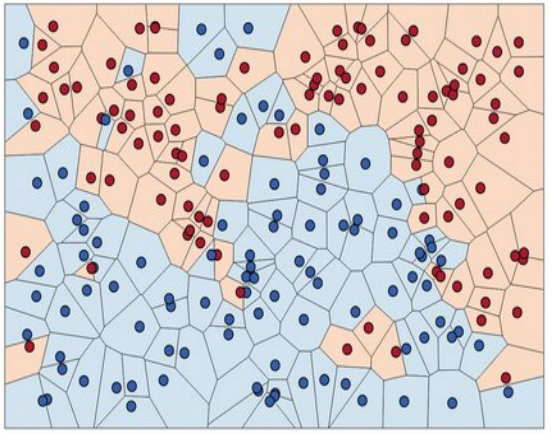
Przykłady przyczyn przeuczenia

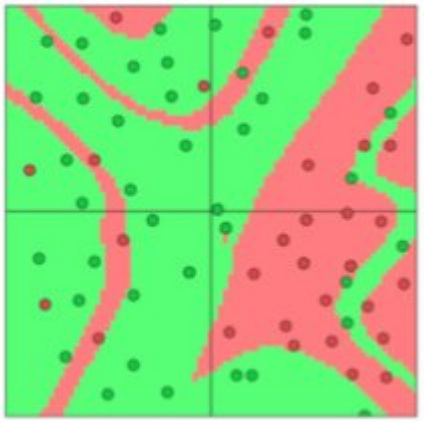
- Zbyt szczegółowa granica decyzyjna (SVM)
- Nadmiernie rozbudowane drzewo decyzyjne (CART)
- Zbyt mały promień przeszukiwania lub ilość sąsiadów w kNN
- Zbyt dużo warstw ukrytych w sieciach neuronowych

(Więcej kolejny wykład)



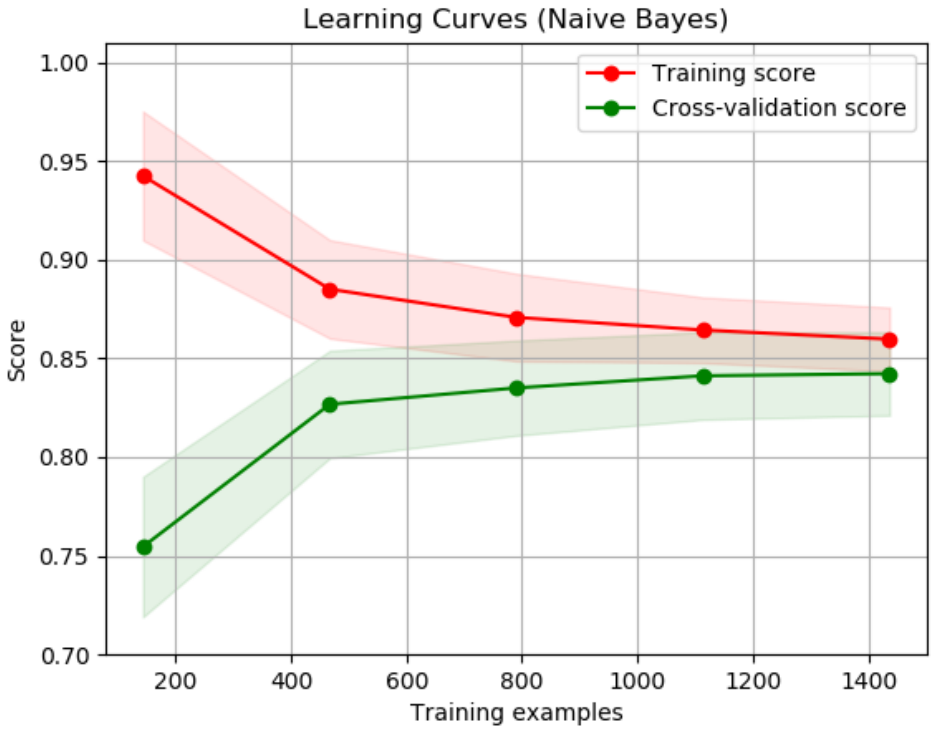






Wielkość zbioru uczącego

- Wielkość zbioru uczącego tylko do pewnego stopnia ma wpływ na błąd modelu. Powyżej pewnej ilości prób nie ma to już większego znaczenia

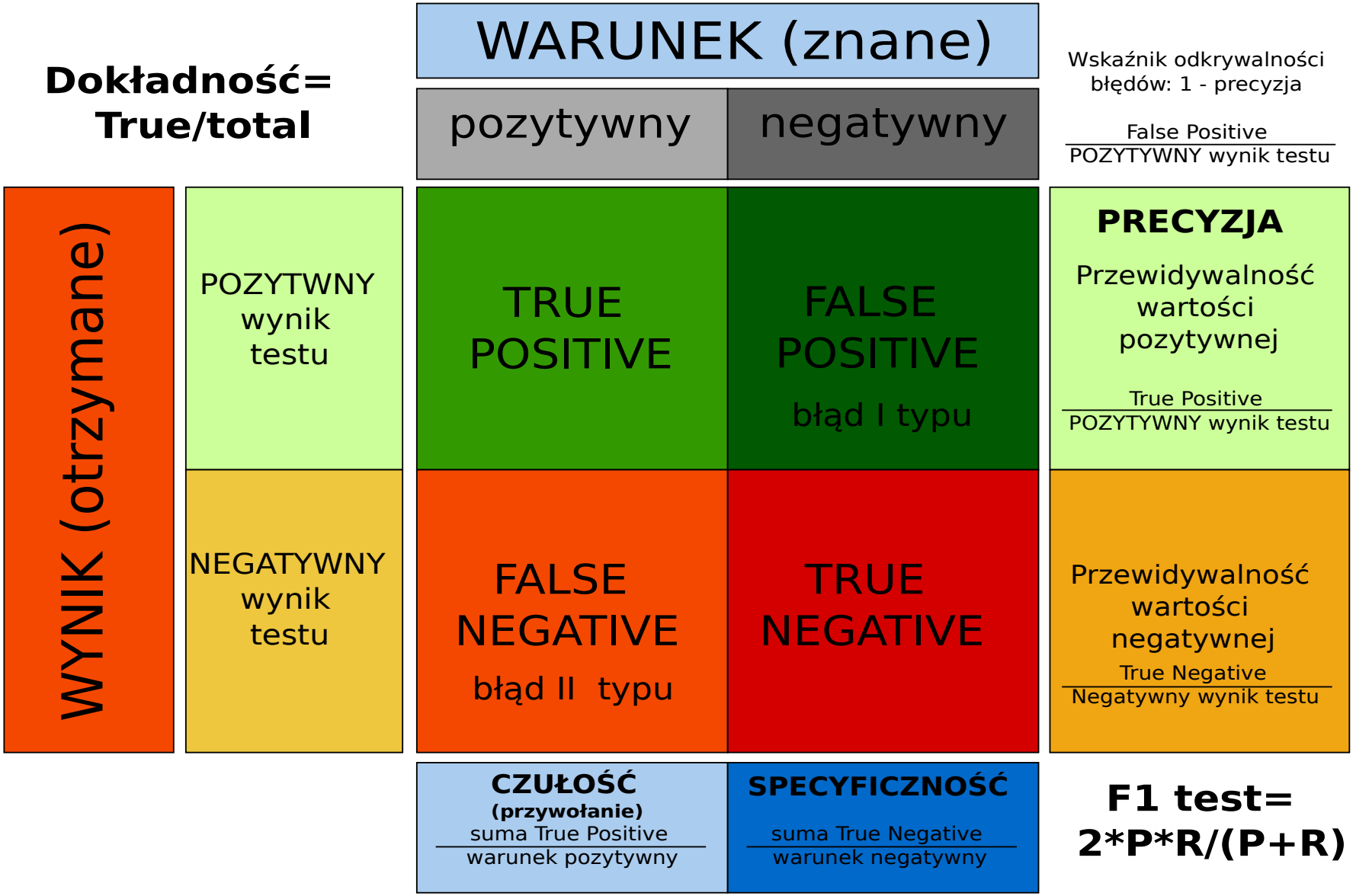


Zrozumienie modelu

- Modele ze zbyt dużą ilością zmiennych wyjaśniających są trudne do zrozumienia
- Modele oparte o małą liczbą zmiennych są prostsze do optymalizacji. Ma to szczególne znaczenie jeżeli celem uczenia maszynowego nie jest wysoki wskaźnik predykcji, ale zrozumienie zależności pomiędzy zmiennymi wyjaśniającymi a wyjaśnianym procesem

Ocena wydajności modeli

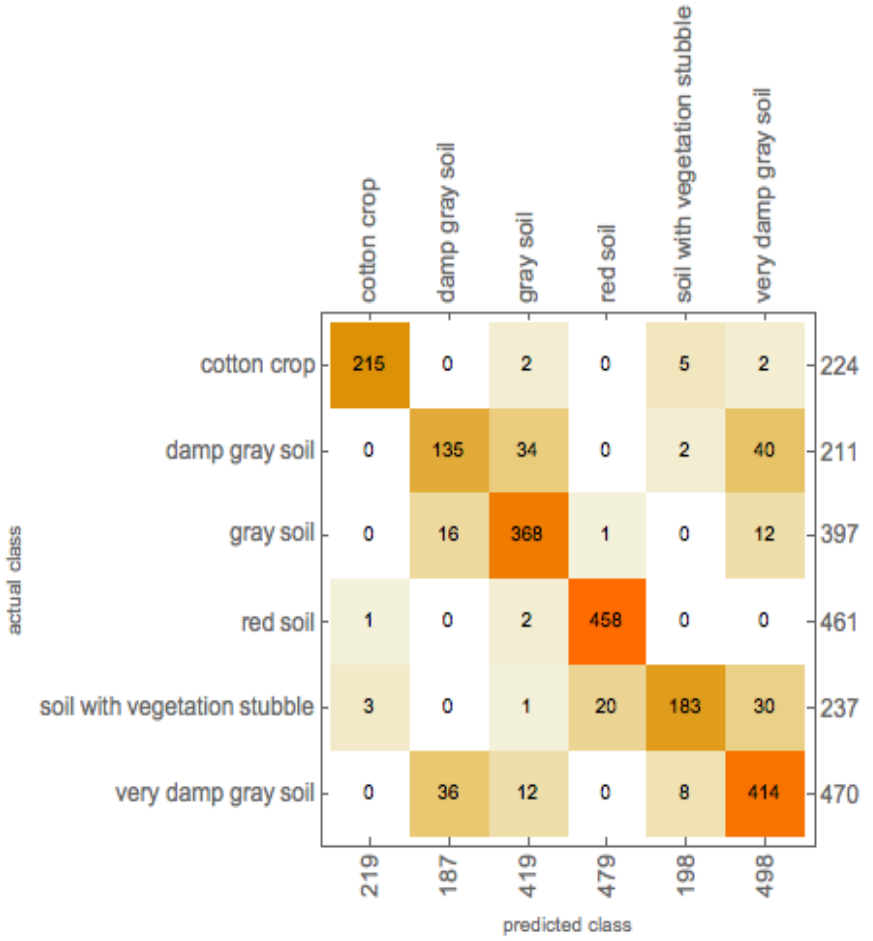
Ocena i porównanie klasyfikatorów



Macierze zmieszania

- Dla dwóch klas i dla wielu klas

		WARUNEK (znane)	
		Pozytywny ***chory***	Negatywny ***zdrowy***
WYNIK (otrzymane)	Pozytywny wynik testu ***chory***	TRUE POSITIVE **stwierdzono chorobę u chorej osoby**	FALSE POSITIVE **stwierdzono chorobę u zdrowej osoby**
	Negatywny wynik testu ***zdrowy***	FALSE NEGATIVE **nie stwierdzono choroby u chorej osoby**	TRUE NEGATIVE **nie stwierdzono choroby u zdrowej osoby**



POSITIVE i NEGATIVE

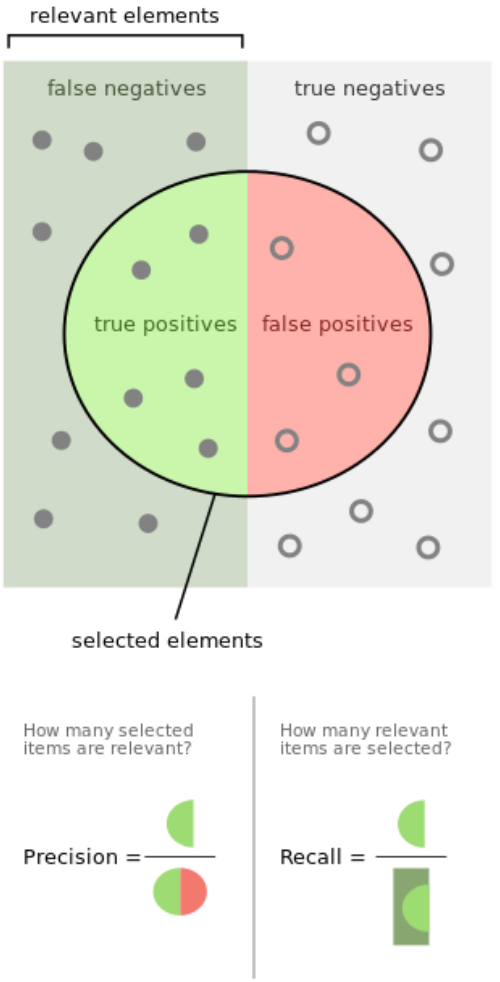
- Obiekty prawidłowo zakwalifikowane jako pozytywne (x należy do C, działanie wywołuje efekt) określa się jako TRUE POSITIVE
- Obiekty prawidłowo zakwalifikowane jako negatywne (x nie należy do C, działanie nie wywołuje efektu) określa się jako TRUE NEGATIVE
- Pojęcia FAŁSZYWIE POZYTYWNY i FAŁSZYWIE NEGATYWNY odnoszą się jedynie do sytuacji binarnych – spełnia założenie nie spełnia założenia, nie odnoszą się do klasyfikacji wieloklasowych, gdzie pomyłki mogą częściej występować pomiędzy niektórymi klasami

Koszt błędu nie zawsze jest jednakowy

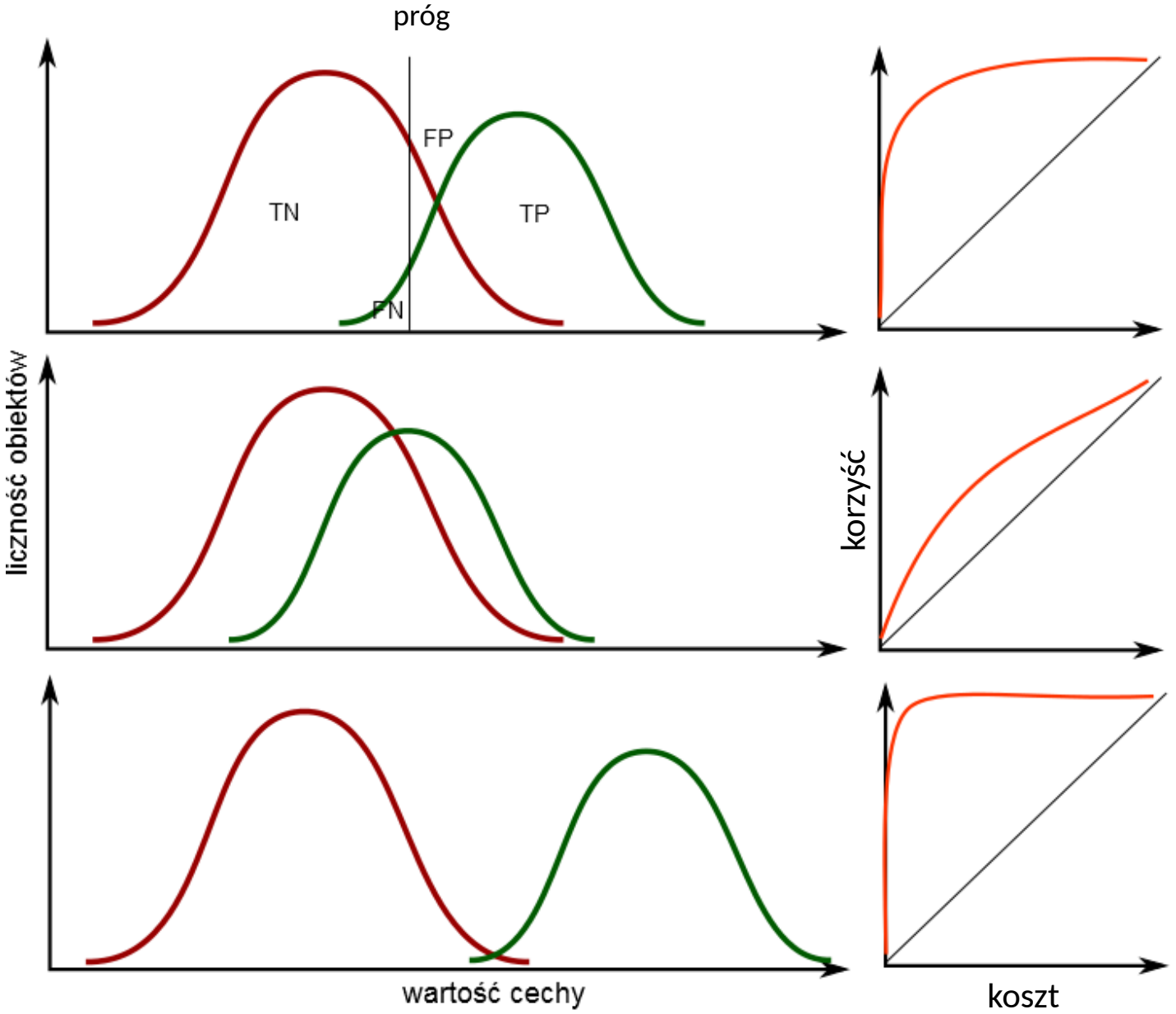
- Badania są tanie (na przykład badania powierzchniowe) – lepiej zbadać wszystko niż coś stracić (minimalizujemy błąd II typu)
- Badania są drogie (na przykład głęboki odwiert) – wybierzmy miejsce gdzie na pewno znajdziemy to co szukamy (minimalizujemy błąd I typu)
- Systemy informatyczne z reguły starają się minimalizować **błąd całkowity** bilansując oba błędy równocześnie

Skuteczność predykcji klasy znaczącej - Precyzja i przywołanie

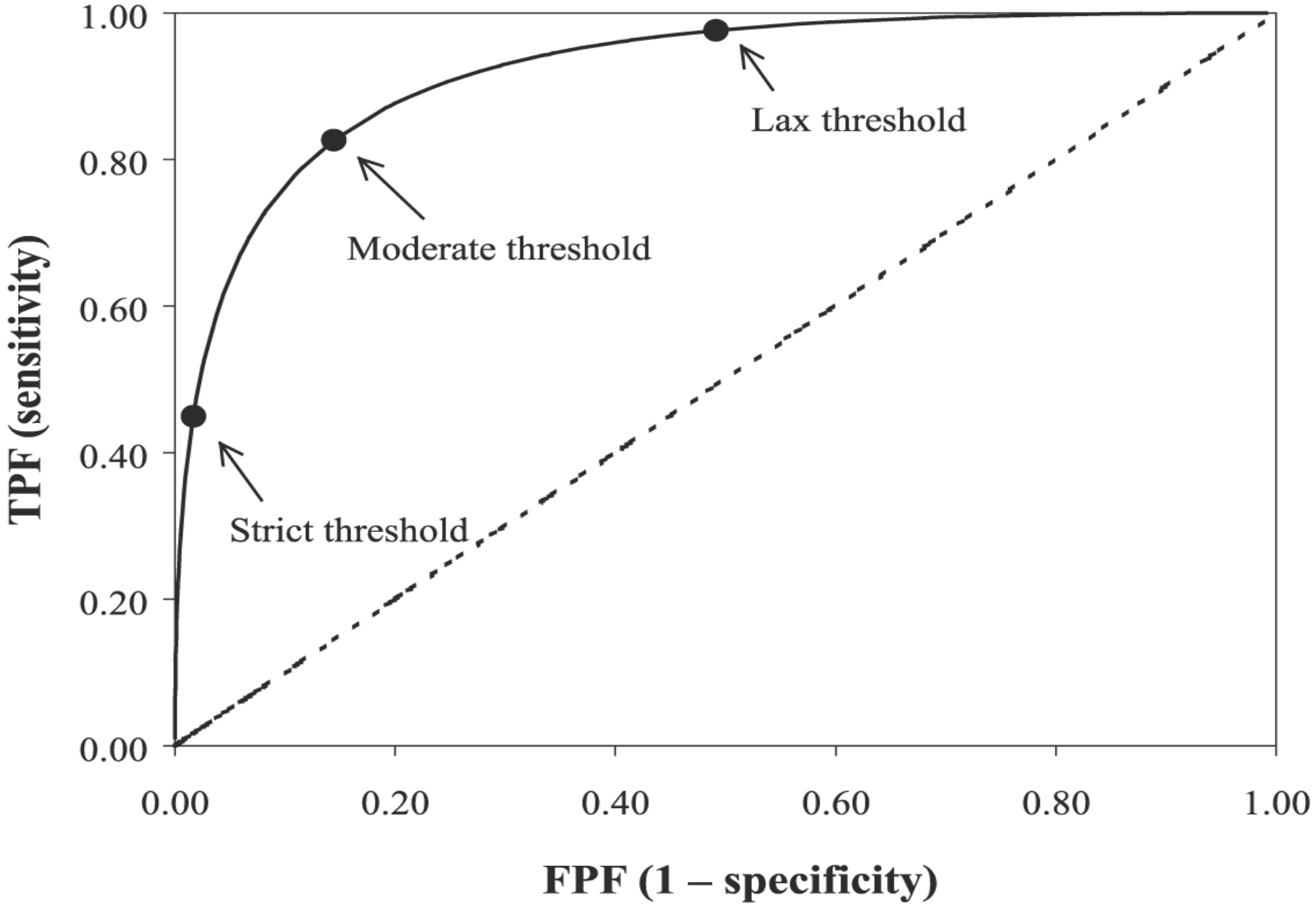
- **Precyzja** – ilość poprawnie sklasyfikowanych obiektów TRUE względem wszystkich sklasyfikowanych jako TRUE – oznacza jak czysta jest grupa obiektów znaczących. Maksymalizacja – precyzji – minimalizacja błędu I typu
- **Przywoływanie** – jak dużo obiektów TRUE udało się poprawnie sklasyfikować – niezależnie od błędnie sklasyfikowanych obiektów FALSE – maksymalizacja przywołania – minimalizacja błędu II typu



Zależność korzyść - koszt

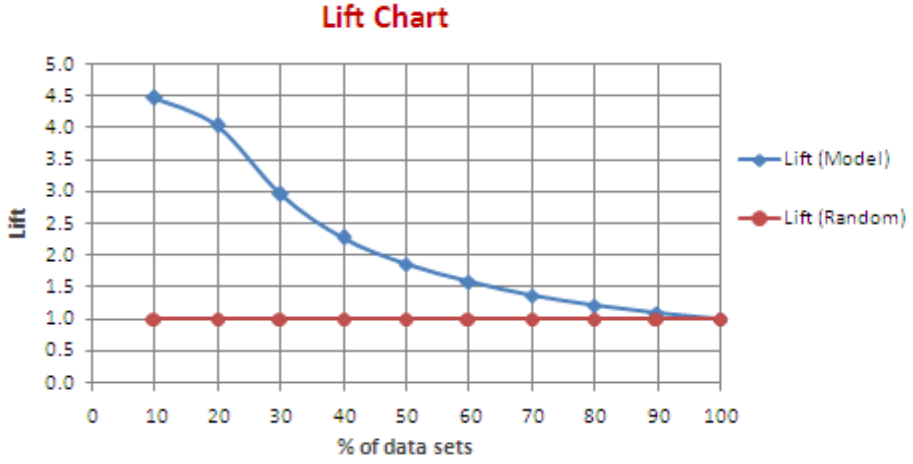


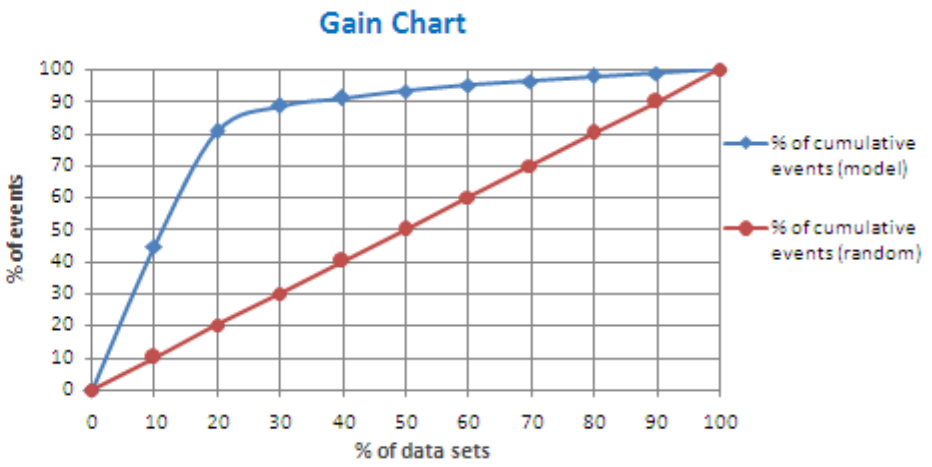
Krzywa ROC


1.00
0.80
0.60
0.40
0.20
0.00
TPF (sensitivity)
Lax threshold
Moderate threshold
Strict threshold
0.00
0.20
0.40
0.60
0.80
1.00
FPF (1 – specificity)

Diagramy wyniesienia (lift)

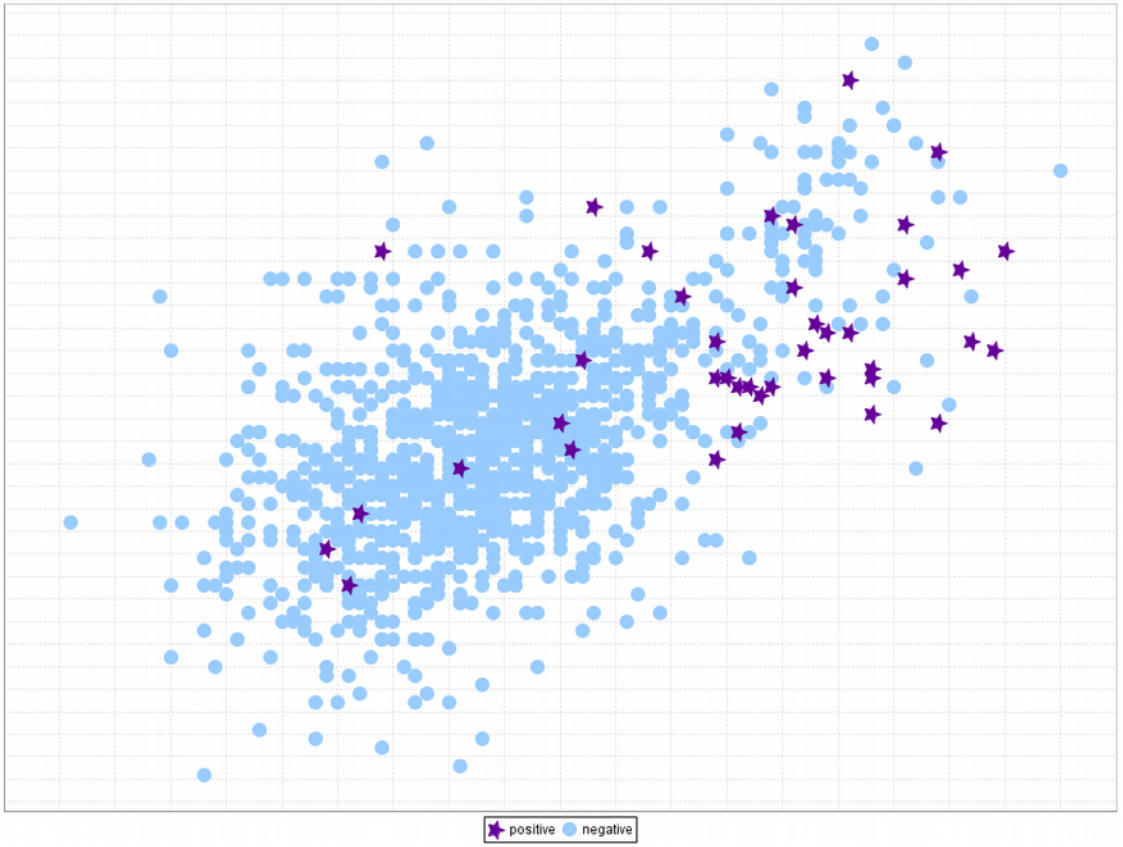
- Diagram lift jest miarą wydajności modelu, wskazującym w jakim stopniu model (1) wzmacnia prawdopodobieństwo sukcesu w porównaniu z wyborem losowym (2). Lift to proporcja 1/2
- Diagram zysku wskazuje jak dużą część populacji musi zostać przeanalizowana aby uzyskać zadowalającą wydajność
- Lift to inna forma wizualizacji krzywej ROC





Zbiory niezbilansowane

- W przypadku gdy jedna z klas (z reguły bardziej znacząca) jest w mniejszości (10x i więcej) w stosunku do klasy większościowej mówimy o zbiorach silnie niezbilansowanych

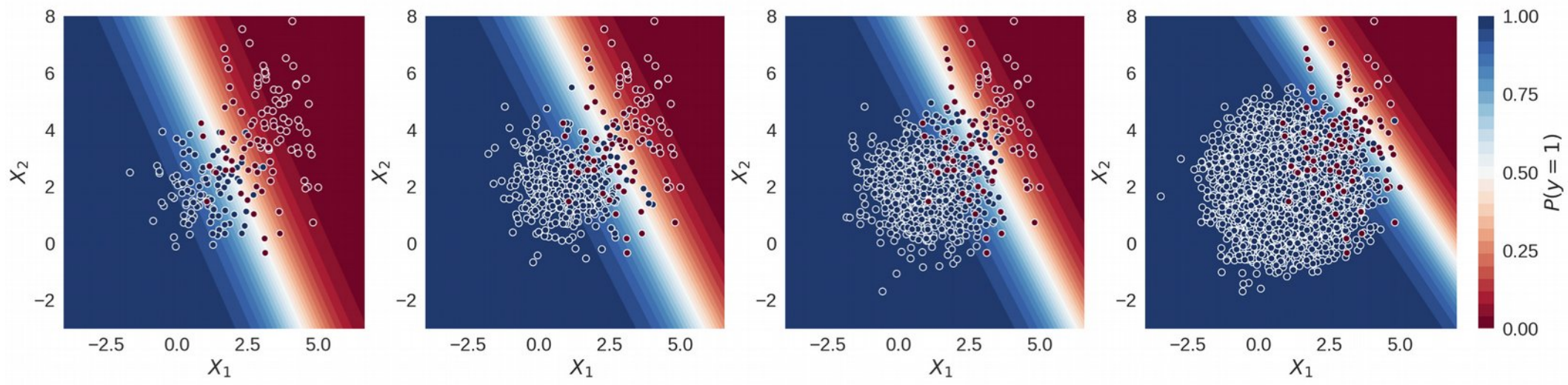


Klasyfikacja zbiorów niezblilansowanych

- Trudności z wyznaczeniem granicy decyzyjnej ze względu na minimalizację zbilansowanego błędu – jak najmniej błędnie klasyfikowanych obiektów niezależnie czy zaliczają się do klasy znaczącej czy nie. W efekcie mamy bardzo duży błąd I typu, przy małym błędzie II typu i małym błędzie całkowitym
- np. 1000 FALSE i 30 TRUE klasyfikator zakwalifikował wszystko jako FALSE i całkowity błąd wynosi mniej niż 3% (!!!). Niestety klasyfikator nie jest w stanie zakwalifikować żadnego obiektu jako TRUE
- Większość zbiorów to zbiory niezbilansowane (oszustwa, zachorowania na rzadkie choroby, itp.)

Uczenie z nadpróbkowaniem

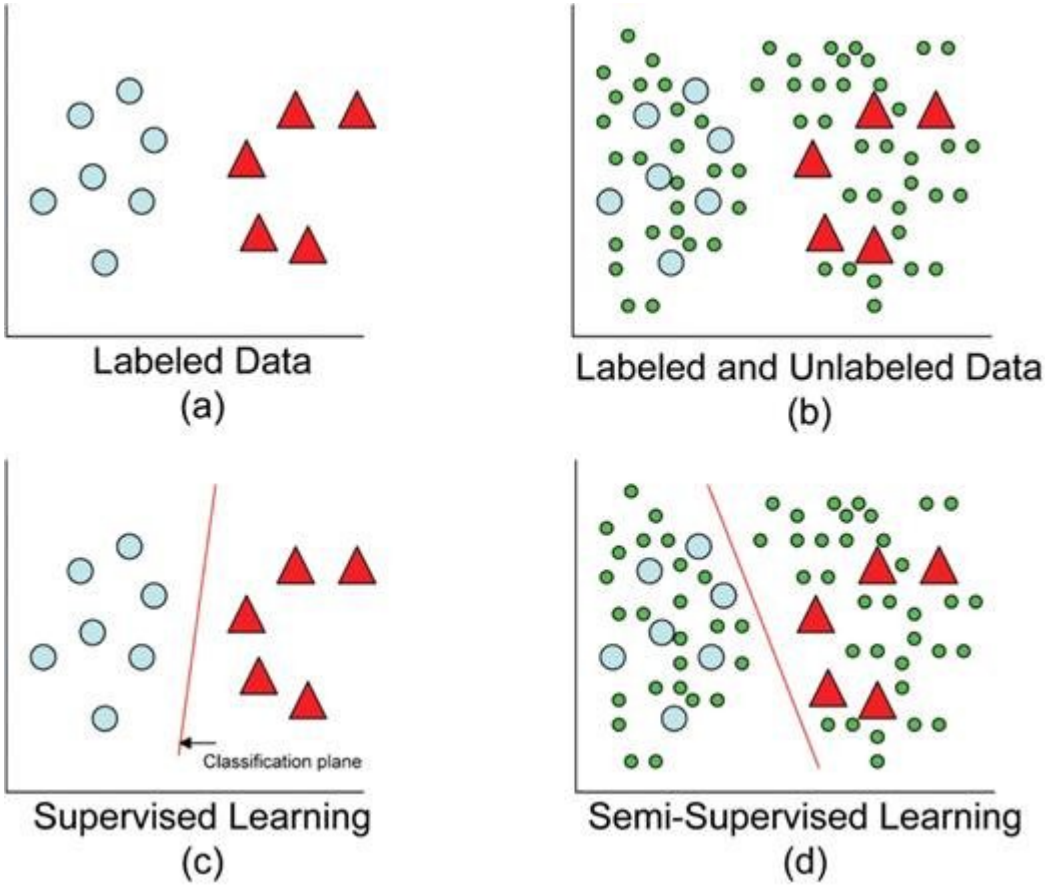
- Nadpróbkowanie klasy mniejszościowej – minimalizuje błąd I typu ale prowadzi do zwiększenia błędu II typu i błędu całkowitego)
- Niemniej jednak pozwala wykrywać rzadkie przypadki (lepiej skierować na badania 100 osób w których 90 okaże się zdrowych niż pozwolić 10 osobom zachorować na raka)

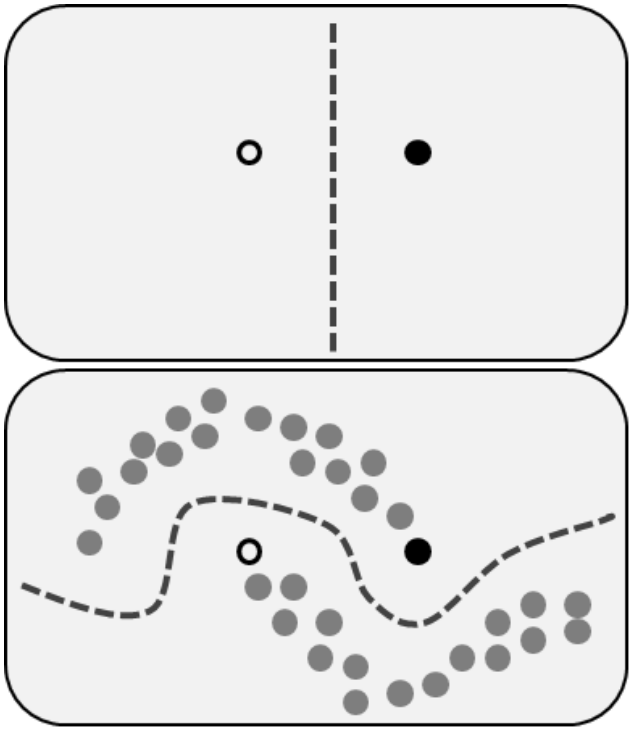


Metody częściowo nadzorowane

- Ang. Semi-supervised
- Metody pomiędzy uczeniem nadzorowanym (wszystkie dane uczące mają etykiety) a nienadzorowanym (brak etykiet)
- Zakładają że jedynie niewielka część zbioru treningowego ma etykiety, przynależność pozostałych jest nieznana
- Nie nadają się do regresji

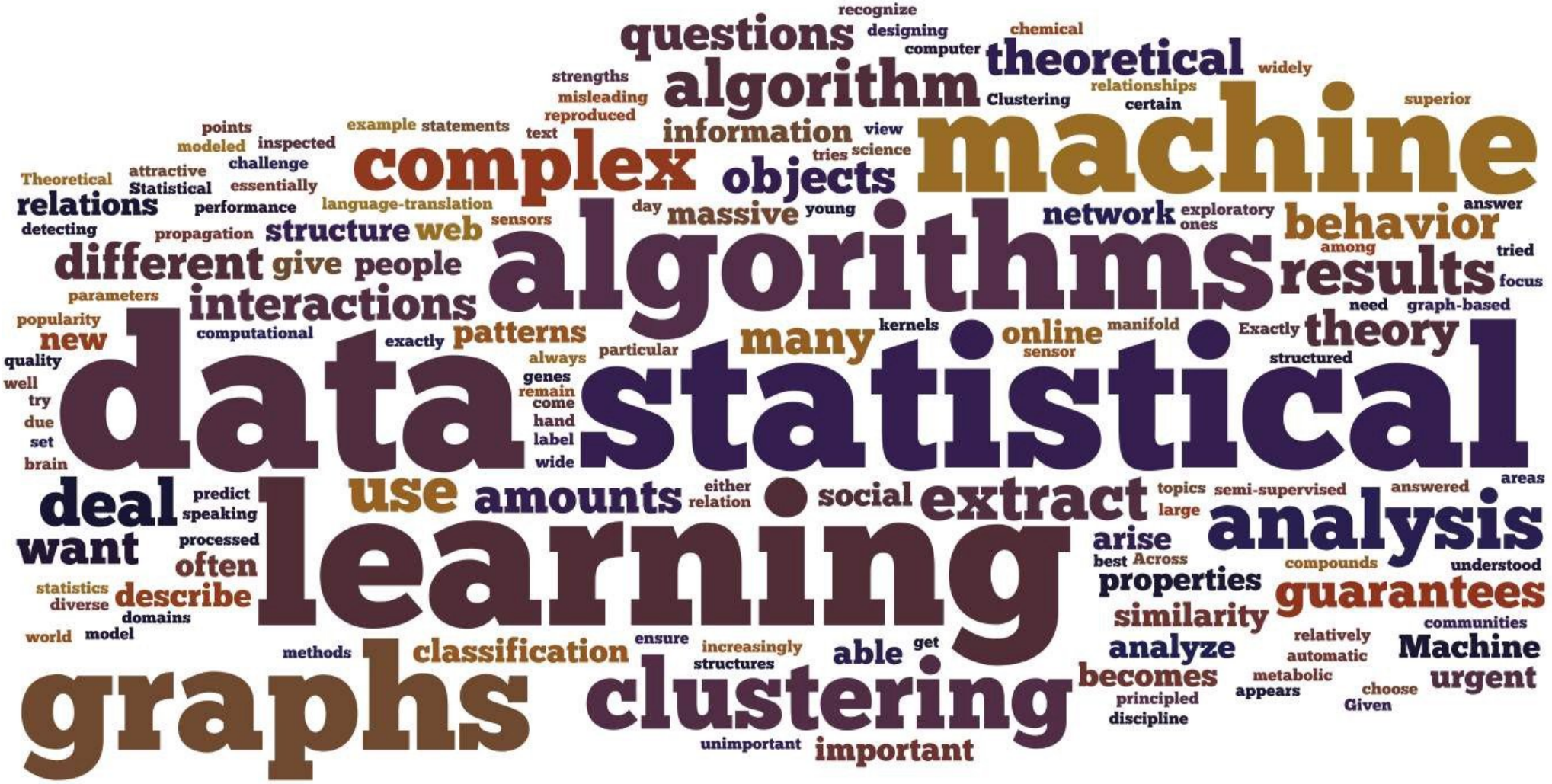
Zasada działania





Zastosowania

- Mało danych treningowych
- Dane treningowe mogą być niereprezentatywne w stosunku do granicy decyzyjnej
- Granice decyzyjne są złożone
- Dane wykazują tendencję do tworzenia skupień


recognize
questions
designing
chemical
computer
theoretical
widely
strengths
algorithm
relationships
misleading
Clustering
certain
superior
reproduced
machine
points
example
statements
information
view
modeled
inspected
text
tries
science
challenge
complex
objects
Theoretical
attractive
Statistical
essentially
relations
performance
language-translation
day
massive
young
network
exploratory
ones
answer
behavior
detecting
propagation
structure
web
sensors
among
tried
different
give
people
algorithms
results
focus
parameters
interactions
need
graph-based
popularity
new
computational
exactly
patterns
kernels
many
online
manifold
Exactly
theory
quality
always
particular
sensor
structured
well
data
statistical
try
genes
remain
come
due
hand
set
label
brain
wide
areas
deal
predict
use
amounts
either
relation
social
extract
topics
semi-supervised
answered
speaking
large
analysis
want
processed
arise
often
best
Across
compounds
understood
properties
guarantees
statistics
diverse
describe
similarity
communities
domains
relatively
world
model
analyze
Machine
methods
classification
ensure
increasingly
structures
able
get
automatic
becomes
metabolic
urgent
graphs
clustering
principled
appears
choose
Given
discipline
unimportant
important