

Bootcamp Data Science

Przemysław Spurek

“The sexy job in the next 10 years
will be statisticians.

People think I'm joking, but who would've guessed that
computer engineers would've been the sexy job of the
1990s?”

Hal Varian, the chief economist at Google

Data Scientist: The Sexiest Job of the 21st Century <https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century>

Drown in data



According to Wikipedia, "Data Science is the extraction of knowledge from data".

According to Wikipedia, "Data Science is the extraction of knowledge from data".

The word "theory" itself can mean many things to many people. Generally, in the machine learning community, theorists are computer scientists, mathematicians, and statisticians, who primarily study algorithms that are provably efficient and provably correct, even if they must rely on unrealistically strong assumptions. Theory papers contain proofs of correctness, proofs of convergence, and guarantees on performance.

[http://www.kdnuggets.com/2015/05/
data-science-machine-learning-scientist-definition-jargon.
html](http://www.kdnuggets.com/2015/05/data-science-machine-learning-scientist-definition-jargon.html)

Toy example

Załóżmy że chcemy nauczyć się rozpoznawać piwo i wino.



Toy example

Musimy zbudować model.



Model → Beer

Toy example

Będziemy potrzebować danych: musimy wybrać parametry do naszego eksperymentu.



COLOR

13.5% Alc/volume

ALCOHOL

Toy example

Będziemy potrzebować danych: musimy zebrać odpowiednie dane.



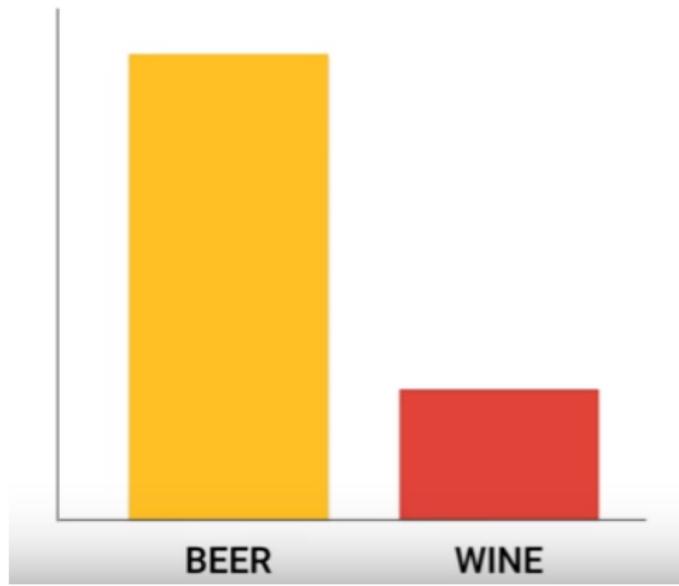
Toy example

Będziemy potrzebować danych: musimy umieścić je w bazie danych.

Color (nm)	Alcohol %	Beer or Wine?
610	5	Beer
599	13	Wine
693	14	Wine

Toy example

Wizualizacja danych.



Toy example

Dzielimy zbiór na część treningową i testową.



Training



Evaluation

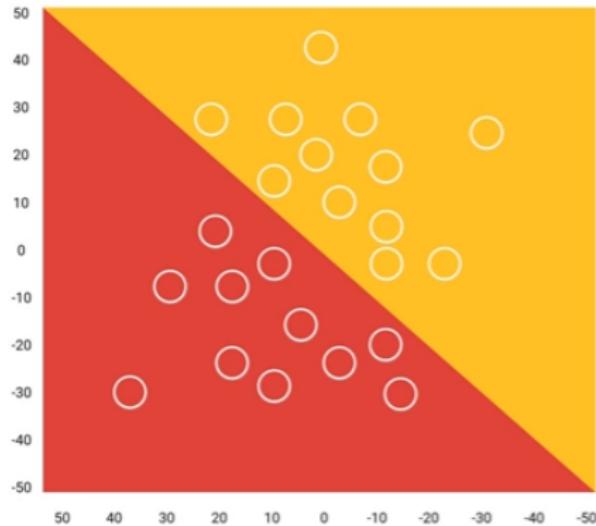
Toy example

Musimy wybrać model.



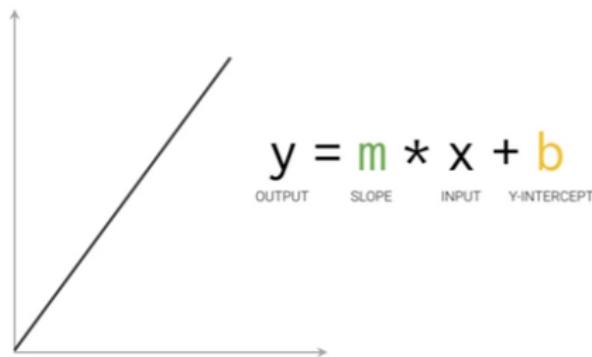
Toy example

Wybierzmy klasyfikator liniowy.



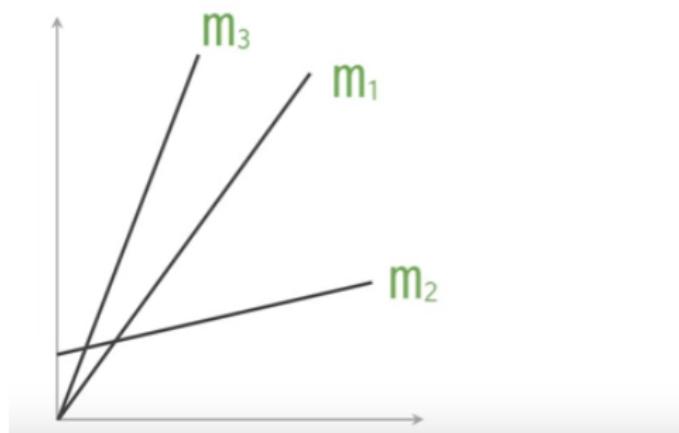
Toy example

Wybierzmy klasyfikator liniowy.



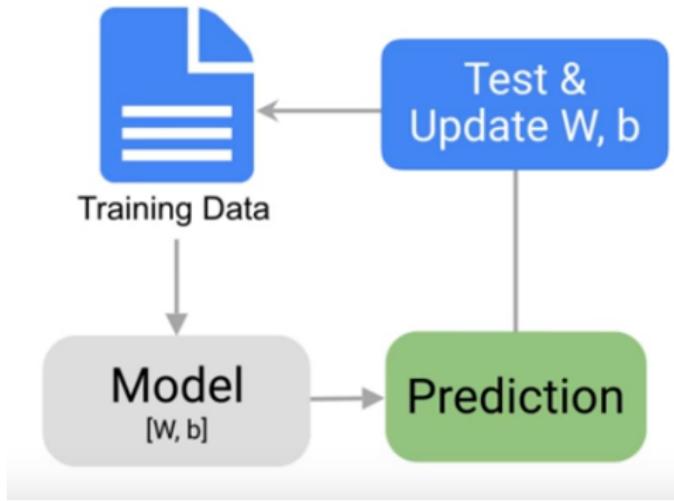
Toy example

Wybierzmy klasyfikator liniowy.



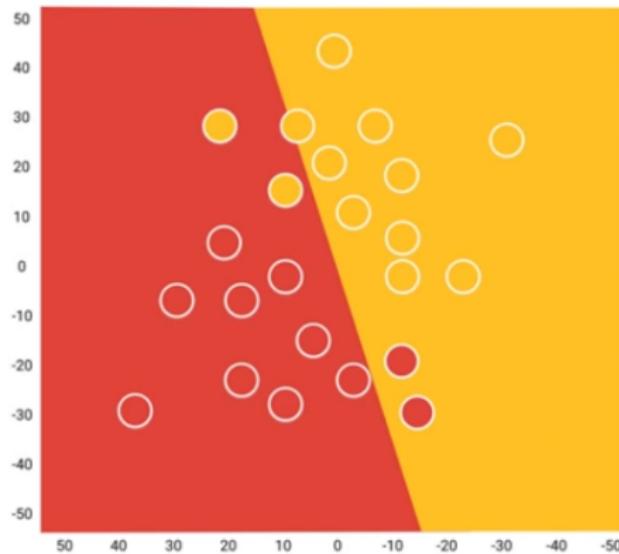
Toy example

Trenujemy nasz model.



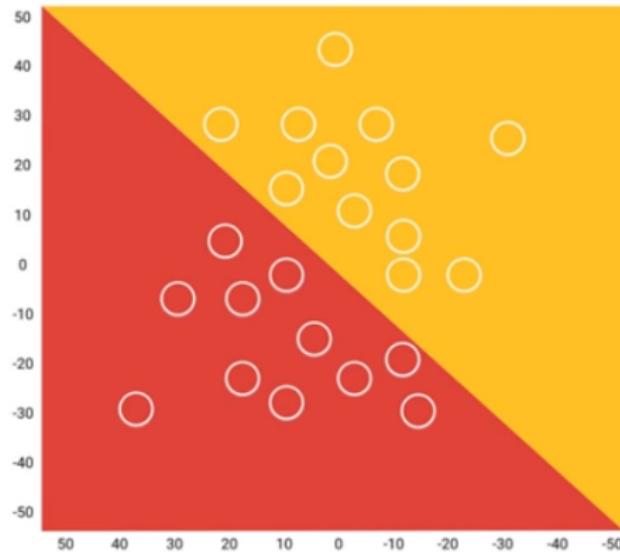
Toy example

Trenujemy nasz model.



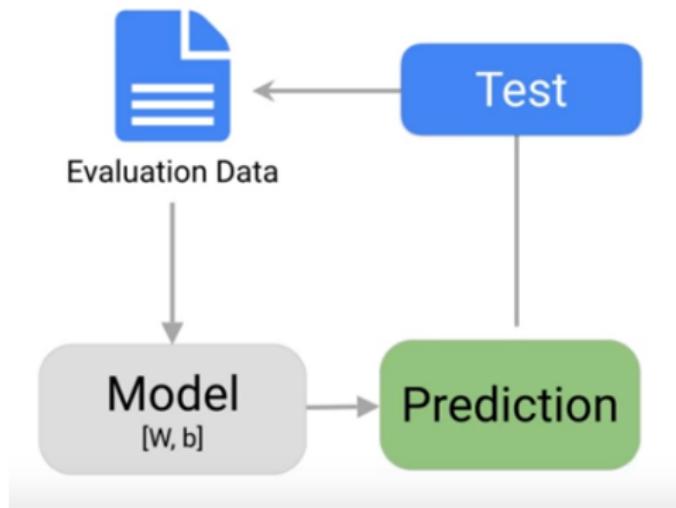
Toy example

Trenujemy nasz model.



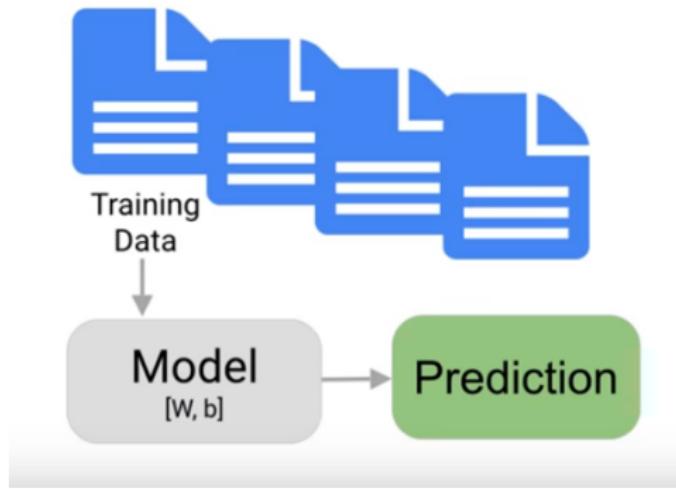
Toy example

Evaluacja modelu.



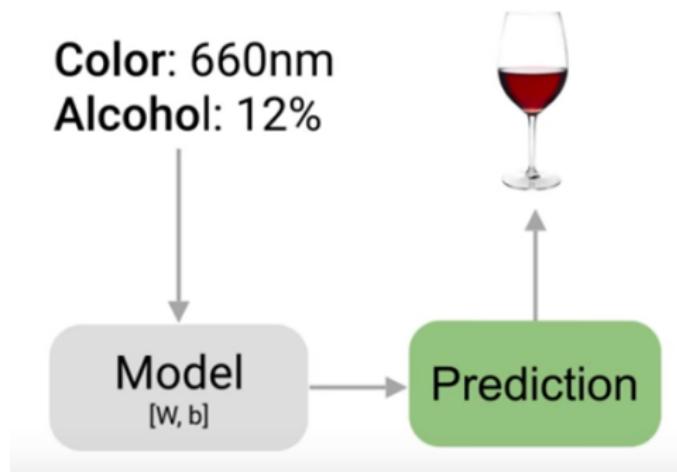
Toy example

Dobieramy najlepsze parametry.



Toy example

Głównym celem była predykcja.



7 Steps of Machine Learning

- Gathering Data
- Preparing that Data
- Choosing a Model
- Training
- Evaluation
- Hyperparameter Tuning
- Prediction

Źródło: <https://medium.com/towards-data-science/the-7-steps-of-machine-learning-2877d7e5548e>

Czym jest uczenie maszynowe?

Jak wybrać feature vectors aby nauczyć program rozróżniać pomarańcze od jabłek?



Czym jest uczenie maszynowe?

A w przypadku obrazów w odcieniach szarości?



Czym jest uczenie maszynowe?

A co jeżeli musimy rozróżnić jeszcze mango?



Czym jest uczenie maszynowe?

Rozważmy inny przykład?



Czym jest uczenie maszynowe?

Naszym celem jest tak skonstruować algorytm uczenia maszynowego by sam dobierał wektory cech.



Czym jest uczenie maszynowe?

Uczenie maszynowe to uczenie z danych

Czym jest uczenie maszynowe?

Uczenie maszynowe to uczenie z danych

- mamy pacjenta w szpitalu po ataku serca; czy, bazując na danych demograficznych i klinicznych, będzie miał następny?

Czym jest uczenie maszynowe?

Uczenie maszynowe to uczenie z danych

- mamy pacjenta w szpitalu po ataku serca; czy, bazując na danych demograficznych i klinicznych, będzie miał następny?
- jaką będzie cena akcji za miesiąc bazując na wynikach firmy i ogólnych trendach ekonomicznych?

Czym jest uczenie maszynowe?

Uczenie maszynowe to uczenie z danych

- mamy pacjenta w szpitalu po ataku serca; czy, bazując na danych demograficznych i klinicznych, będzie miał następny?
- jaką będzie cena akcji za miesiąc bazując na wynikach firmy i ogólnych trendach ekonomicznych?
- rozpoznaj, pisane ręcznie, cyfry w kodach pocztowych,

Czym jest uczenie maszynowe?

Uczenie maszynowe to uczenie z danych

- mamy pacjenta w szpitalu po ataku serca; czy, bazując na danych demograficznych i klinicznych, będzie miał następny?
- jaką będzie cena akcji za miesiąc bazując na wynikach firmy i ogólnych trendach ekonomicznych?
- rozpoznaj, pisane ręcznie, cyfry w kodach pocztowych,
- jakie są zakażenia bakteryjne we krwi pacjenta na podstawie najprostszych testów?

Czym jest uczenie maszynowe?

Uczenie maszynowe to uczenie z danych

- mamy pacjenta w szpitalu po ataku serca; czy, bazując na danych demograficznych i klinicznych, będzie miał następny?
- jaką będzie cena akcji za miesiąc bazując na wynikach firmy i ogólnych trendach ekonomicznych?
- rozpoznaj, pisane ręcznie, cyfry w kodach pocztowych,
- jakie są zakażenia bakteryjne we krwi pacjenta na podstawie najprostszych testów?
- wykryj współczynniki ryzyka dla raka prostaty,

Czym jest uczenie maszynowe?

Uczenie maszynowe to uczenie z danych

- mamy pacjenta w szpitalu po ataku serca; czy, bazując na danych demograficznych i klinicznych, będzie miał następny?
- jaką będzie cena akcji za miesiąc bazując na wynikach firmy i ogólnych trendach ekonomicznych?
- rozpoznaj, pisane ręcznie, cyfry w kodach pocztowych,
- jakie są zakażenia bakteryjne we krwi pacjenta na podstawie najprostszych testów?
- wykryj współczynniki ryzyka dla raka prostaty,
- jaki jest współczynnik postępu i niebezpieczeństwa nawrotu raka wątroby na podstawie badań patologicznych fragmentów organu wyciętych podczas operacji,

Czym jest uczenie maszynowe?

Uczenie maszynowe to uczenie z danych

- mamy pacjenta w szpitalu po ataku serca; czy, bazując na danych demograficznych i klinicznych, będzie miał następny?
- jaką będzie cena akcji za miesiąc bazując na wynikach firmy i ogólnych trendach ekonomicznych?
- rozpoznaj, pisane ręcznie, cyfry w kodach pocztowych,
- jakie są zakażenia bakteryjne we krwi pacjenta na podstawie najprostszych testów?
- wykryj współczynniki ryzyka dla raka prostaty,
- jaki jest współczynnik postępu i niebezpieczeństwa nawrotu raka wątroby na podstawie badań patologicznych fragmentów organu wyciętych podczas operacji,
- etc.

Real-time event detection:

- <https://www.youtube.com/watch?v=QcCjmWwEUgg>
- <https://www.youtube.com/watch?v=-1errutWwLY>

Self Driving Cars:

- <https://www.youtube.com/watch?v=lL16AQItG1g>
- <https://www.youtube.com/watch?v=DjAJnQoNdMA>

Real-time face recognition

- <https://www.youtube.com/watch?v=B4m2RVFLbME>
- <https://www.youtube.com/watch?v=88HdqNDQsEk>

Machine learning OCR

- <https://www.youtube.com/watch?v=Rq6t-002WVA>

Go

- <https://www.youtube.com/watch?v=g-dKX0lsf98>
- https://www.youtube.com/watch?v=KsbQ_HNX6Pg

Deep Learning

- <https://www.youtube.com/watch?v=Bui3DWs02h4>

Uczenie maszynowe czy statystyczne

Uczenie maszynowe czy statystyczne

Nie udawajmy, że uczenie maszynowe jest wynalazkiem informatyków i powstało w ostatnich latach. Znacznie wcześniej zostało wykryte w statystyce.

Uczenie maszynowe czy statystyczne

Nie udawajmy, że uczenie maszynowe jest wynalazkiem informatyków i powstało w ostatnich latach. Znacznie wcześniej zostało wykryte w statystyce.

Jakie są różnice?

Uczenie maszynowe czy statystyczne

Nie udawajmy, że uczenie maszynowe jest wynalazkiem informatyków i powstało w ostatnich latach. Znacznie wcześniej zostało wykryte w statystyce.

Jakie są różnice?

Żadne jedno i drugie starają się odpowiedzieć na pytanie jak i czego można się nauczyć z danych?

Uczenie maszynowe czy statystyczne

Nie udawajmy, że uczenie maszynowe jest wynalazkiem informatyków i powstało w ostatnich latach. Znacznie wcześniej zostało wykryte w statystyce.

Jakie są różnice?

Żadne jedno i drugie starają się odpowiedzieć na pytanie jak i czego można się nauczyć z danych?

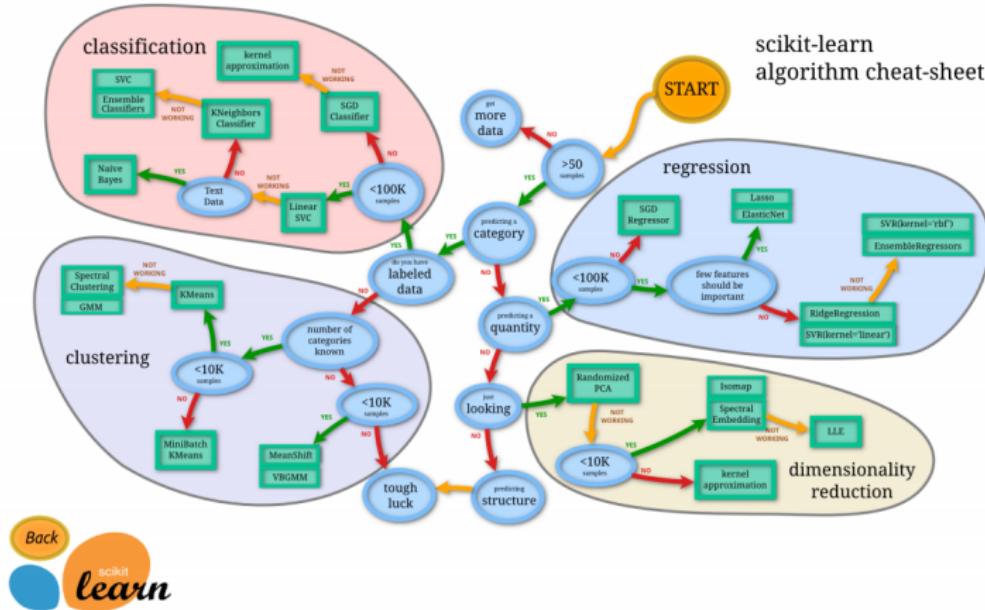
Statystyka:

przedziały ufności, testy hipotez, optymalne estymatory w niskich wymiarach, analiza przeżycia, szeregi czasowe, wielokrotne testowanie

Nauczanie maszynowe:

predykcja w wysokich wymiarach, uczenie online, semi-nadzorowane, rozmaitości, aktywne, boosting.

Uczenie maszynowe

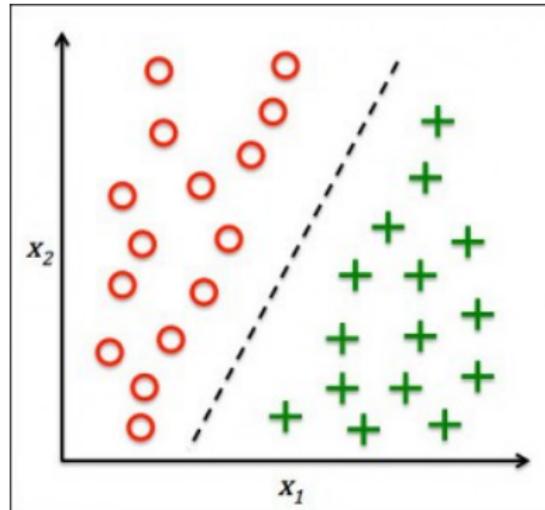


No free lunch theorem

Wszystkie modele są niepoprawne, chociaż niektóre z nich są bardziej przydatne.

- nie ma najlepszego modelu
- dla każdego zadania odpowiedni może być inny model
- dla każdego modelu potrzebujemy wiele algorytmów uczących

Klasyfikacja

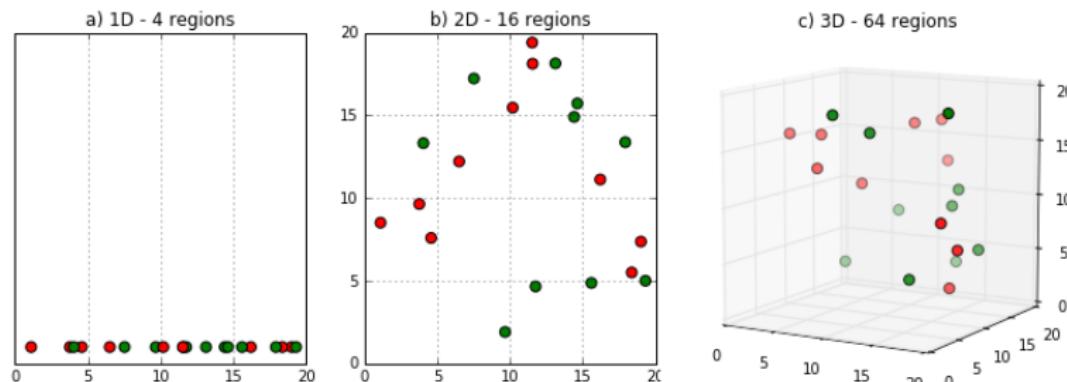


https://github.com/przem85/bootcamp/blob/master/introduction/I_1.ipynb

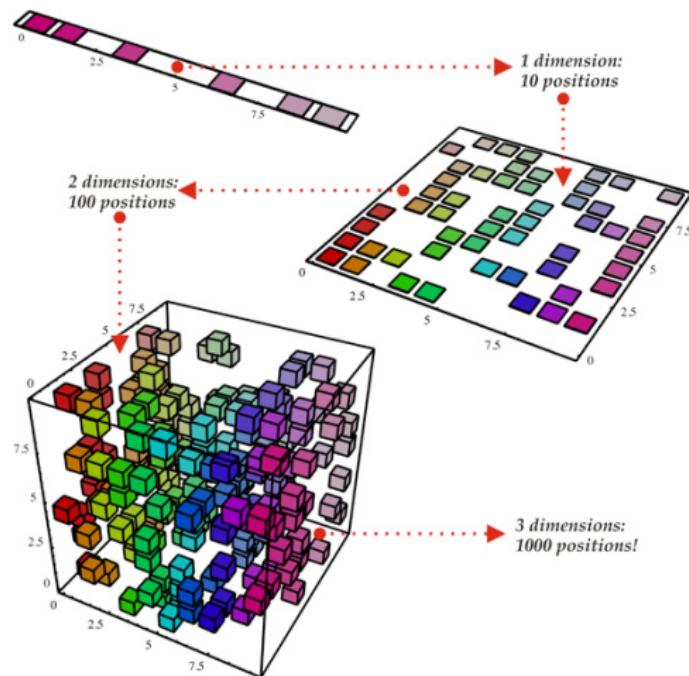
KŁĄTWA (PRZEKLEŃSTWO) WYMIAROWOŚCI

Curse of Dimensionality

W miarę wzrostu liczby wymiarów (zmiennych) liczba obiektów (obserwacji) potrzebnych do wiarygodnego oszacowania parametrów lub funkcji rośnie wykładniczo.



KŁĄTWA (PRZEKLEŃSTWO) WYMIAROWOŚCI

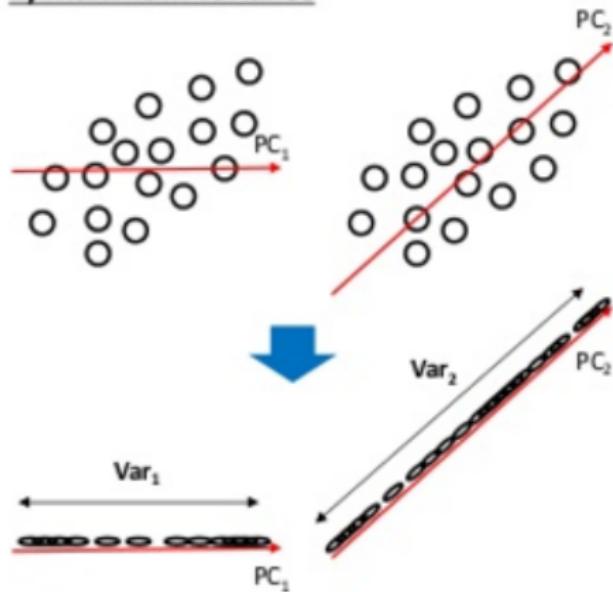


https://github.com/przem85/bootcamp/blob/master/introduction/I_5.ipynb

Redukcja wymiarowości

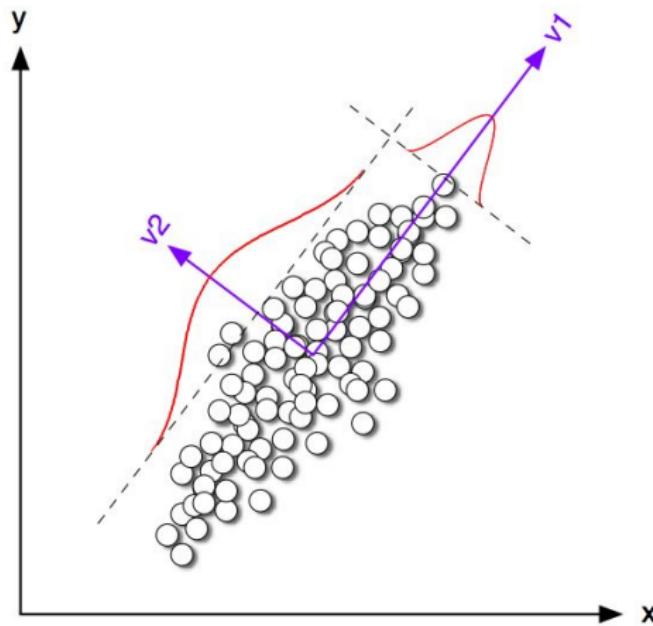
Principal Component Analysis

1) Maximum variance

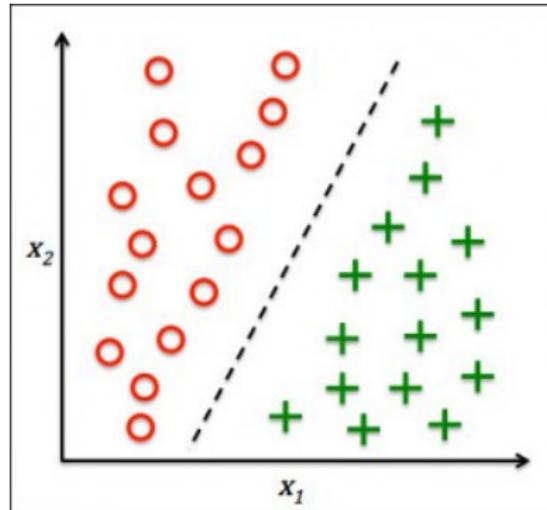


Redukcja wymiarowości

Principal Component Analysis

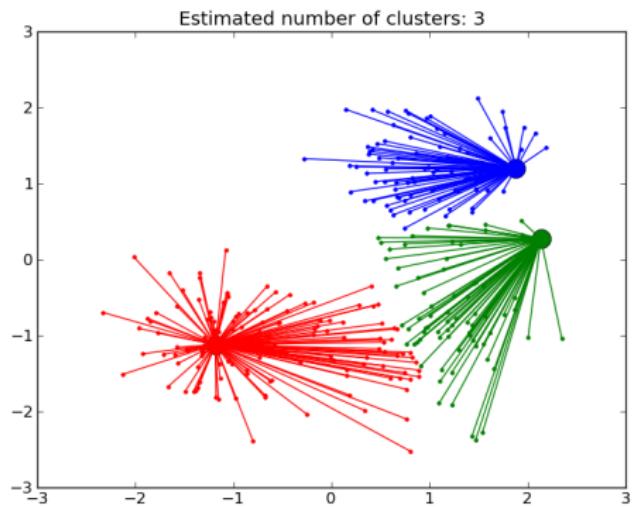


Klasyfikacja



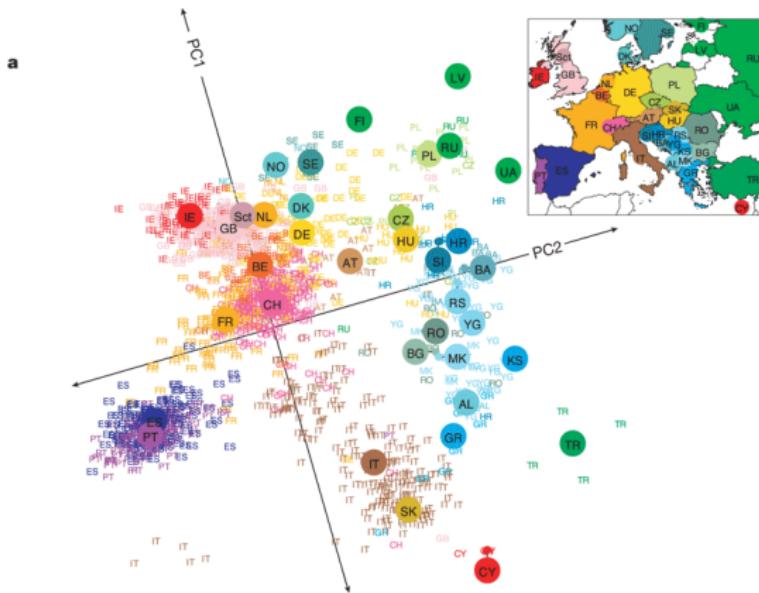
https://github.com/przem85/bootcamp/blob/master/introduction/I_2.ipynb

Klastrowanie

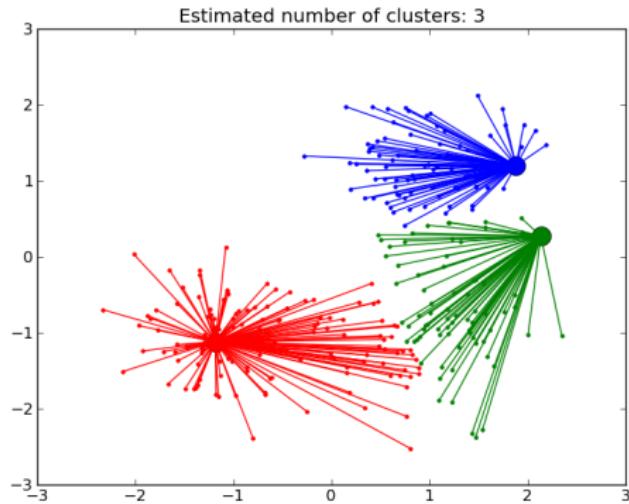


Klastrowanie

Population structure within Europe

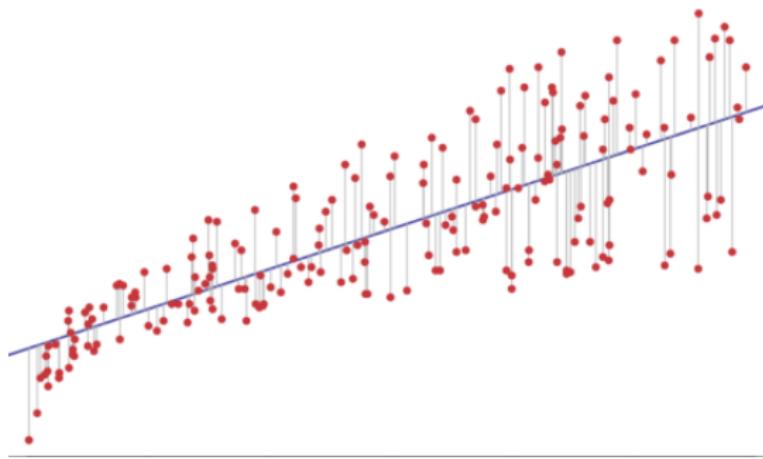


Klastrowanie



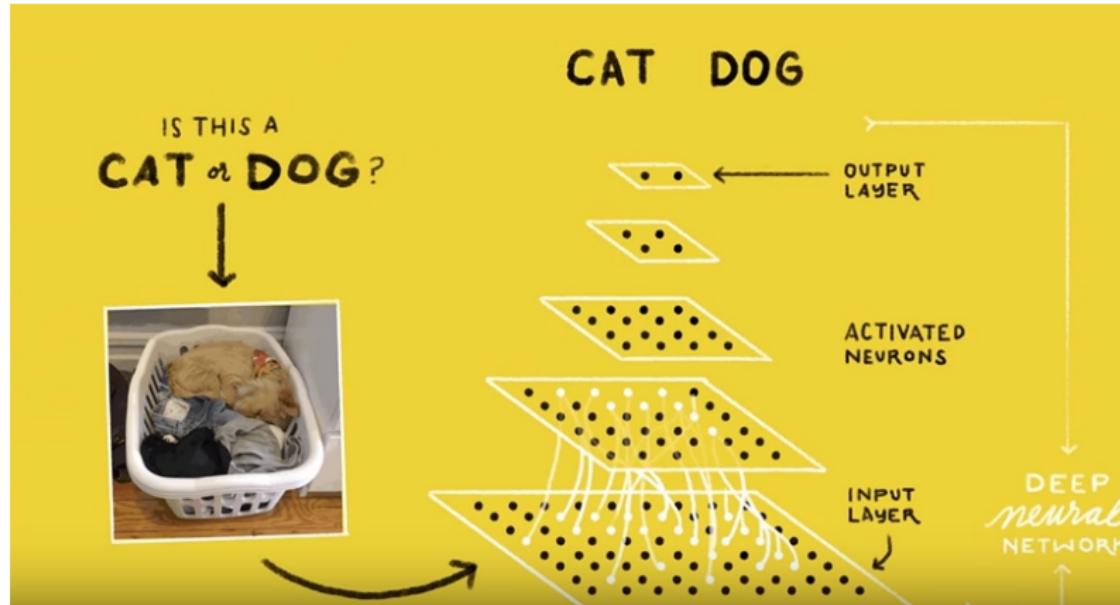
https://github.com/przem85/bootcamp/blob/master/introduction/I_3.ipynb

Regresja



https://github.com/przem85/bootcamp/blob/master/introduction/I_4.ipynb

Deep Learning



Deep Learning

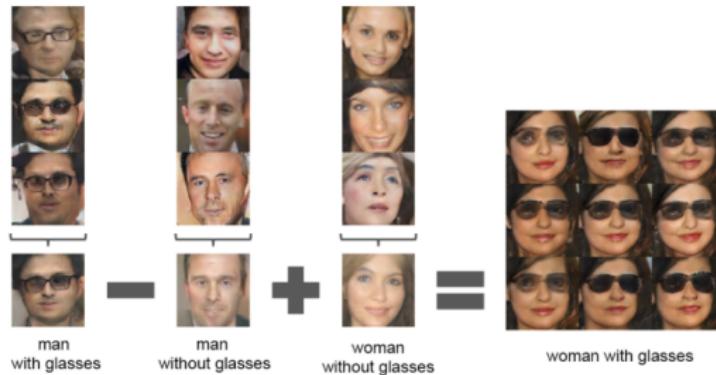


<https://www.youtube.com/watch?v=-R9bJGNH1tQ>

<https://www.youtube.com/watch?v=Uxax5EKg0zA>

word2vec / GloVe

$$v(king) - v(man) + v(woman) \approx v(queen) :$$



https://www.youtube.com/watch?v=xMwx2A_o5r4

https://www.youtube.com/watch?v=BD8wPsr_DAI&t=4s

Prawdziwa praca Analityka Danych

Prawdziwa praca Analityka Danych

Our goal is to identification of compounds acting on two biological receptors 5-HT_{1a} and 5-HT₆, the proteins responsible for the regulation of central nervous system.

Prawdziwa praca Analityka Danych

Our goal is to identification of compounds acting on two biological receptors 5-HT_{1a} and 5-HT₆, the proteins responsible for the regulation of central nervous system.

Compounds classified by a learning system as active in virtual screening process are usually further examined and the most promising ones could be used in drug designing.

The activity level is measured by a positive real valued number K_i :

- if $K_i \leq 100$, then a compound is active,
- if $K_i > 1000$, then a compound is inactive
- if $100 < K_i \leq 1000$, then a compound is not classified to any of these groups and they are usually eliminated from a training stage.

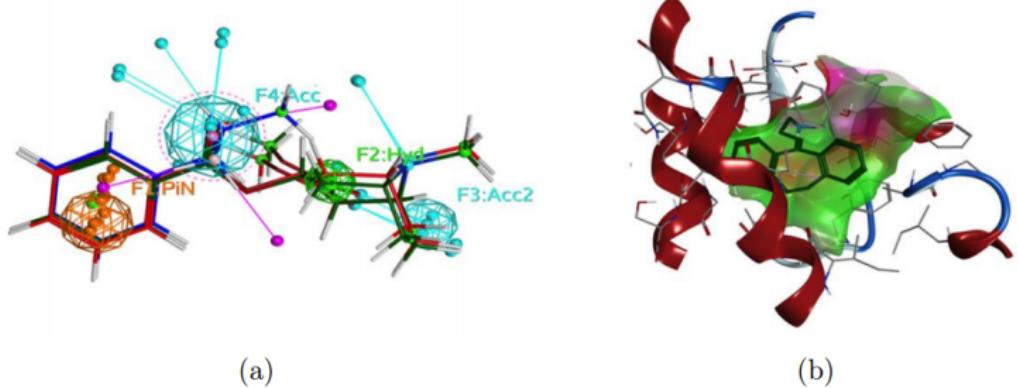


Figure 7: a) Ligand-based drug design focuses on specific properties of a molecule, e.g. employing a pharmacophore. b) Structure-based drug design utilizes the ligand-binding pocket amino acid side chains of the target receptor. Figure from [18].

Most of the ML models require input data represented in a subset of \mathbb{R}^d space.

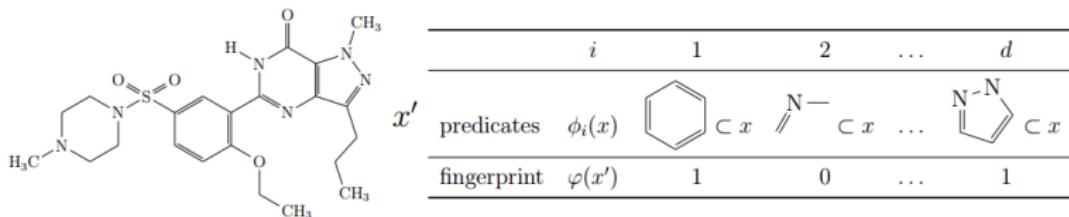


Figure 8: Sample fingerprint of the chemical molecule x' . $|A|_x$ denotes the number of atoms or substructures A in x , so in particular $A \subset x \iff |A|_x \geq 1$.

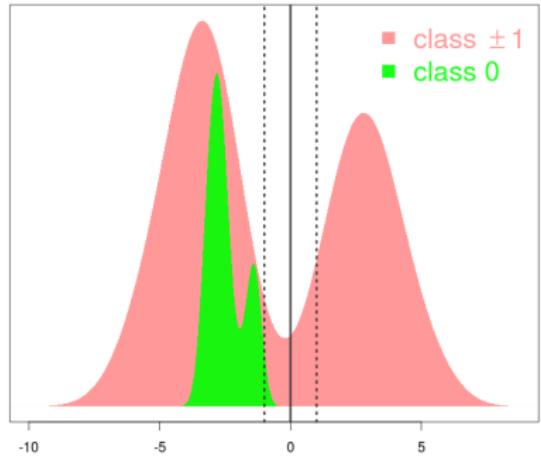
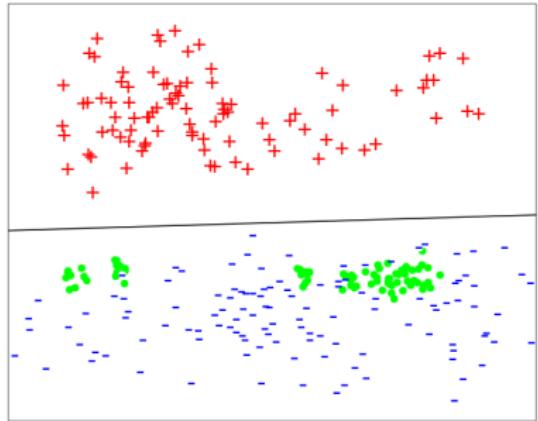
<https://www.youtube.com/watch?v=TTtrkOUe-Cg>

One of the machine learning paradigms states that one should take into account all existing information in building a learning framework.

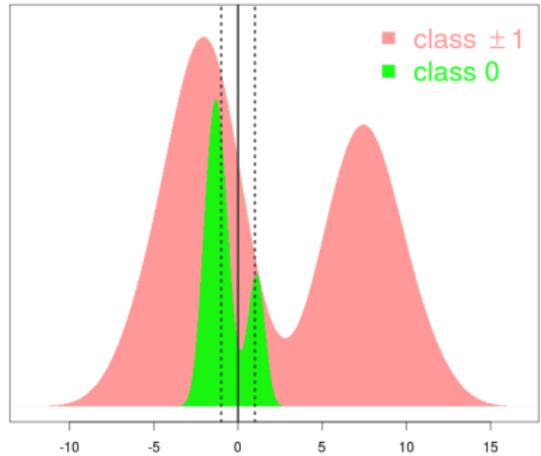
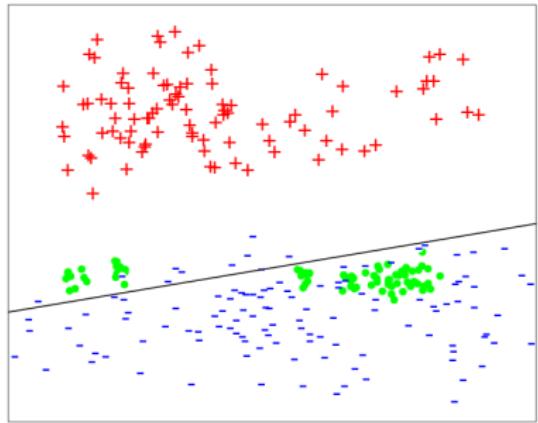
One of the machine learning paradigms states that one should take into account all existing information in building a learning framework.

Main problem of our work

Our goal is to construct a model which uses 3 classes.

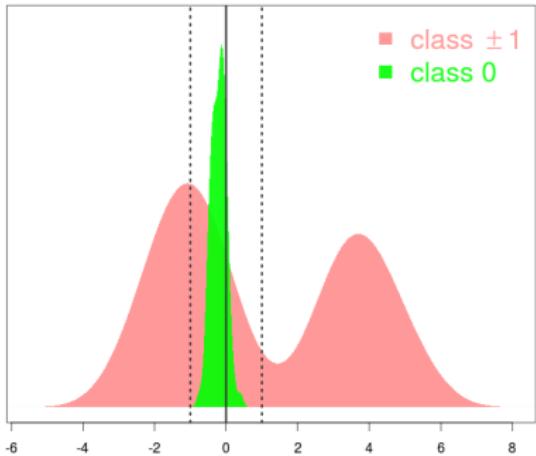
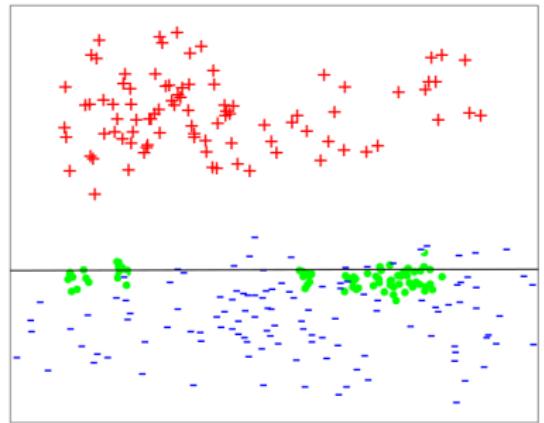


Rysunek: SVM



Rysunek: S3VM

Our approach



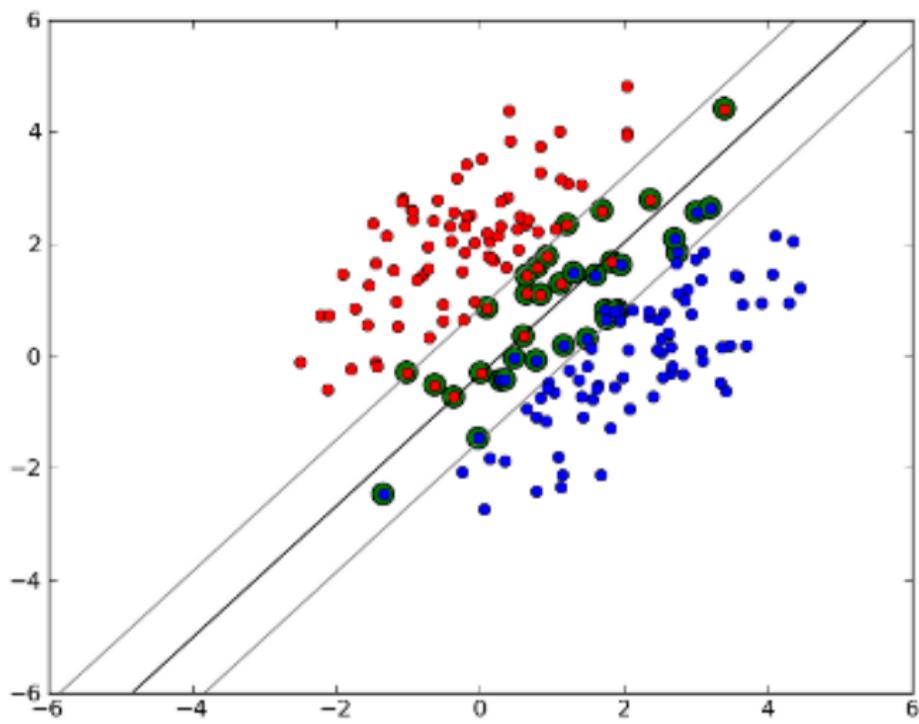
Rysunek: SVM_0

Let us recall that SVM [Chapter 2.3](Large-scale kernel machines, Bottou and all, 2007) aims at finding such an affine function $x \rightarrow v^T x + b$ which minimizes the cost function given by

$$\begin{aligned} \text{SVM}(v, b) = & \frac{1}{2} \|v\|^2 + C \sum_{y_i=-1} \max(0, 1 + (v^T x_i + b)) \\ & + C \sum_{y_i=1} \max(0, 1 - (v^T x_i + b)), \end{aligned} \tag{1}$$

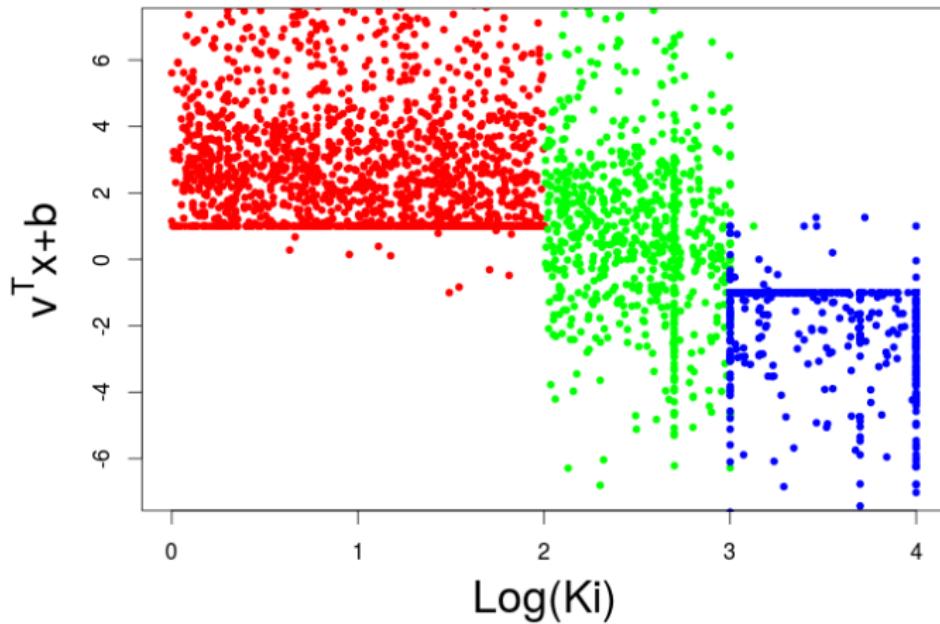
where $X = (x_i)_i$ is a dataset and $y_i = \pm 1$ denotes the class membership of x_i . The first term $\frac{1}{2} \|v\|^2$ plays the regularization role, while the expression $\max(0, 1 - y_i(v^T x_i + b))$ measures a distance of the point $v^T x_i + b$ from the set $[1, +\infty)$, for $y_i = +1$ (or from $(-\infty, -1]$, for $y_i = -1$).

Our approach

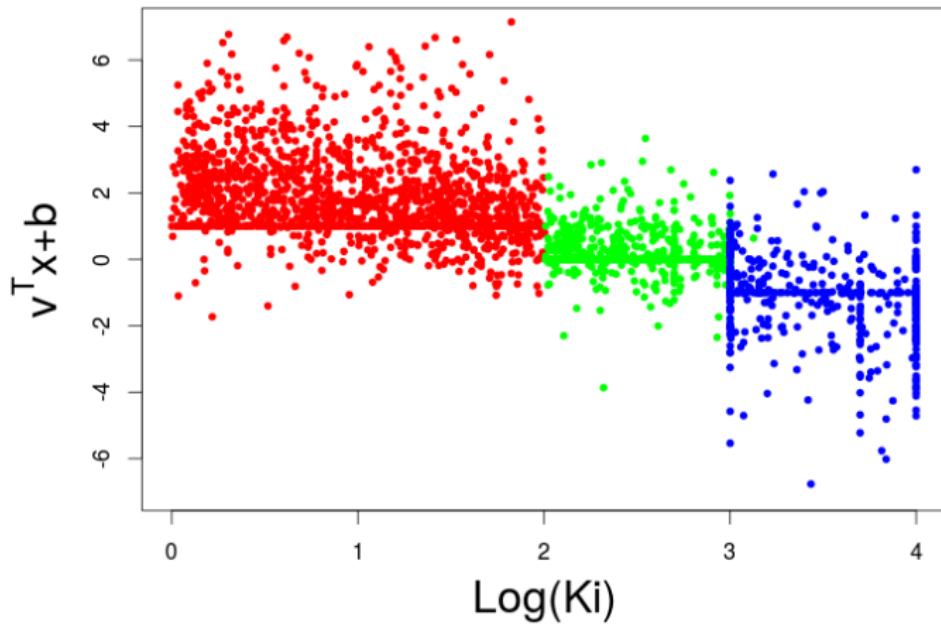


Our model is a small modification

$$SVM_0(v, b) = \text{SVM}(v, b) + C \sum_{i:y_i=0} |v^T x_i + b|.$$



Rysunek: SVM



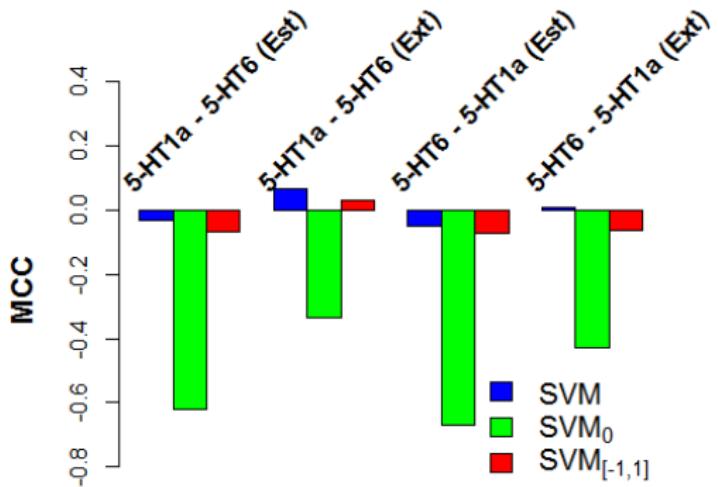
Rysunek: svmo

Detection of active compounds

Tablica: MCC scores reported on test sets for binary classification task.

Dataset	SVM	S3VM	SVM_0
Heart Disease	0.75 ± 0.02	0.80 ± 0.01	0.78 ± 0.02
Housing	0.86 ± 0.02	0.83 ± 0.02	0.87 ± 0.01
<hr/>			
5-HT1a (Ext)	0.59 ± 0.02	0.59 ± 0.01	0.58 ± 0.01
5-HT6 (Ext)	0.77 ± 0.02	0.74 ± 0.02	0.75 ± 0.01

Chemical space exploration



Rysunek: MCC scores of classification – the classifier was trained on one receptor and tested on the other.

Sprawy organizacyjne

Czym jest bootcamp:

Sprawy organizacyjne

Czym jest bootcamp:

- charakterystyka pracy,

Sprawy organizacyjne

Czym jest bootcamp:

- charakterystyka pracy,
- jak skorzystać w największym stopniu z tego kursu

Sprawy organizacyjne

Czym jest bootcamp:

- charakterystyka pracy,
- jak skorzystać w największym stopniu z tego kursu
- jak będzie wyglądały poszczególne zajęcia i jak się do nich przygotować

Sprawy organizacyjne

Program bootcampu:

Sprawy organizacyjne

Program bootcampu:

- Python - podstawy

Sprawy organizacyjne

Program bootcampu:

- Python - podstawy
- Python zaawansowany

Sprawy organizacyjne

Program bootcampu:

- Python - podstawy
- Python zaawansowany
- Analiza danych

Sprawy organizacyjne

Program bootcampu:

- Python - podstawy
- Python zaawansowany
- Analiza danych
- Podstawy statystyki i modele statystyczne

Sprawy organizacyjne

Program bootcampu:

- Python - podstawy
- Python zaawansowany
- Analiza danych
- Podstawy statystyki i modele statystyczne
- Modelowanie danych

Sprawy organizacyjne

Program bootcampu:

- Python - podstawy
- Python zaawansowany
- Analiza danych
- Podstawy statystyki i modele statystyczne
- Modelowanie danych
- Podstawy i metody uczenia maszynowego

Sprawy organizacyjne

Program bootcampu:

- Python - podstawy
- Python zaawansowany
- Analiza danych
- Podstawy statystyki i modele statystyczne
- Modelowanie danych
- Podstawy i metody uczenia maszynowego
- Wprowadzenie do Deep Learning

Program bootcampu:

- Python - podstawy
- Python zaawansowany
- Analiza danych
- Podstawy statystyki i modele statystyczne
- Modelowanie danych
- Podstawy i metody uczenia maszynowego
- Wprowadzenie do Deep Learning
- Apache Spark i Big Data

Sprawy organizacyjne

Program bootcampu:

- Python - podstawy
- Python zaawansowany
- Analiza danych
- Podstawy statystyki i modele statystyczne
- Modelowanie danych
- Podstawy i metody uczenia maszynowego
- Wprowadzenie do Deep Learning
- Apache Spark i Big Data
- Projekt finalny i szkolenie z rekruterem