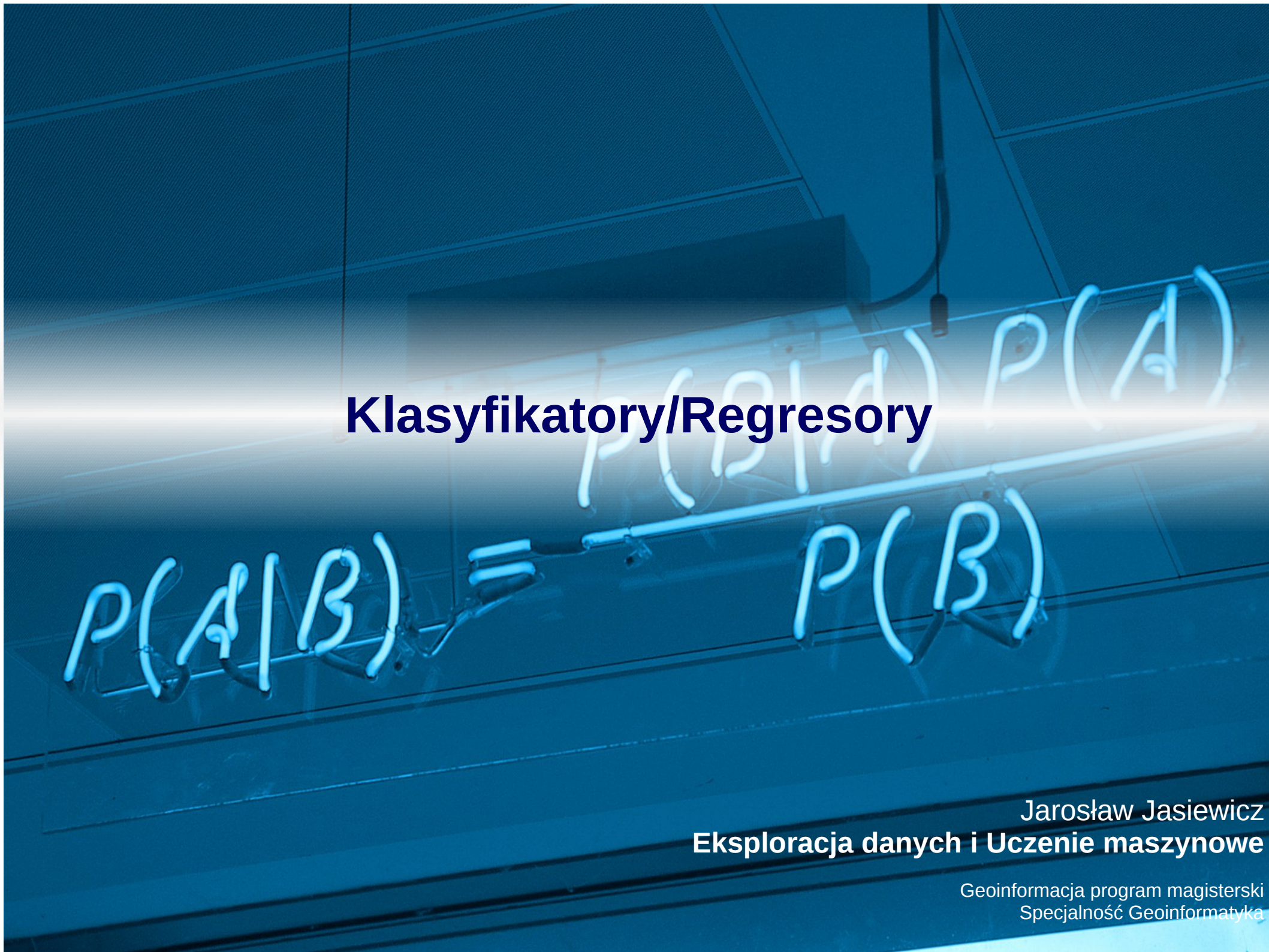


Klasyfikatory/Regresory

Jarosław Jasiewicz

Eksploracja danych i Uczenie maszynowe

Geoinformacja program magisterski
Specjalność Geoinformatyka

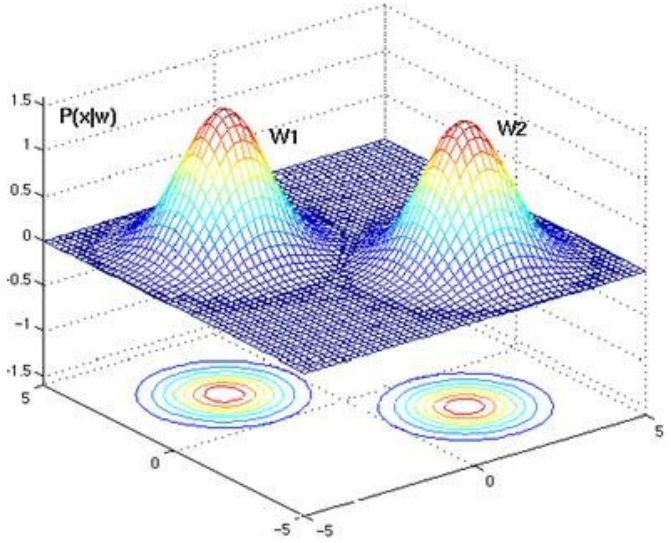
Wybrane metody regresji/klasyfikacji

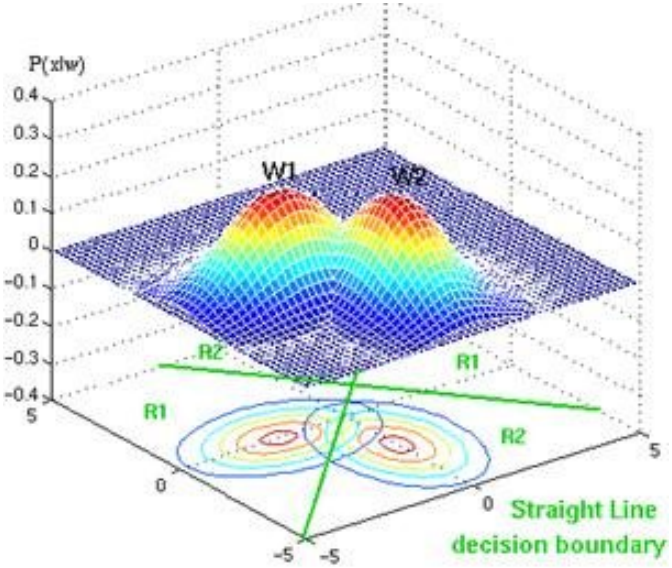
- Naiwny klasyfikator Bayesa
- Metody najbliższego sąsiada
- Metody dyskryminacyjne
- Metoda częściowych najmniejszych kwadratów
- Drzewa decyzyjne
- Metody adaptacyjne
- Sieci neuronowe
- Wektory wsparcia
- Metody wzmacniane

Naiwne klasyfikatory Bayesa

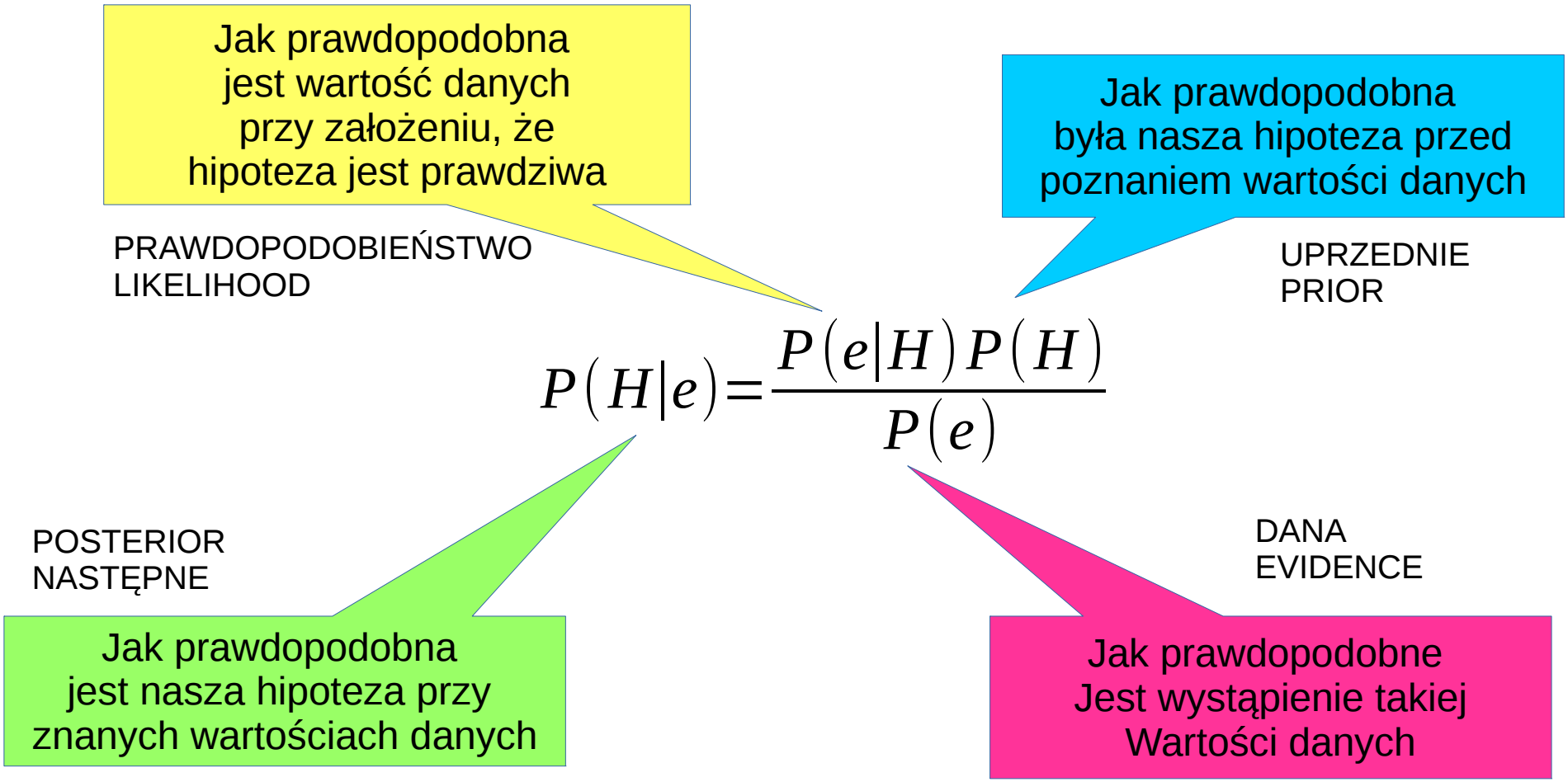
Grupa prostych klasyfikatorów probabilistycznych wykorzystująca (naiwne) założenie o **silnej niezależności** pomiędzy zmiennymi uczącymi. Nie jest stosowany w regresji

- Zalety:
  - szybkość i skalowalność, prostota
  - Łatwość treningu na małym zbiorze danych
- Wady
  - Założenie o niezależności predyktorów (zmiennych wyjaśniających)
  - Nie potrafi uwzględnić zależności między cechami





Co jest czym w prawie Bayesa



Przykład

Pacjent ma gorączkę, czy ma grypę?

- POSTERIOR: Jakie jest prawdopodobieństwo że pacjent z gorączką ma grypę ???
- LIKELIHOOD: jakie jest prawdopodobieństwo gorączki u chorego na grypę **96%**
- PRIOR: jakie jest prawdopodobieństwo złapania grypy? **1.2%**
- EVIDENCE: jakie jest prawdopodobieństwo wystąpienia gorączki **3.4%**

Odpowiedź: (0.96*0.012)/0.034 = **33.9%**

Dlaczego:

Pomimo, że gorączka w czasie grypy jest prawie pewna, to jednak występuje ona 3x częściej w innych przypadkach nie związanych z grypą

Przykład z teledetekcji

Klasyfikacja obrazu teledetekcyjnego

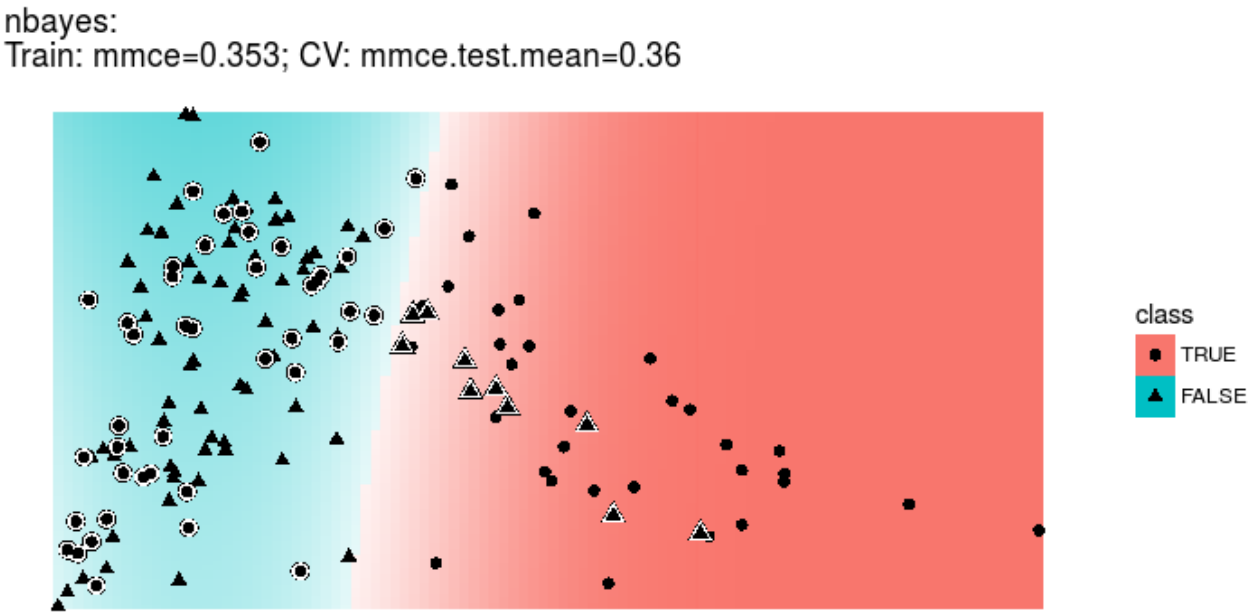
- POSTERIOR: Jaka jest skuteczność klasyfikatora do wykrywania budynków ???
- LIKELIHOOD: jaka jest ogólna skuteczność klasyfikatora **71%**
- PRIOR: jaka jest częstość budynków? **12.7%**
- EVIDENCE: jakie jest odbicia spektralnego typowego dla budynku **12.9%**

Odpowiedź: (0.71*0.127)/0.129 = **69.8%**

Dlaczego:

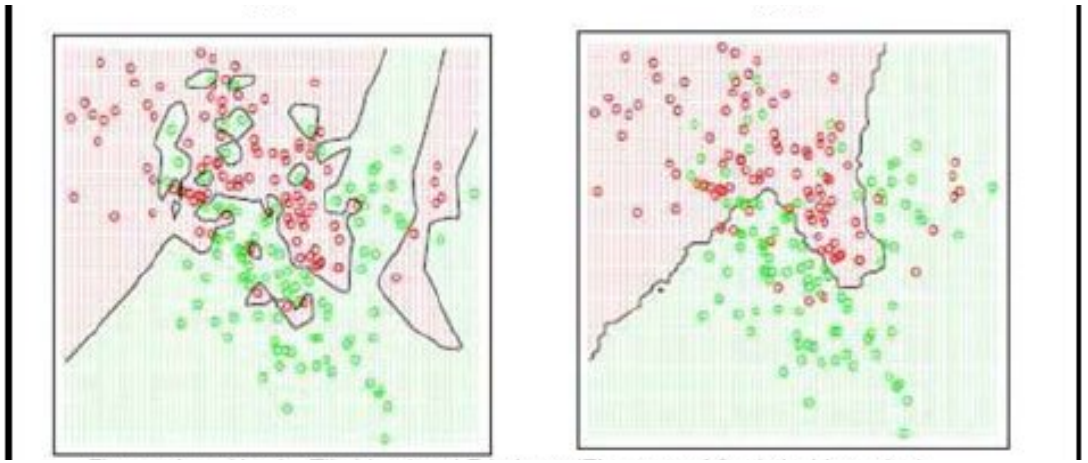
Cecha na podstawie której klasyfikator rozpoznaje budynki występuje z podobnym pokryciem do budynków

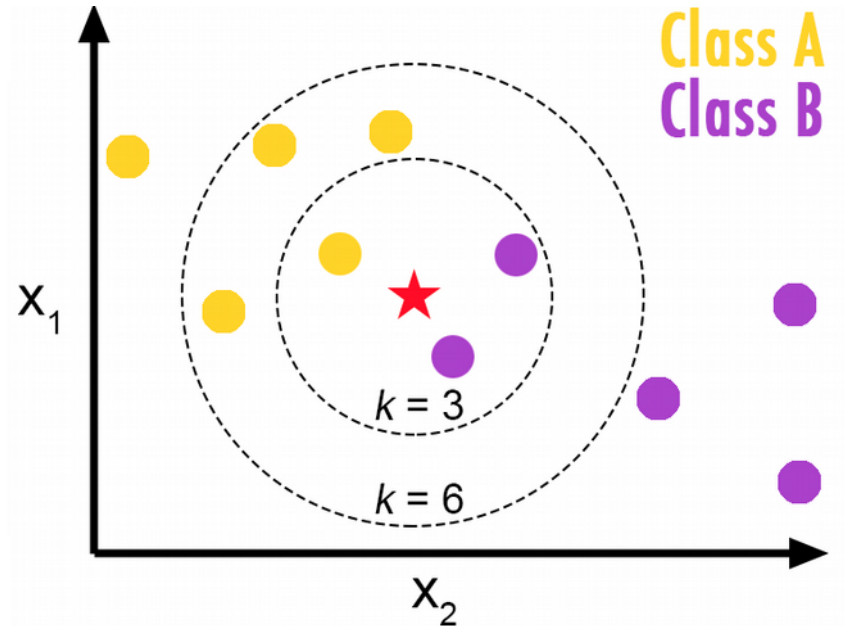
Przestrzeń decyzyjna Naive Bayes



Algorytm najbliższego sąsiada

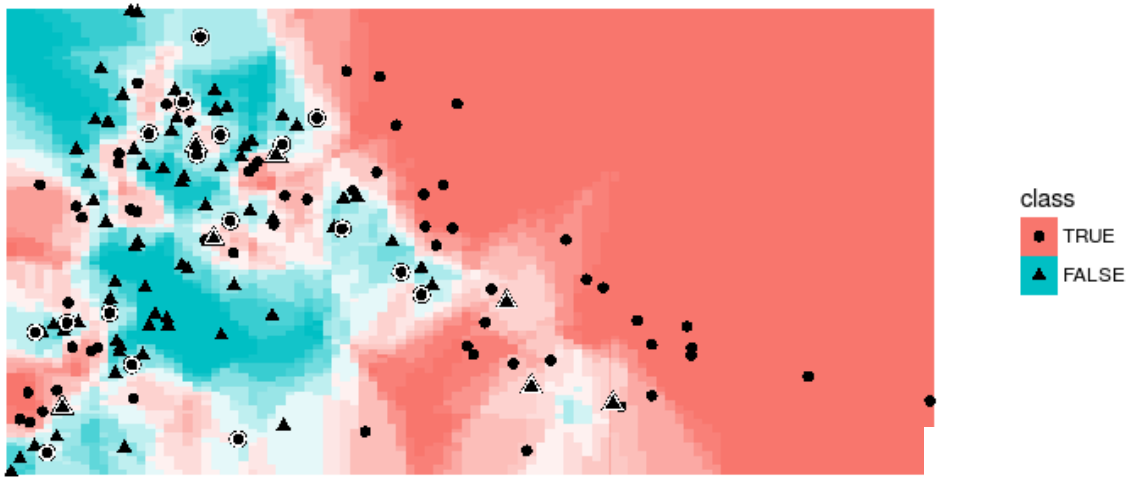
- Nieparametryczny algorytm podejmujący decyzję na podstawie właściwości otoczenia w n-wymiarowej przestrzeni, na podstawi k najbliższych sąsiadów. W przypadku klasyfikacji jest to wybór na podstawie większości, w przypadku regresji – wartość średnia
- Klasyfikator łatwy do przeuczenia jeżeli k małe (np. 1), przy wzrastającym k rośnie generalizacja ale też czas obliczeń



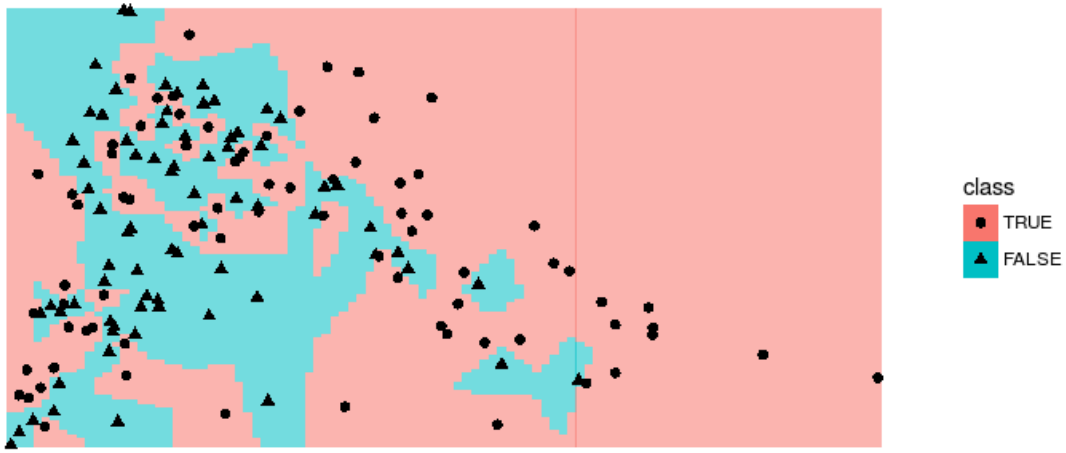


Przestrzeń decyzyjna kNN

kknn: k=5; scale=FALSE
Train: mmce=0.16; CV: mmce.test.mean=0.46



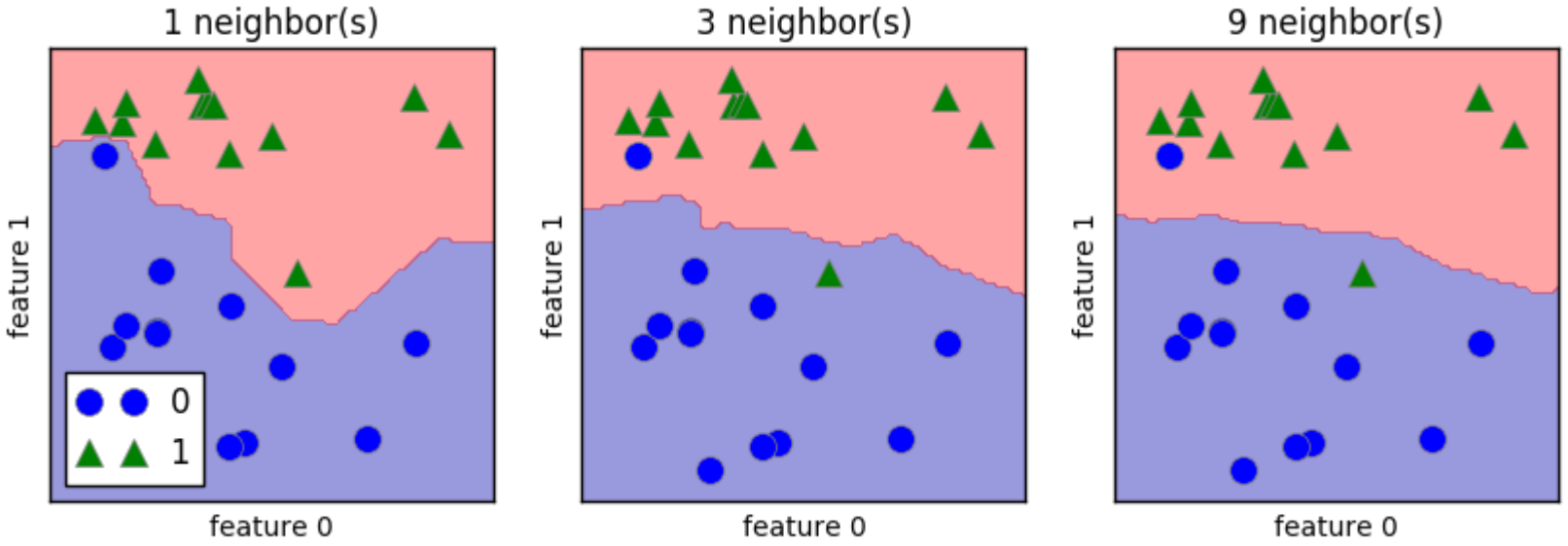
kknn: k=1; scale=FALSE
Train: mmce= 0; CV: mmce.test.mean=0.46



Wersja przeuczona k =1

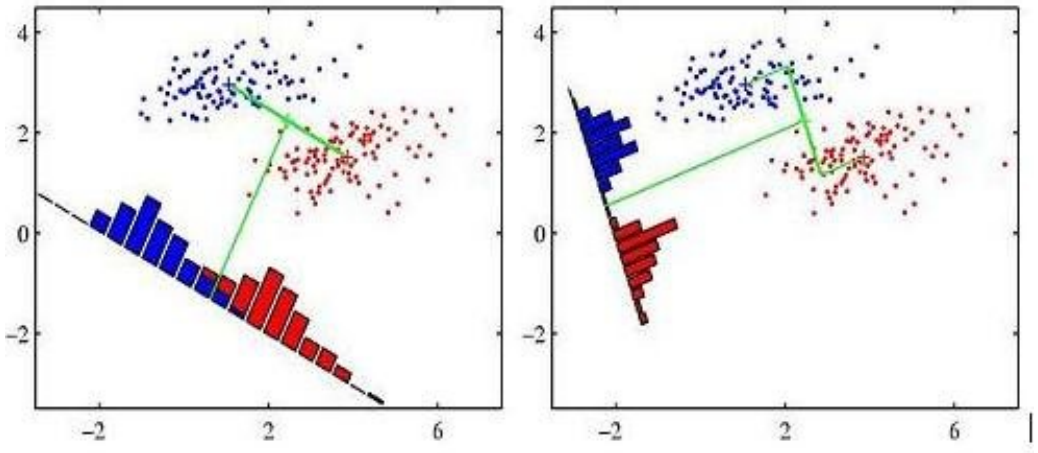
Przeuczenie kNN

overfitting



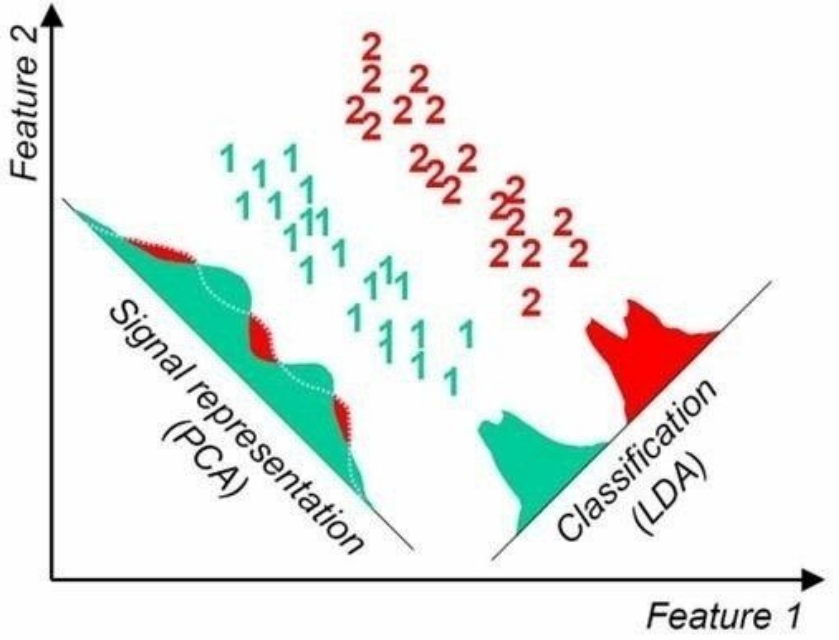
Analiza dyskryminacyjna

- (Linear) **Discriminant Analysis**
- Algorytm, który poszukuje liniowej kombinacji cech, dla której dseparowanie powyższych klas jest największe.
- Stosowana w **klasyfikacji** (nie nadaje się do **regresji**), lub też jako metoda redukcji wymiarowości
- Oprócz dyskryminacji liniowej stosuje się też odmiany potęgowe i kernelowe

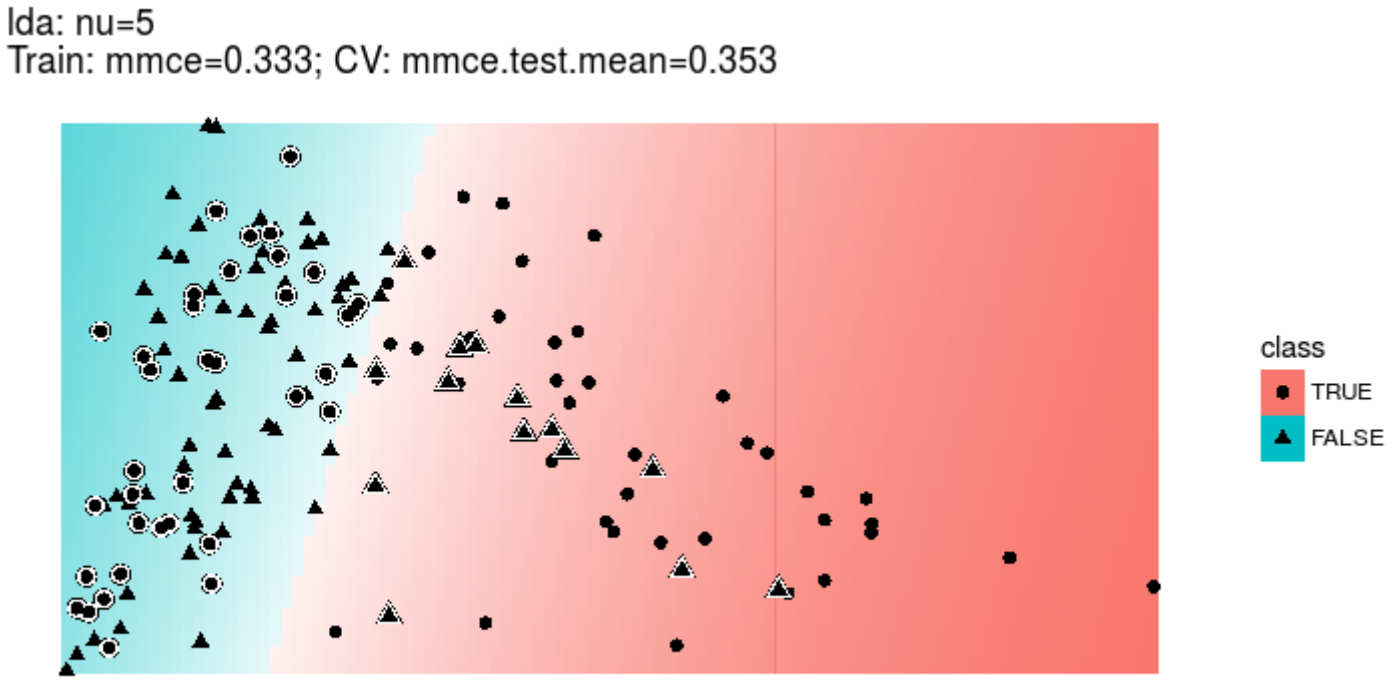


LDA i PCA

- LDA ma wiele związków z PCA, jako metoda poszukiwania liniowej kombinacji zmiennych, tak aby najlepiej wyjaśnić zmienność w obrębie struktury danych
- Analiza dyskryminacyjna wymaga zmiennej zależnej



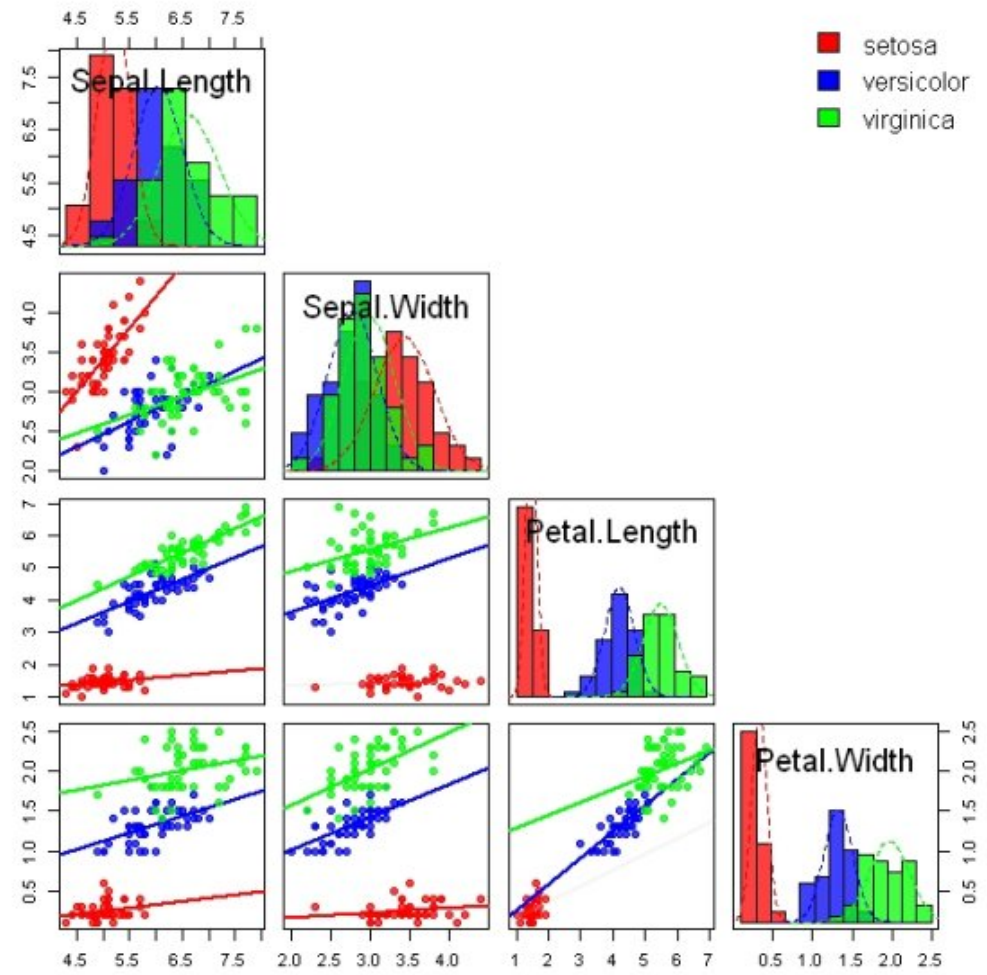
Przestrzeń decyzyjna LDA

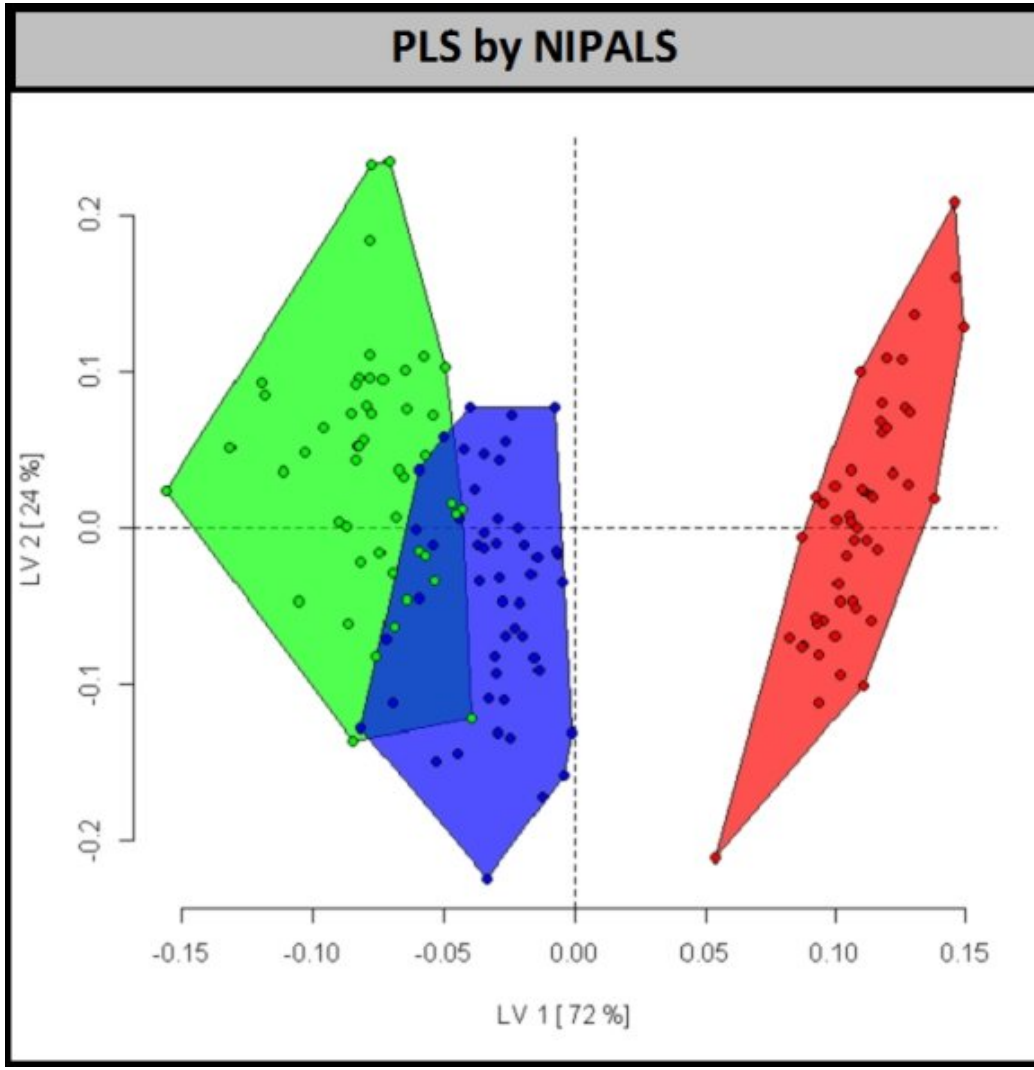


Metoda częściowych najmniejszych kwadratów

- **Partial least squares** – metoda powiązana z PCA i LDA. W przeciwieństwie do LDA nie szuka hiperpłaszczyzny maksymalizującej wariancję w zmiennych niezależnych dla poszczególnych klas poszukuje zależności liniowej poprzez projekcję zmiennych zależnych i niezależnych do innej przestrzeni (jak PCA)
- Partial least squares -DA odmiana metody dla danych kategoryzacyjnych
- Metoda ma zastosowanie, gdy mamy więcej zmiennych niż obserwacji (typowe dla danych ekologicznych)

Jak działa PLS

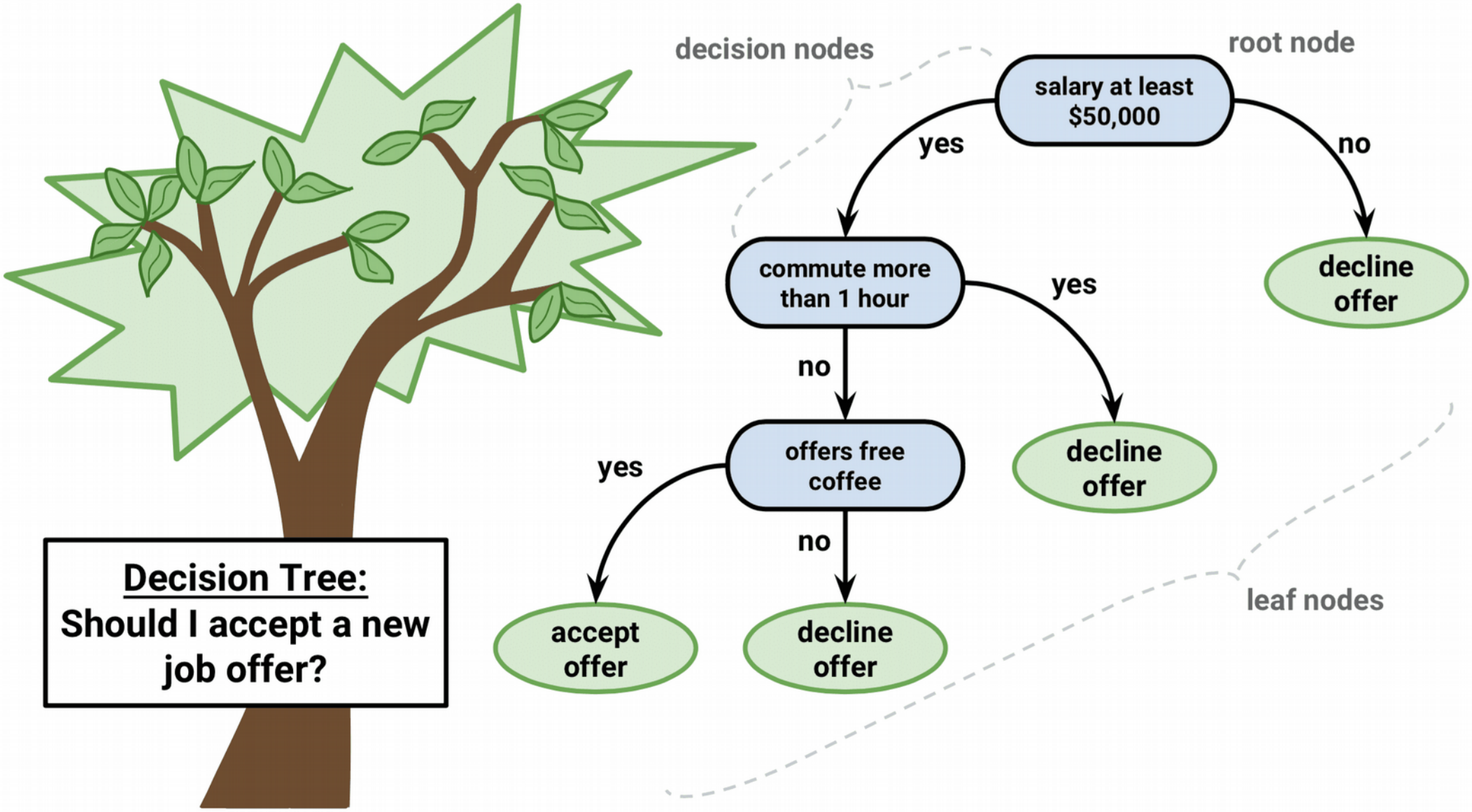




Drzewa klasyfikacyjno-regresyjne

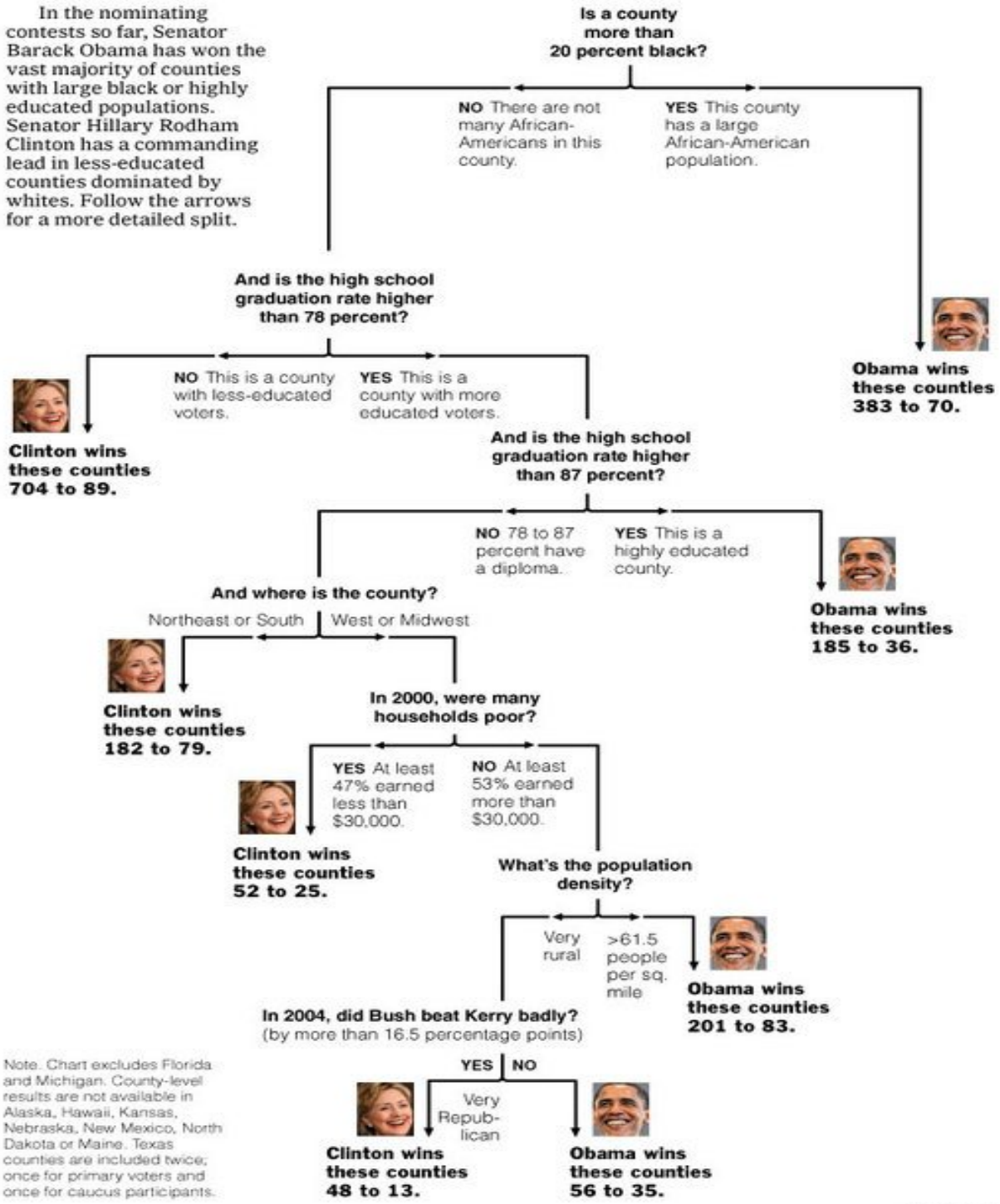
- **Classification and regression trees**
- Popularna metoda uczenia, polega na przewidzeniu klasy (**drzewo klasyfikacyjne**) lub wartości (**drzewo regresyjne**) zmiennej zależnej na podstawie znajdowania reguł w zmiennych wyjaśniających
- Prosta metoda polegająca na kolejnych podziałach, gdzie każdy podział maksymalizuje różnice pomiędzy klasami docelowymi
- Proces podziałów kończy się, kiedy liść zawiera albo czystą klasę, albo dalsze podziały nie są możliwe. W celu uniknięcia przeuczenia (generalizacji modelu) stosuje się przycinanie (pruning)
- Mocne strony:
  - Szybka metoda
  - Przejrzyste kryteria decyzyjne
- Słabe strony
  - Zachłanny algorytm
  - Łatwość przeuczenia
- Odmiany: C45, C50, Qubist

Drzewa decyzyjne



Decision Tree: The Obama-Clinton Divide

In the nominating contests so far, Senator Barack Obama has won the vast majority of counties with large black or highly educated populations. Senator Hillary Rodham Clinton has a commanding lead in less-educated counties dominated by whites. Follow the arrows for a more detailed split.



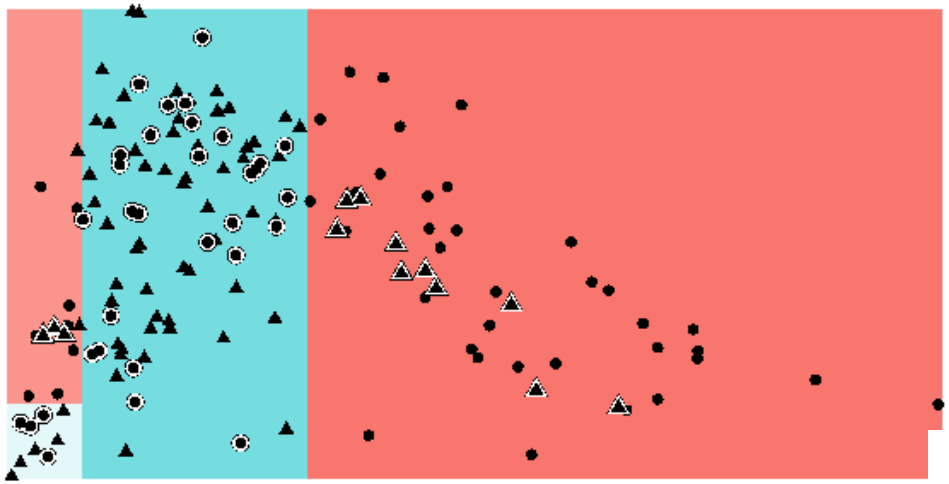
Note. Chart excludes Florida and Michigan. County-level results are not available in Alaska, Hawaii, Kansas, Nebraska, New Mexico, North Dakota or Maine. Texas counties are included twice; once for primary voters and once for caucus participants.

Sources: Election results via The Associated Press; Census Bureau; Dave Leip's Atlas of U.S. Presidential Elections

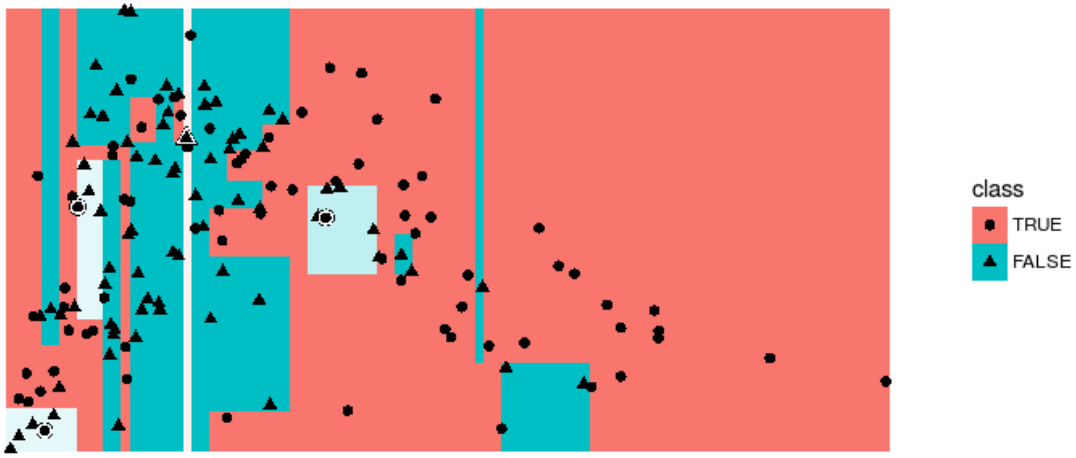
AMANDA COX/
THE NEW YORK TIMES

Przestrzeń decyzyjna CART

rpart: xval=0
Train: mmce= 0.3; CV: mmce.test.mean=0.433

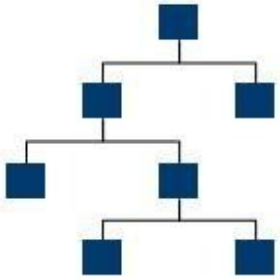
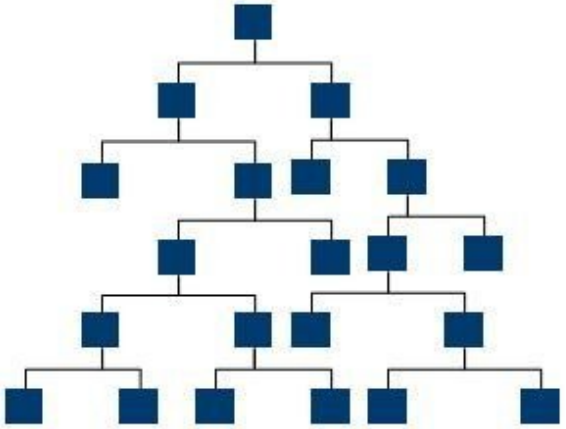
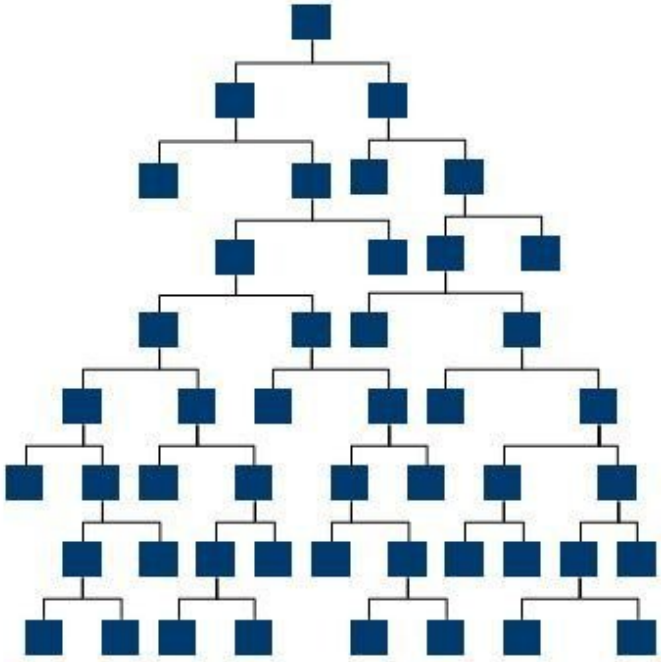


rpart: xval=20; minsplit=3; minbucket=1; cp=0.001
Train: mmce=0.0267; CV: mmce.test.mean=0.567



Wersja przeuczona – nadmiar podziałów

Przeuczenie drzew decyzyjnych

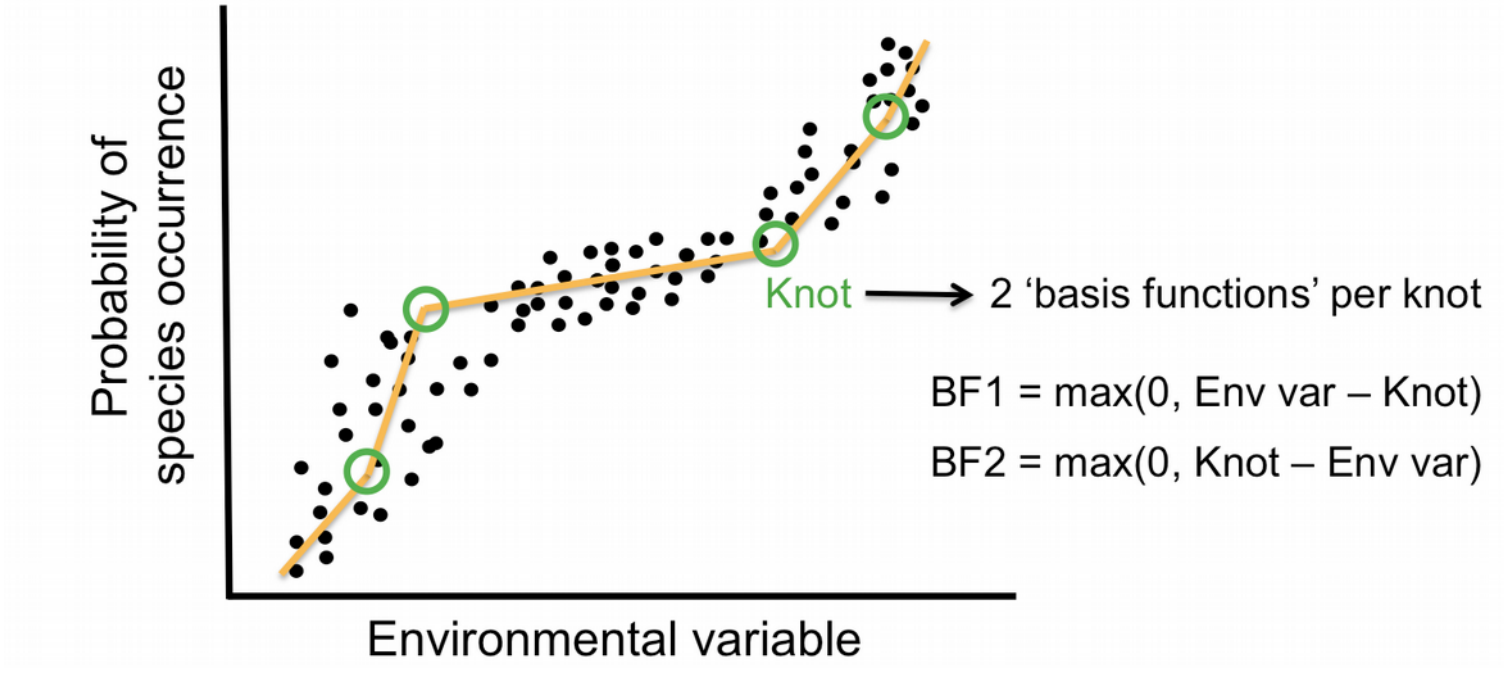
Underfit tree	Optimal tree	Overfit tree
		
Accuracy on training = 50%	Accuracy on training = 70%	Accuracy on training = 90%
Accuracy on test = 50%	Accuracy on test = 70%	Accuracy on test = 65%

Regresja nieliniowa i metody adaptacyjne

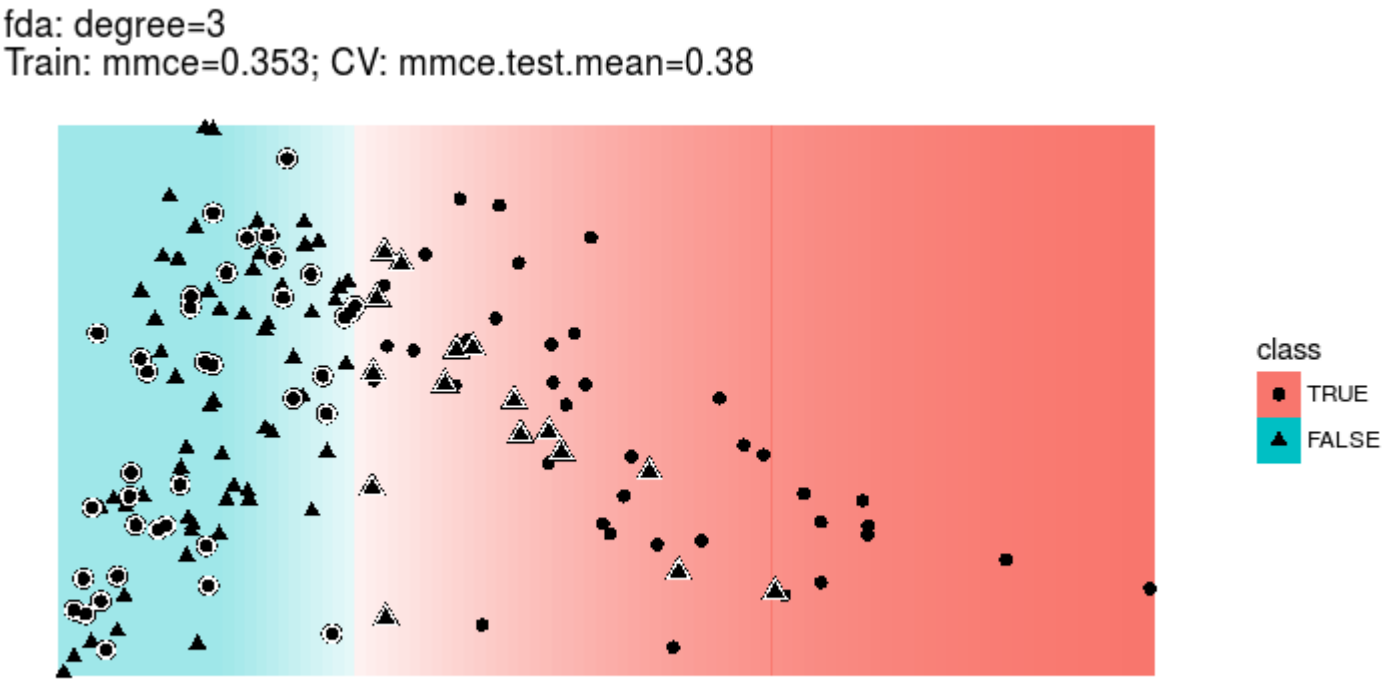
- Technika nieparametrycznej regresji, automatycznie modelująca nieliniowe zależności pomiędzy zmiennymi
- Zaletą metody jest możliwość stosowania zarówno predykatorów dyskretnych jak i ciągłych
- Prosty do zrozumienia i interpretacji, interpretowany
- Nie wymaga transformacji danych wejściowych
- Samodzielnie dobiera zmienne wyjaśniające na podstawie ich wag
- Popularne metody **MARS/Earth, Spline, GAM** (General Addtitive model)

MARS

- Algorytm dokonuje podziału na danych na podgrupy rozdzielone węzłami, czyli obiektami wyznaczającymi przedziały podgrup
- Zastosowaniu prostych modeli spline dla każdego przedziału
- Analizie zmiennych i wyborze oraz przypisaniu wag tym, które przenoszą najwięcej informacji



Przestrzeń decyzyjna - MARS



Modele penalizowane

- Normowane modele regresji liniowej, nie nadają się do klasyfikacji
- Modele penalizowane zapobiegają przeuczeniu
- Przyczyną przeuczenie jest złożoność modelu wyrażająca się wysokimi wartościami niektórych współczynników – **tych które nie są w stanie w sposób prosty wyjaśnić modelu**
- Funkcja kosztu – błąd dopasowania

$$L = \sum(\hat{Y}_i - Y_i)^2$$

- Ogólna postać modelu regresji (dla jednej zmiennej:

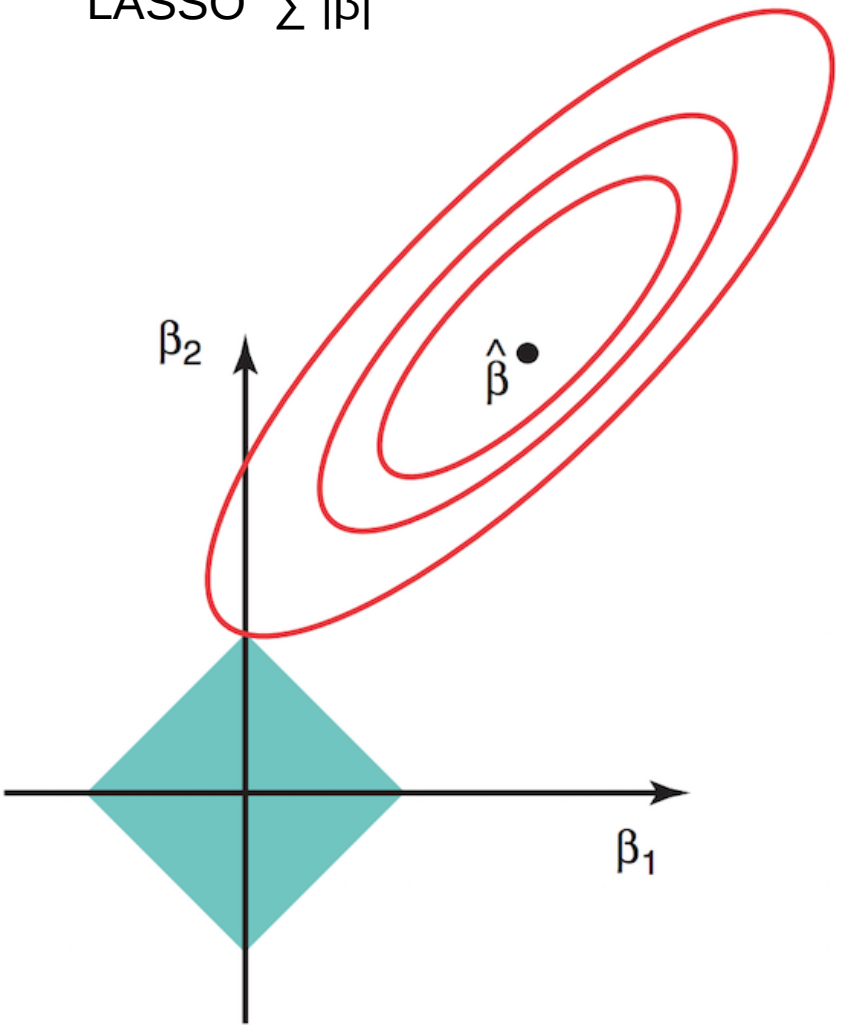
$$\hat{Y} = \beta_0 + \beta_1 X_1$$

- Funkcja kosztu z karą – parametrem normującym: - suma parametrów modelu

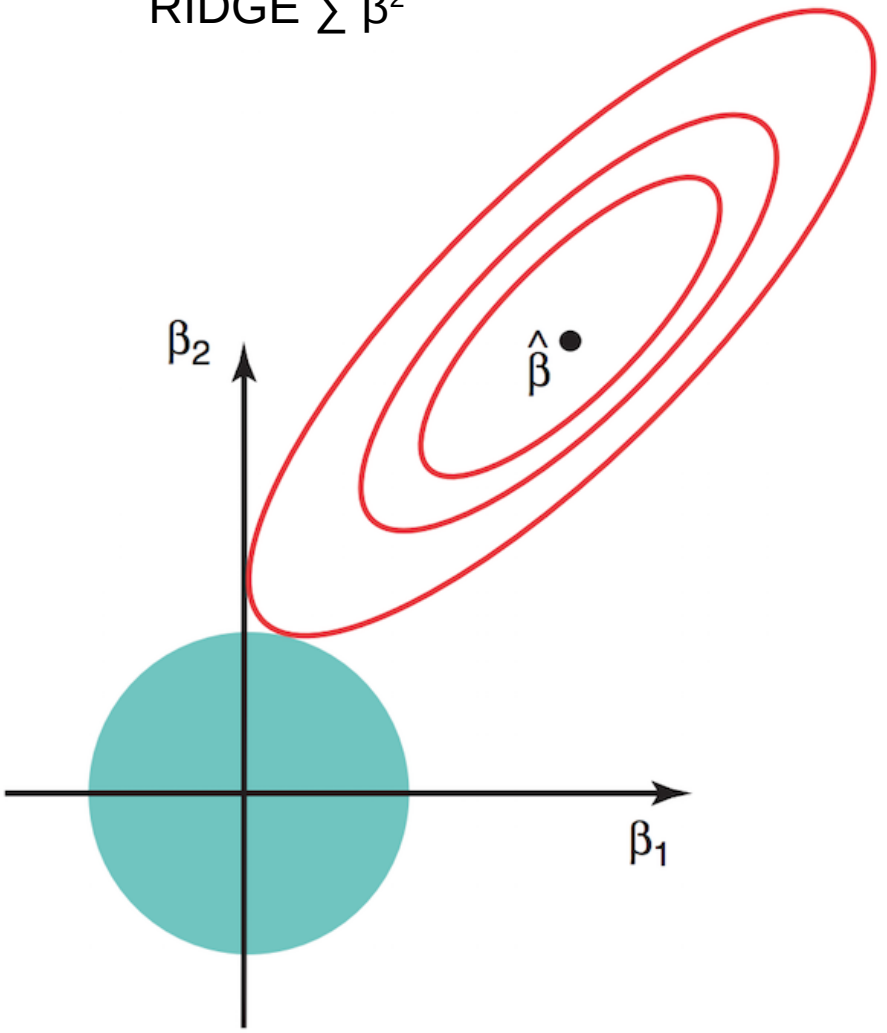
$$L = \sum(\hat{Y}_i - Y_i)^2 + \lambda \sum \beta$$

Lasso vs. Ridge

LASSO $\sum |\beta|$



RIDGE $\sum \beta^2$

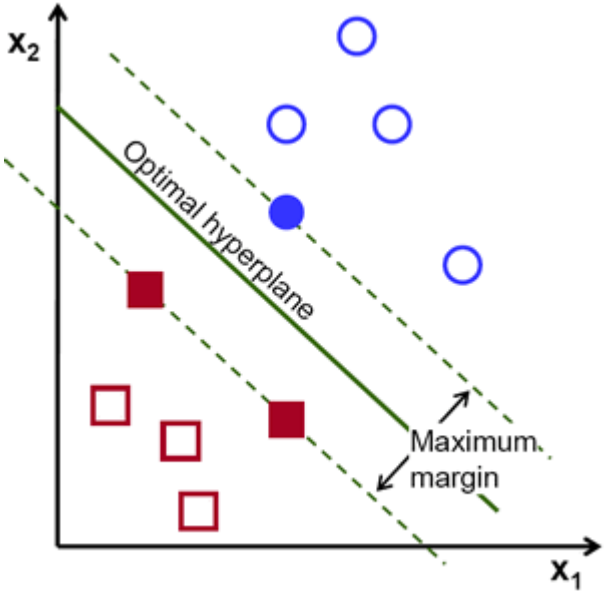


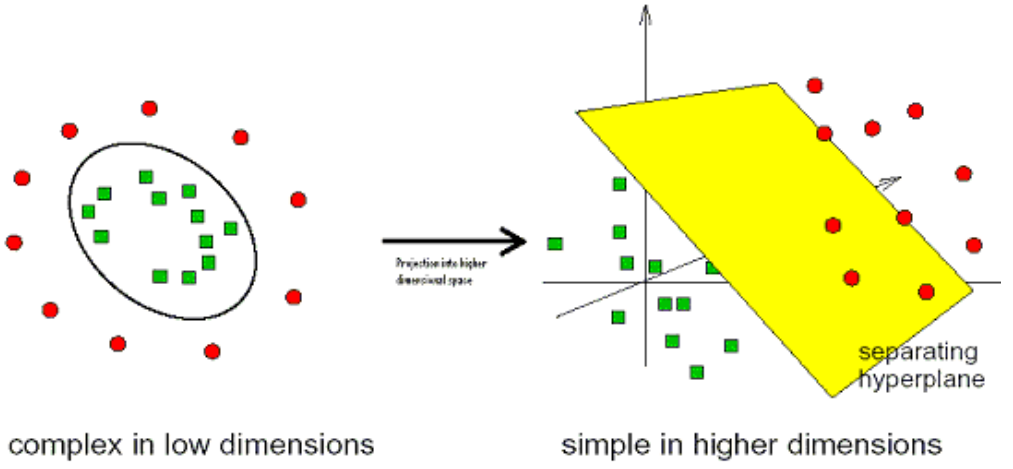
Popularne modele

- Regresja **Ridge** (Norma L2) – nie usuwa nieistotnych cech, jedynie zmniejsza wartości największych (mało istotnych) współczynników regresji
- Regresja **Lasso** (Norma L1) – ustawia współczynniki wysokie współczynniki cech na 0 w ten sposób usuwa nieistotne cechy (shrinkage – kurczenie modelu)
- Regresja **Elastic Net** (połączenie obu norm)
- Norma to inaczej natężenie wektora cech (zbioru zmiennych)

•Maszyny wektorów wsparcia

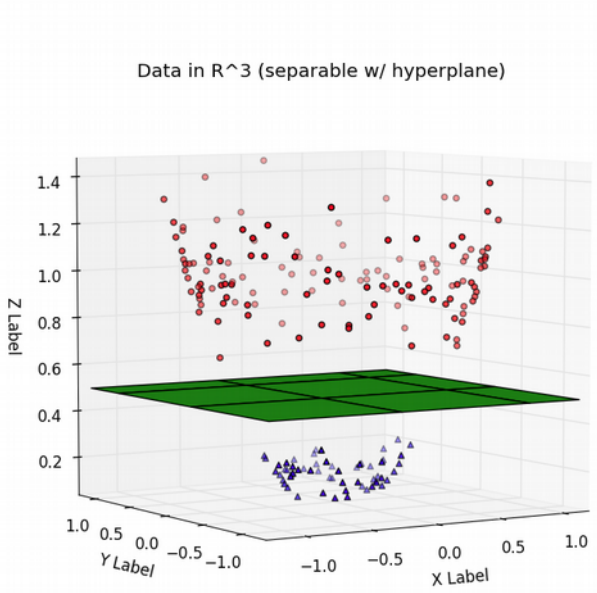
- **Support vector machines**
- Klasyfikatory liniowe, podstawą działania jest maksymalizacja marginesu pomiędzy dwoma klasami – odległości pomiędzy dwoma wektorami podpierającymi płaszczyznę rozdzielającą
- Jeżeli klasy nie są możliwe do rozdzielenia liniowo stosuje się funkcje jądrowe poprzez przeniesienie problemu z mniejszej do większej liczby wymiarów
- Obiekty trudne do liniowego rozdzielenia w mniejszej liczbie wymiarów są łatwo rozdzielane hiperpłaszczyzną w większej

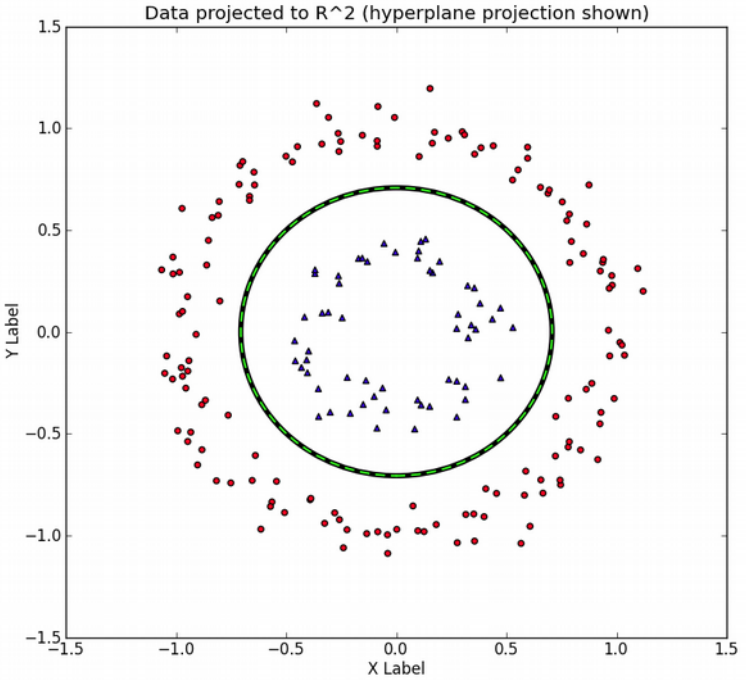




Funkcje kernelowe (jądrowe)

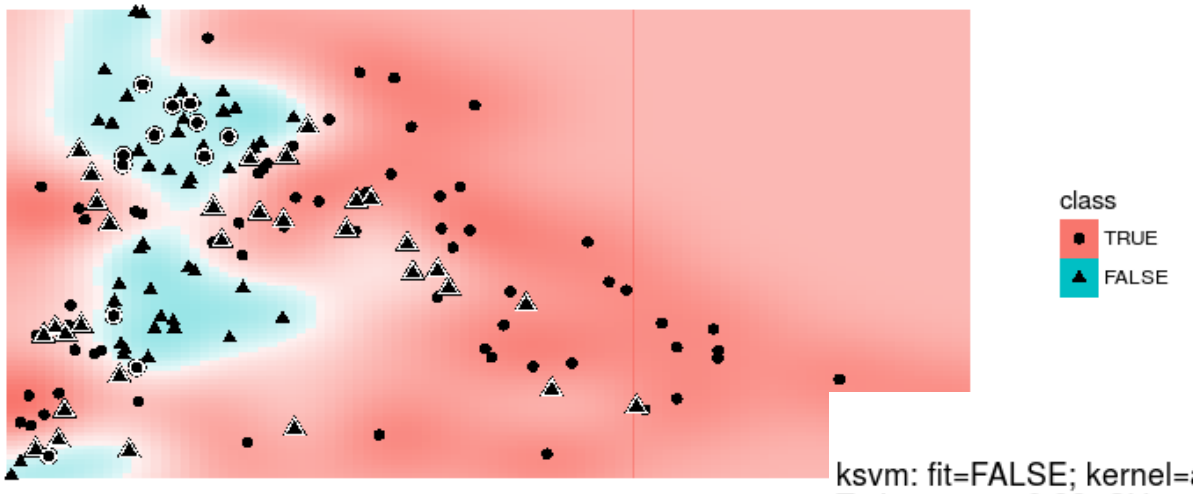
- Funkcje jądrowe (wielomianowa, gaussowska, sigmoidalna i inne) działają na zasadzie dodania dodatkowego wymiaru
- np: dla x i y dodajemy nowy wymiar z, zależny od x i y



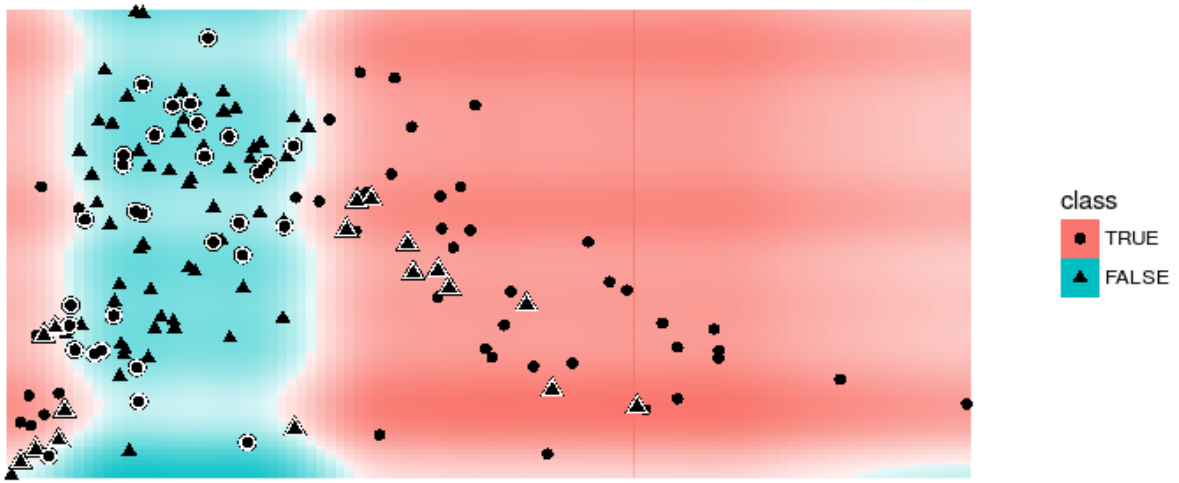


Przestrzeń decyzyjna – SVM

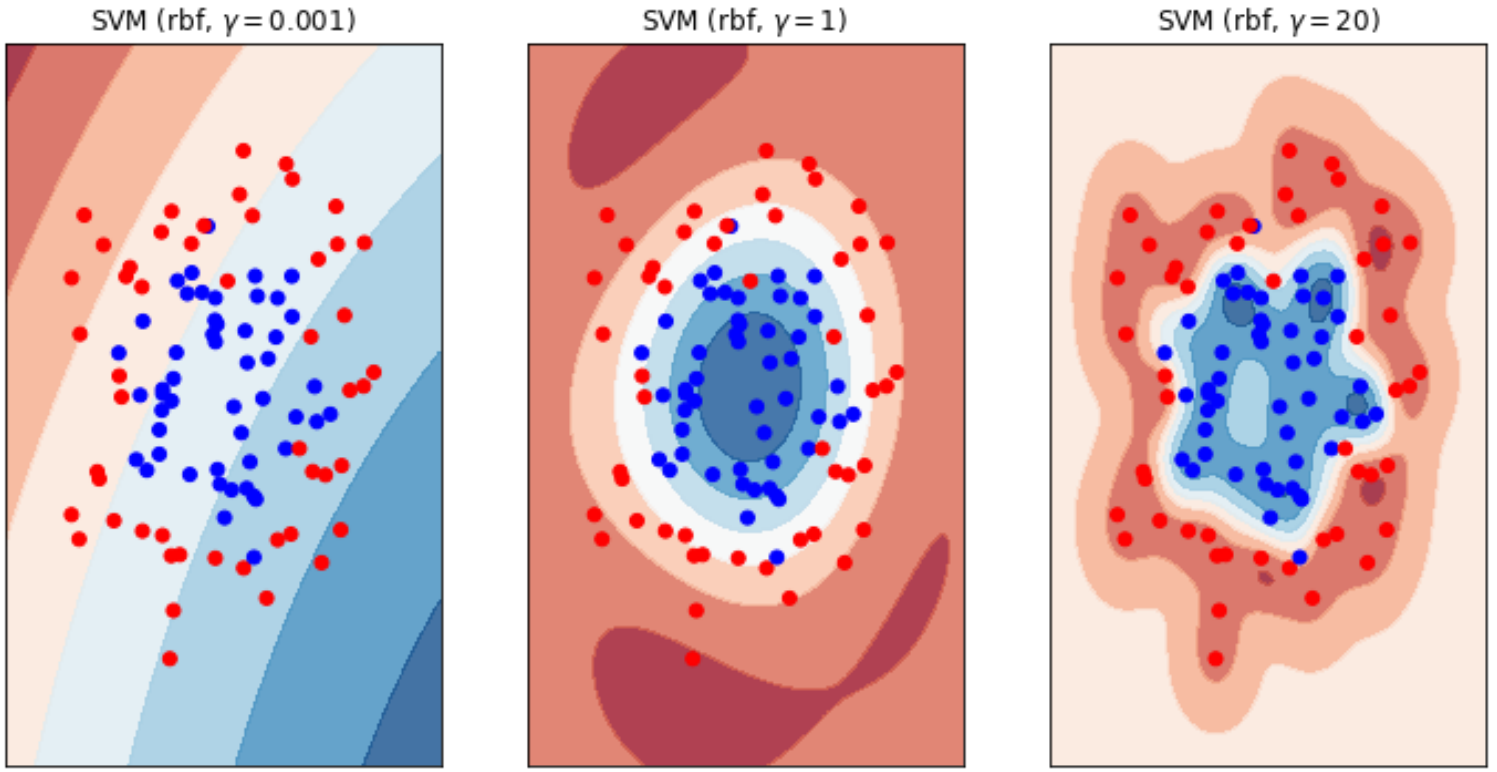
ksvm: fit=FALSE; kernel=rbfdot; sigma=5
Train: mmce=0.287; CV: mmce.test.mean=0.44



ksvm: fit=FALSE; kernel=anovadot; sigma=5
Train: mmce=0.32; CV: mmce.test.mean= 0.4

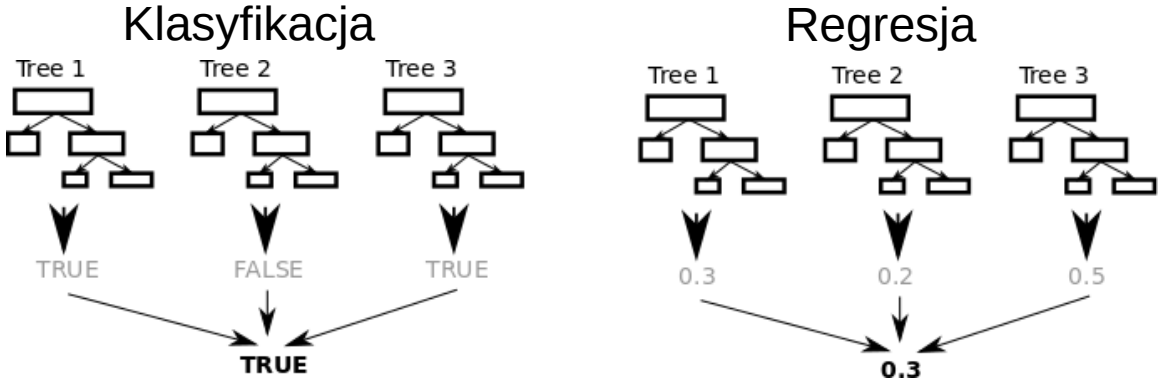


Przeuczenie SVM



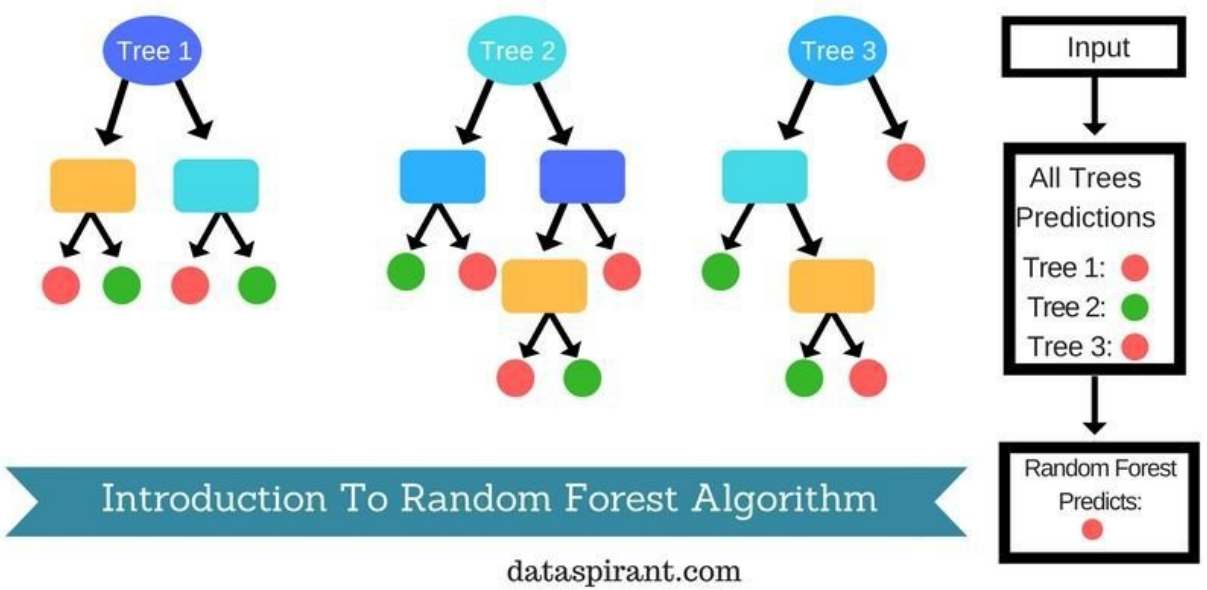
Metody agregujące

- Meta algorytmy, których celem jest jednoczesna redukcja błędu systematycznego i losowego (**bias i variance**). Zakłada że wiele „słabych” klasyfikatorów/regresorów może zostać połączona w jeden silny
- **Bagging** – algorytm agregujący rodzinę klasyfikatorów (np. CART), gdzie wynik klasyfikacji opiera się na głosowaniu większościowym
- **Boosting** – metoda konstruowania kolejnych wersji klasyfikatorów na podstawie losowych ciągów uczących i przypisywaniu wag obiektom z ciągów uczących. Wagi te określają prawdopodobieństwo wylosowania w kolejnej iteracji. Waga wzrasta jeżeli obiekt został błędnie zakwalifikowany. Obiekty błędnie klasyfikowane są częściej losowane co jest pożądane, ponieważ z reguły znajdują się w pobliżu granicy decyzyjnej.



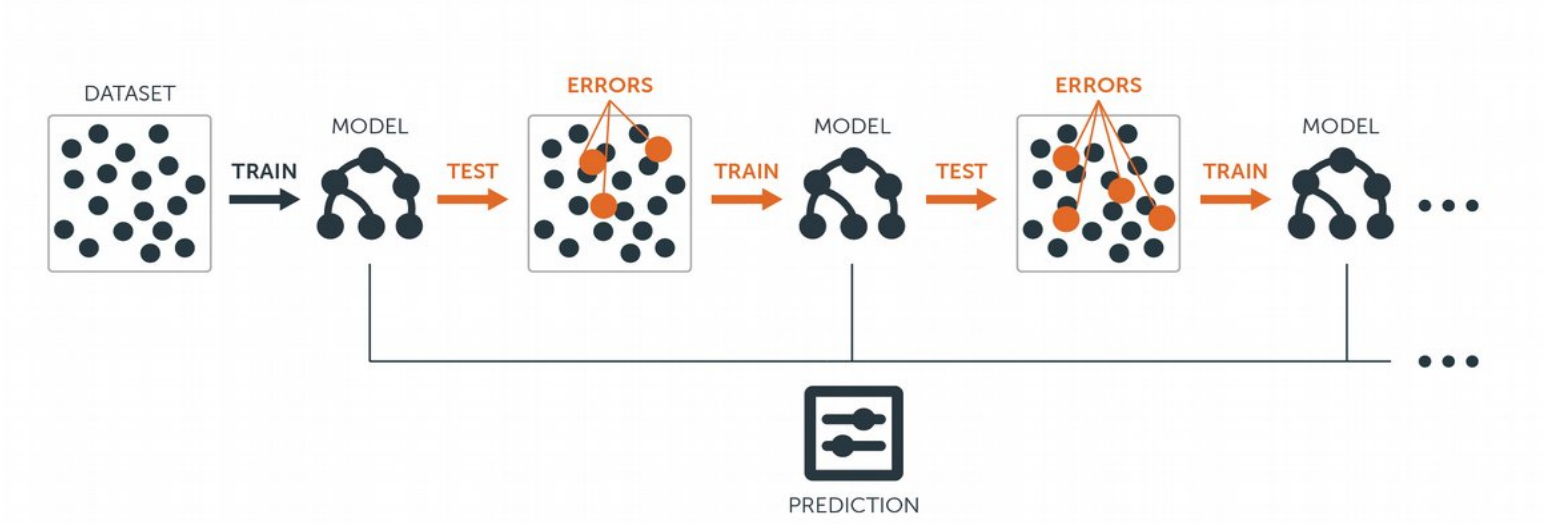
Random Forest - bagging

- Agregujące uogólnienie drzew decyzyjnych, końcowa decyzja jest podejmowana na podstawie głosowania lub uśredniania (regresja).
- Tworzone są małe drzewa
- Losowaniu podlegają zmienne uczące jak i przypadki. Losuje się ograniczoną liczbę cech, dzięki czemu mogą być stosowane zbiory o bardzo dużej liczbie zmiennych wyjaśniających



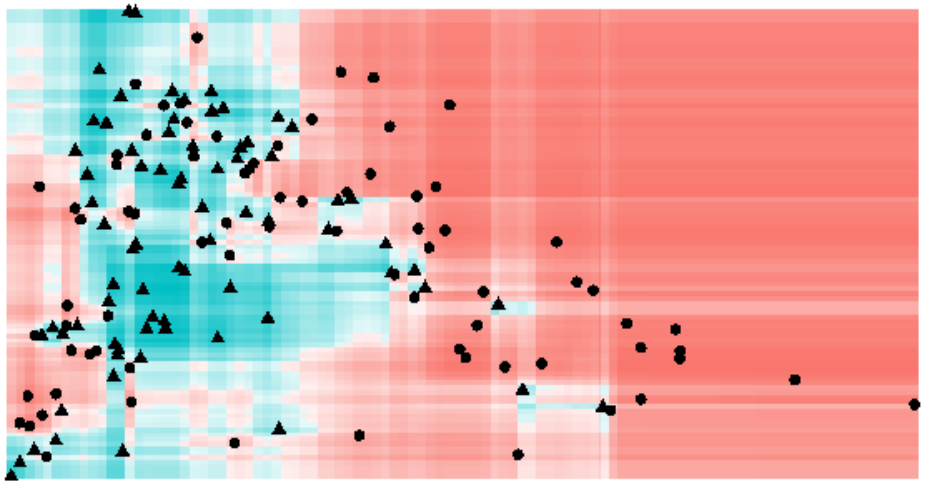
Boosted trees - boosting

- Model tworzy wstępne drzewo na podstawie wartości inicjalnych podziałów dla których klasyfikuje obiekty/ wyznacza wartość
- Błędnie zaklasyfikowane obiekty zostają użyte do wyznaczenia korekty, która poprawi klasyfikację
- Korekta pozwala wyznaczyć gradient zmian, który doprowadzi do wyznaczenia nowych, lepszych parametrów podziału drzewa
- Proces jest powtarzany aż do osiągnięcia zamierzonego celu, ilości założonych iteracji lub nie można wyznaczyć korekty

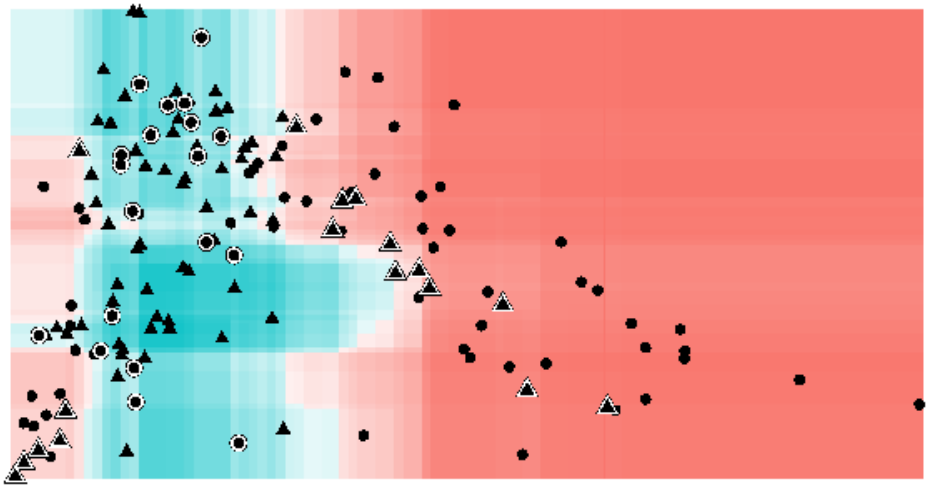


Przestrzeń decyzyjna - metody wzmacniane

rf:
Train: mmce= 0; CV: mmce.test.mean=0.487



ada: xval=0
Train: mmce=0.24; CV: mmce.test.mean=0.44

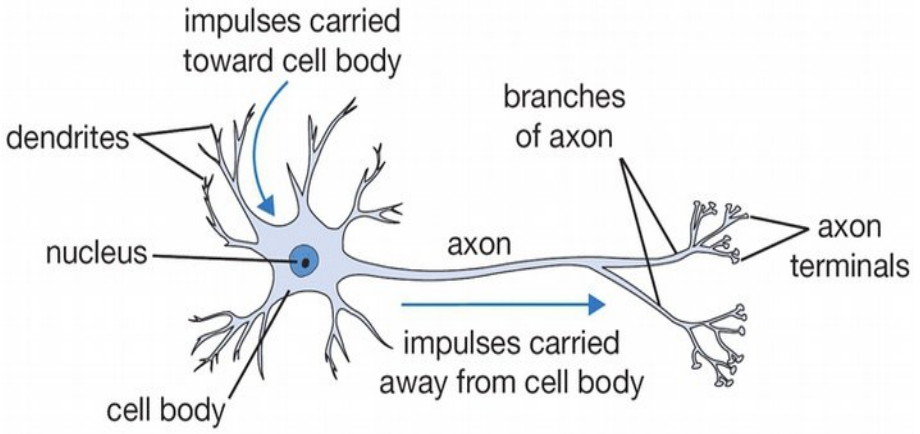


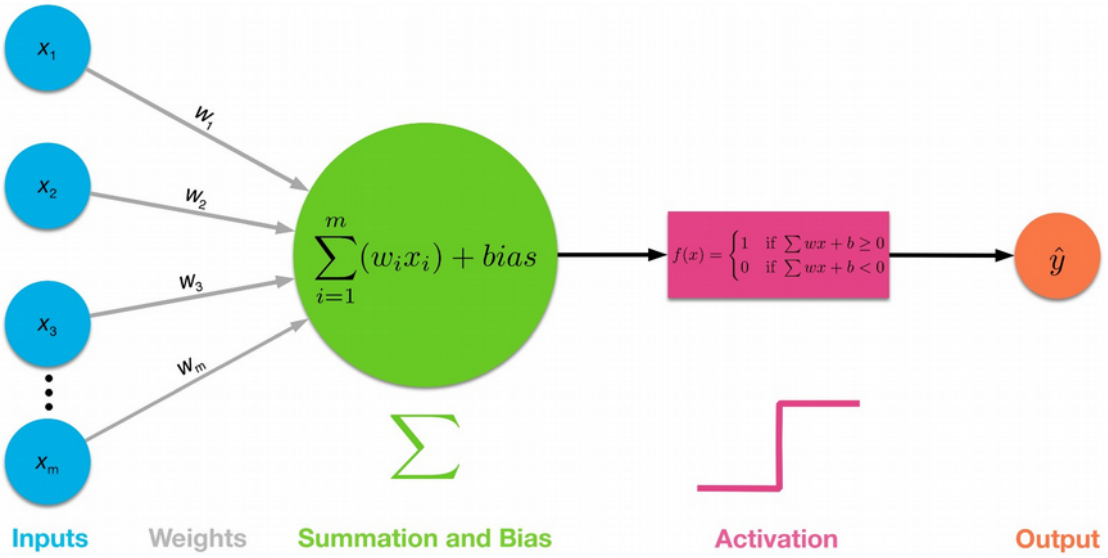
Sieci neuronowe

- Narzędzia przetwarzające sygnał (zmienne wyjaśniające) poprzez rząd elementów zwany sztucznymi neuronami. Każdy neuron wykonuje podstawową operację ważonego sumowania
- Sieci jednokierunkowe – bez sprzężeń zwrotnych
- Sieci rekurencyjne połączenia między neuronami mają charakter cykliczny
- Głębokie sieci – wielowarstwowe sieci
- Samoorganizujące się mapy to też sieci neuronowe

Neuron

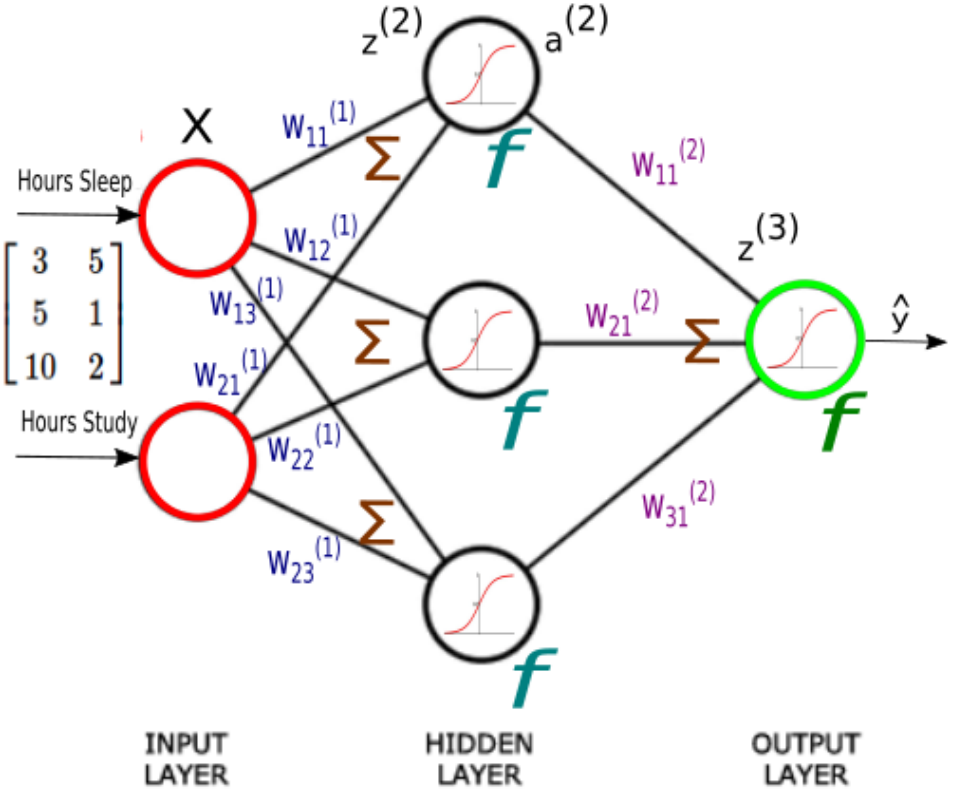
- Uczenie się sieci polega na iteracyjnym dobieraniu wag na wejściu do sztucznego neuronu, tak aby suma wartości wejściowych pomnożonych przez wagi dawała optymalną decyzję
- Sieci neuronowe mogą też być strukturami fizycznymi (sprzętowymi)



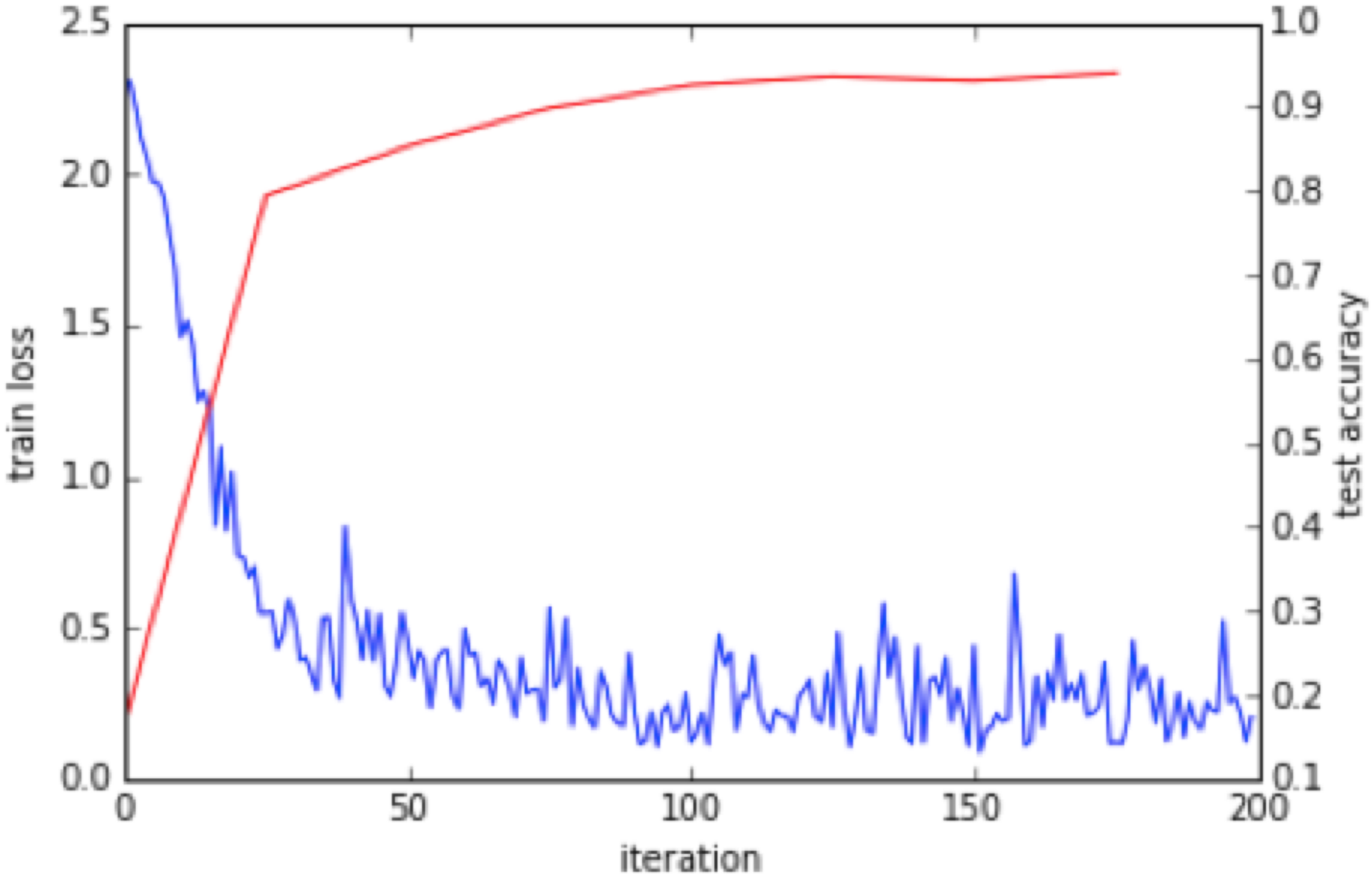


Jak działa sieć neuronowa?

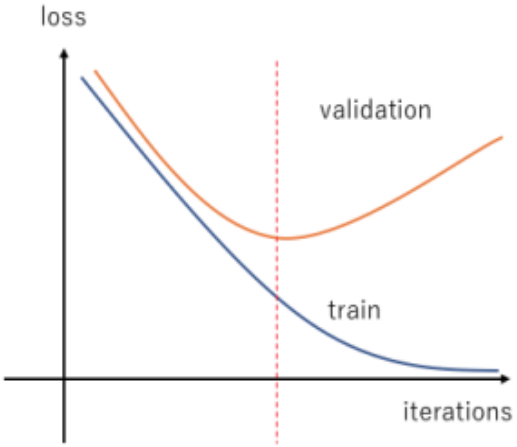
- Działanie sieci opiera się na decyzjach podejmowanych przez poszczególne neurony a następnie uwspólnieniu decyzji



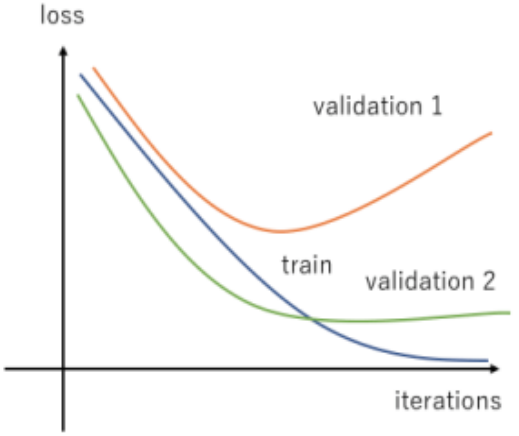
Optymalizacja sieci neuronowych



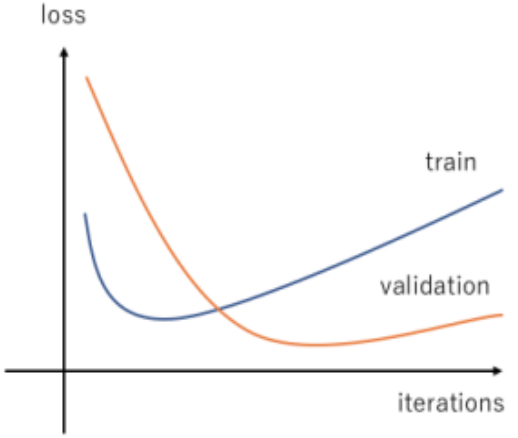
Przeuczenie sieci



Sytuacja klasyczna

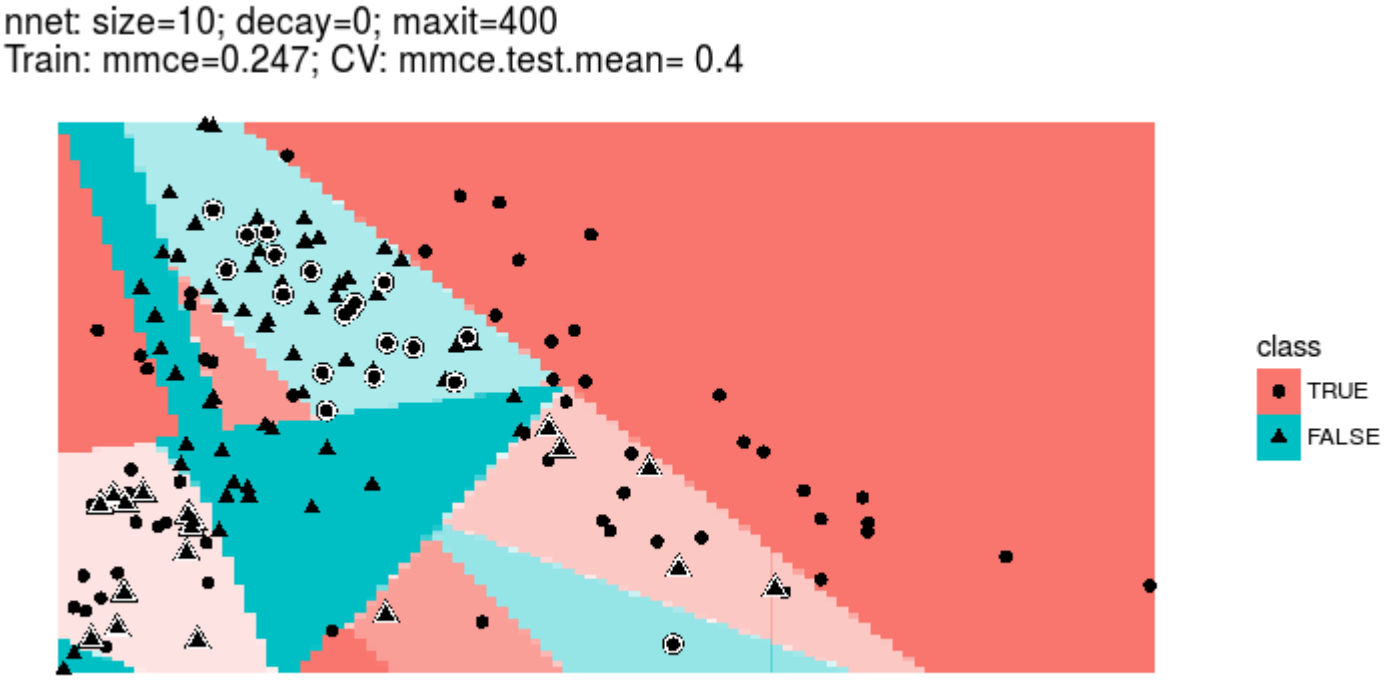


Niepewność co do zbiorów

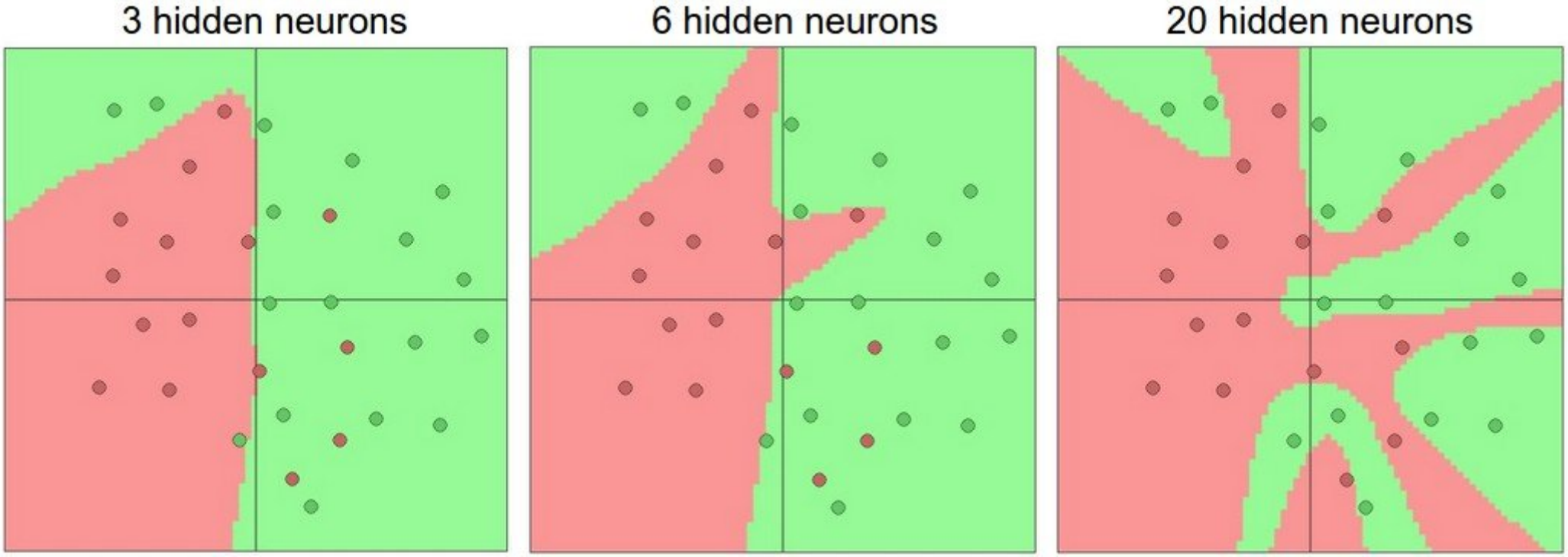


Błędna optymalizacja

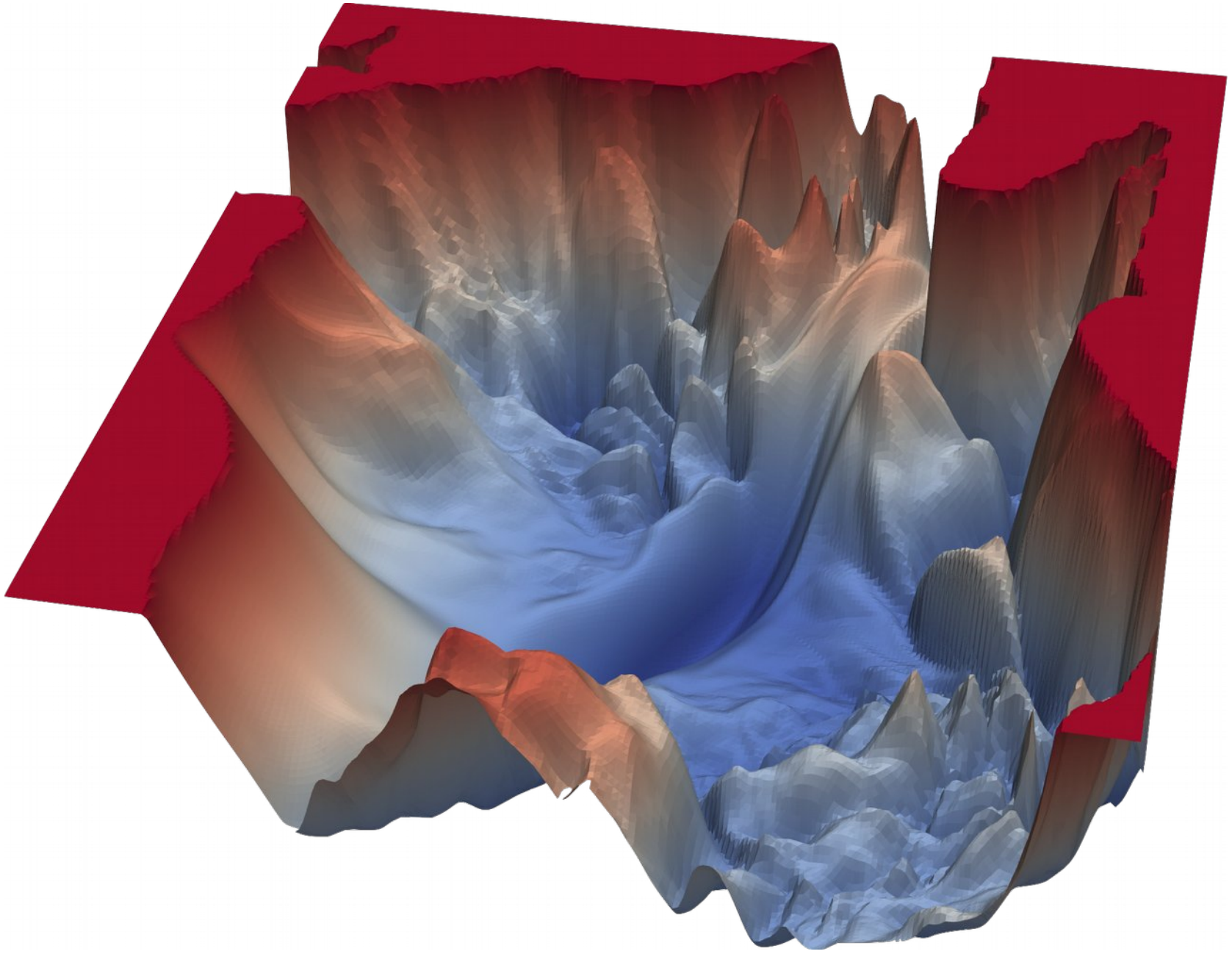
Sieci neuronowe – przestrzeń decyzyjna



Przeuczenie sieci neuronowych

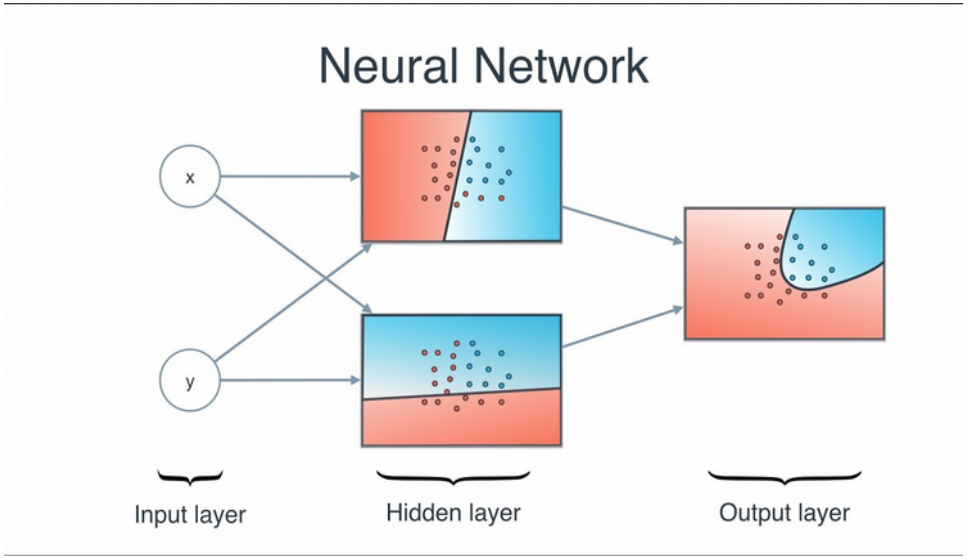


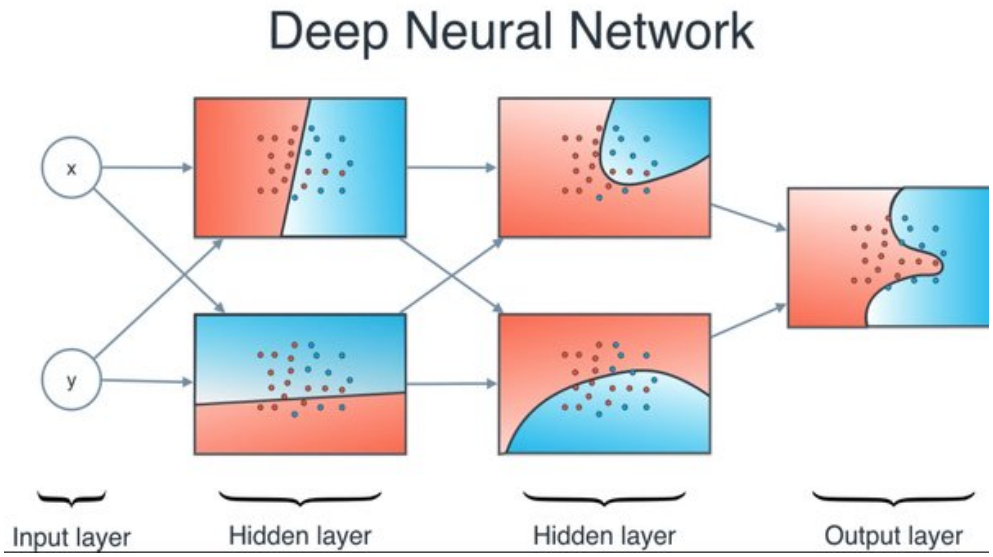
Krajobraz funkcji kosztu

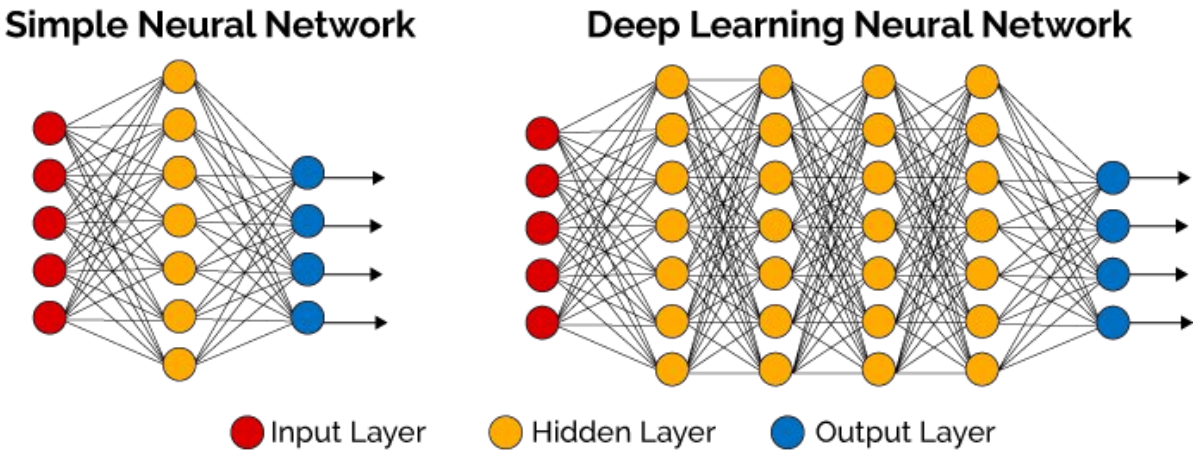


Standardowa Sieć a Deep Learning

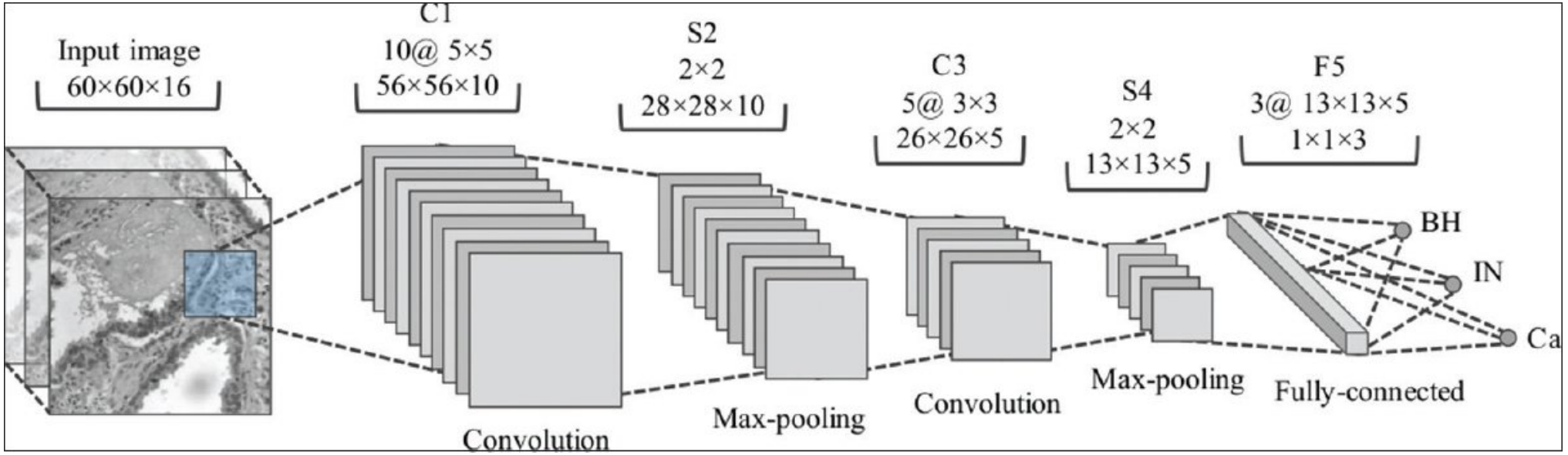
Liczne słabe klasyfikatory



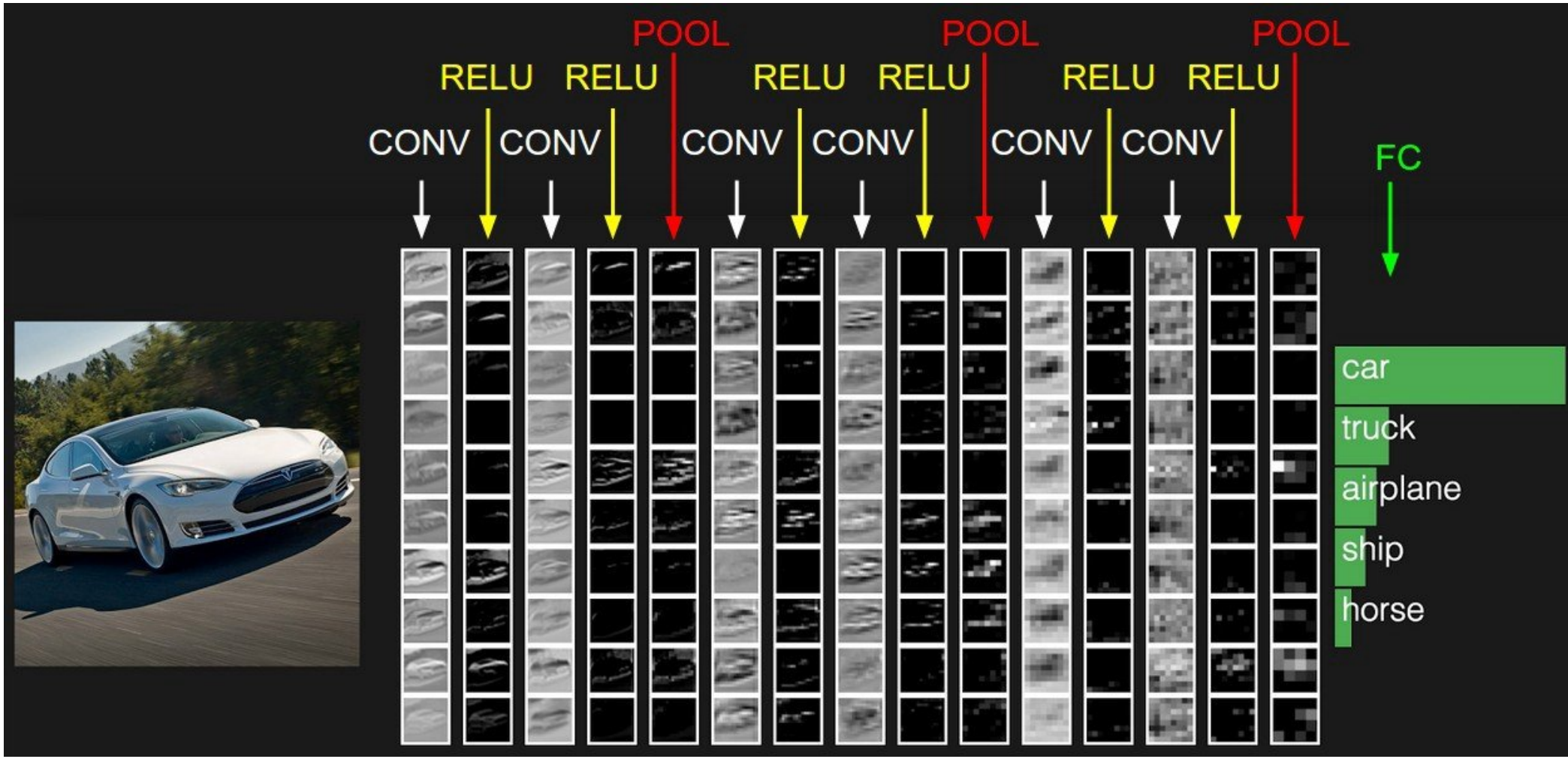




Sieci konwolucyjne



Ekstrakcja cech w sieciach konwolucyjnych




LEARNING
MODEL
REINFORCEMENT
DOPAMINE
COMPUTATIONAL
NEGATIVE
DECISION
CONTROL
PATIENTS
BASAL
PARTICIPANTS
MAKING
WORKING
FRONTAL
DOPAMINERGIC
INDIVIDUALS
LEVELS
NETWORK
NEUROSCIENCE
SUBJECTS
STUDIES
RESULTS
IMPAIRED
CONTROLS
NOGO
BASIS
PFC
GENETIC
ACTIVITY
FRAMEWORK
CONDITIONS
HOWEVER
BG
DA
BEHAVIOR
PROBABILISTIC
SYMPTOMS
REWARD
MECHANISMS
NEUROCOMPUTATIONAL
SCHIZOPHRENIA
PARKINSON'S
REPRESENTATIONS
INFORMATION
FEEDBACK
TIME
CIRCUITS
PROVIDE
UNDERSTANDING
BETTER
GO
STIMULUS
FUNCTION
TASKS
VALUE
USING
TEST
DEFICITS
CORTEX
COGNITIVE
SYSTEM
EFFECTS
BEHAVIORAL
ACTIONS
ERRORS
D2
STN
WM
DIFFERENCES
LEVEL
TWO
AVOIDANCE
CHOICES
MEMORY
STRIATAL
PREFRONTAL
PERFORMANCE
GANGLIA
POSITIVE
MOTOR
ROLE
TRAINING
ONE
THOUGHT
PREDICTIONS
FUNCTIONS
NOVEL
PREDICTED
RESPONSES
ACTION
STRIATUM
TASK
PROCESSES
PUNISHMENT
RESPONSE
OUTCOMES
CONSISTENT
THETA
VIA
REVERSAL
ABILITY
RELATIVE
UNDERLYING
BRAIN
EVIDENCE
SZ
LEARN
STIMULI
DIFFERENT
SHOW
DATA
CONFLICT
NEURAL
SUPPORT
FINDINGS
INCREASED
SUGGEST
INDIVIDUAL
TESTED
CHOICE
STRUCTURE
MECHANISTIC
SYSTEMS
RESULTS
USED
DECISIONS
SELECTION
ERROR
HUMANS
PREDICTION
RELATED
HUMAN
HIGHER
LARGER
SYMPTOM
WHEREAS
INTERACTIONS
PATIENTS
BIAS
SIGNALS
STUDY
HIERARCHICAL
DISEASE
MAY
SINGLE
PHASE
MODELING
ERN
GROUP
WHETHER