

Statystyka cz.2 i model regresji liniowej

Michał Więtczak

Doświadczenie w projektach komercyjnych dla firm o globalnych zasięgu z uczenia maszynowego i uczenia głębokiego

Szkolenia z zakresu Data Science, Deep learning, Machine Learning, Natural Language Processing, Computer Vision, Python

Testowanie hipotez

Type of dependent variable	Type of independent variable						
	Ordinal/categorical				Normal/interval (ordinal)	More than 1	None
	Two groups		More groups				
	Paired	Unpaired	Paired	Unpaired			
2 categories	McNemar Test, Sign-Test	Fisher Test, Chi-squared-Test	Cochran's Q-Test	Fisher Test, Chi-squared-test	(Conditional) Logistic Regression	Logistic Regression	Chi-squared-Test
Nominal	Bowker Test	Fisher Test, Chi-squared-Test		Fisher Test, Chi-squared-test	Multinomial logistic regression	Multinomial logistic regression	Binomial Test
Ordinal	Wilcoxon Test, Sign-Test	Wilcoxon-Mann-Whitney Test	Friedman-Test	Kruskal-Wallis Test	Spearman-rank-test	Ordered logit	Median Test
Interval	Wilcoxon Test, Sign-Test	Wilcoxon-Mann-Whitney Test	Friedman-Test	Kruskal-Wallis Test	Spearman-rank test	Multivariate linear model	Median Test
Normal	t-Test (for paired)	t-Test (for unpaired)	Linear Model (ANOVA)	Linear Model (ANOVA)	Pearson-Correlation-test	Multivariate Linear Model	t-Test
Censored Interval	Log-Rank Test		Survival Analysis, Cox proportional hazards regression				
None	Clustering, factor analysis, PCA, canonical correlation						

Test Chi-kwadrat zgodności

Test zgodności chi-kwadrat (inaczej zwany testem Pearsona) służy do porównania ze sobą zaobserwowanego rozkładu naszej zmiennej z jakimś teoretycznym rozkładem.

Test zgodności chi-kwadrat w praktyce można wykorzystać przynajmniej na dwa sposoby

1. sprawdzenie równoliczności grup
2. porównanie występowania obserwacji z ich teoretycznym występowaniem

1) Równoliczność grup

Przykład:

Badacz chciał sprawdzić, czy w swoim badaniu była równa liczba kobiet i mężczyzn (statystycznie równa, nieistotne statystycznie różnice). W badaniu przebadał 480 mężczyzn oraz 520 kobiet. Wynik okazał się nieistotny statystycznie (dla $p > 0,05$). Oznacza to, że badacz może przyjąć, że przebadał podobną liczbę kobiet i mężczyzn (mówiąc językiem statystyki).

Test chi-kwadrat zgodności

Test ten stosuje się również w przypadku sprawdzania, czy któraś z udzielanych odpowiedzi była najczęściej udzielana

Przykład:

Badacz zadał pytanie respondentom czy bardziej im smakuje napój A czy napój B. 36 osób badanych udzieliło odpowiedzi A, a 64 osoby udzieliło odpowiedzi B.

Badacz założył, że gdyby napoje nie różniły się preferencją to powinien uzyskać podobne wyniki w obu grupach, po 50 osób. Przeprowadził test zgodności chi-kwadrat i (dla poziomu $p < 0,05$) ocenił, że rozkład udzielanych odpowiedzi nie jest równy, przeważa preferencja napoju B.

Test chi - kwadrat

Za pomocą testu niezależności χ^2 (chi-kwadrat) można sprawdzić czy pomiędzy dwiema cechami jakościowymi występuje zależność. Układ hipotez jest następujący:

- H_0 : zmienne są niezależne,
- H_1 : zmienne nie są niezależne.

W programie R test niezależności można wywołać za pomocą funkcji `chisq.test()` z pakietu `stats`. Jako argument tej funkcji należy podać tablicę kontyngencji. W przypadku operowania na danych jednostkowych można ją utworzyć poprzez funkcję `table()`. Jeżeli wprowadzamy liczebności ręcznie to należy zadbać o to, żeby wprowadzony obiekt był typu `matrix`.

Przykład

Czy pomiędzy zmienną płeć, a zmienną przynależność do związków zawodowych istnieje zależność?

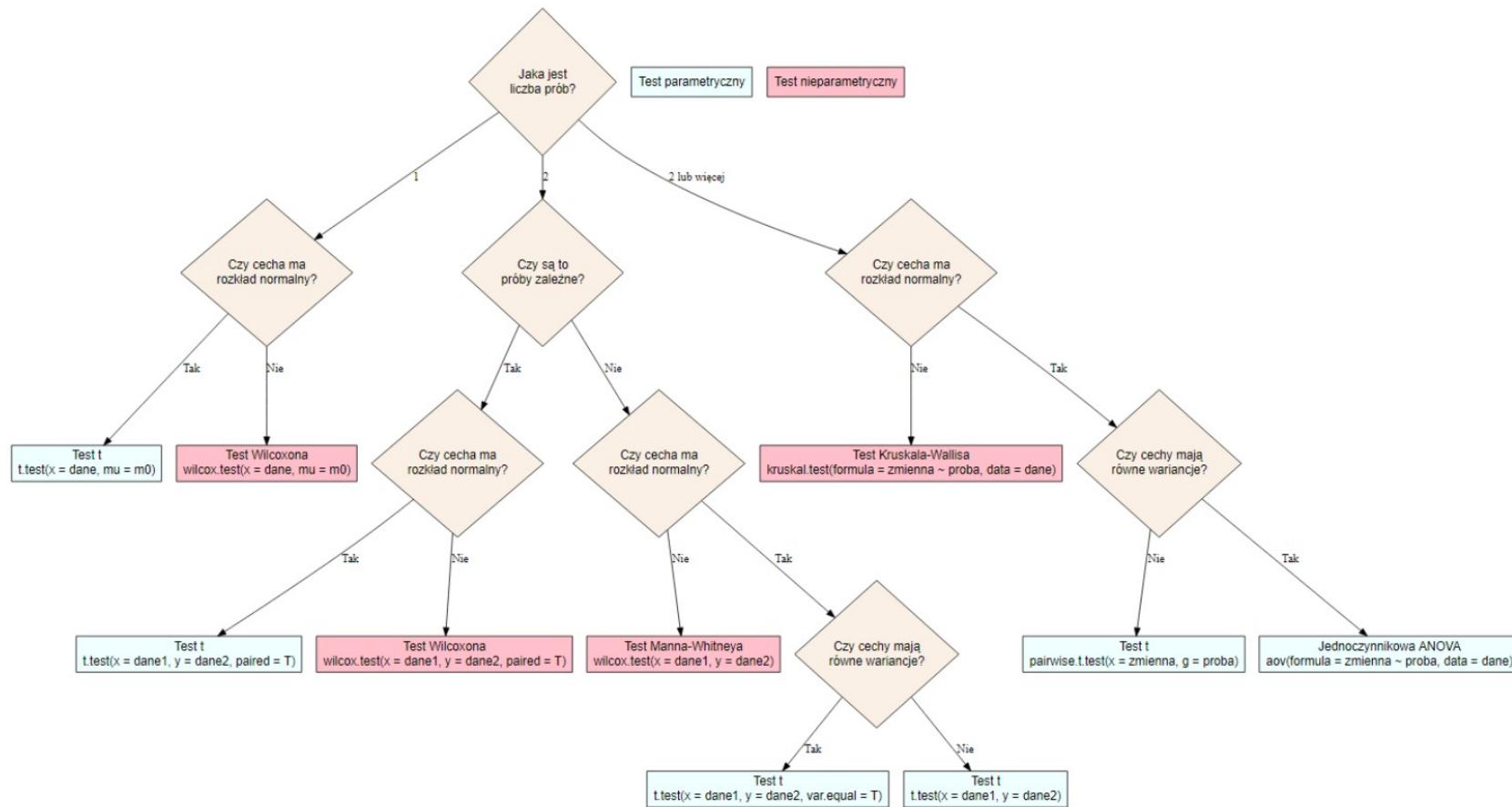
W pierwszym kroku określamy hipotezy badawcze:

H_0 : pomiędzy płcią a przynależnością do związków nie ma zależności

H_1 : pomiędzy płcią a przynależnością do związków jest zależność

oraz przyjmujemy poziom istotności - weźmy standardową wartość $\alpha = 0,05$.

W pierwszej kolejności popatrzmy na tabelę krzyżową (kontyngencji) zawierającą liczebności poszczególnych kombinacji wariantów.



ANOVA - analiza wariancji

Analiza wariancji ANOVA jest jedną z najbardziej popularnych i najczęściej stosowanych analiz statystycznych. Dokładniej - analizą wariancji określa się grupę analiz, służących do badania wpływu czynników (zmiennych niezależnych) na zmienną zależną.

Moglibyśmy podzielić analizę wariancji na trzy grupy analiz:

- **jednoczynnikowa analiza wariancji**
 - *wpływ jednego czynnika międzygrupowego na zmienną zależną*
- **wieloczynnikowa analiza wariancji**
 - *wpływ kilku czynników międzygrupowych na zmienną zależną*
- **analiza wariancji dla czynników wewnętrzgrupowych**
 - *wpływ czynnika wewnętrzgrupowego na zmienną zależną, tzw. "powtarzane pomiary"*

Porównanie testów

Zmienna ilościowa,
rozkład normalny

Test
parametryczny

- test t-Studenta

Zmienna ilościowa,
rozkład inny niż
normalny

Test
nieparametryczny

- test U Mano-
Whitneya

Zmienna
jakościowa

Tabele
kontyngencji

- test chi-kwadrat
Pearsona
- dokładny test
Fishera

Porównanie metod

1. Na początku zbieramy dane i sprawdzamy czy rozkład jest normalny w każdej z grup danych jeżeli mamy ich kilka (np. wyniki finansowe przedsiębiorstwa w każdym miesiącu)
2. Jeżeli rozkład jest normalny to:
 - mamy 1 lub 2 grupy to test t-studenta (1 lub 2 stronny w zależności ile prób (ang samples)) - zmienne są tu zależne (zależą nawzajem od siebie) wówczas porównujemy wariancje
 - jeżeli zmienne są niezależne to porównujemy średnie
 - jeżeli mamy więcej niż 2 grupy to analiza ANOVA
3. Jeżeli zmienne nie mają rozkładu normalnego:
 - a. dla 2 grup:
 - jeżeli zmienne są niezależne to porównujemy mediany ze sobą testem rank Wilcooxona
 - jeżeli zmienne są zależne to porównujemy testem znaków Wilcooxona
 - b. dla więcej niż 2 grup:
 - tzw. nieparametryczna ANOVA (tu porównanie median przy parametrycznej była średnia!!)
4. Jeżeli dane są dyskretne:
 - a. dla 2 grup (zmienne niezależne)
 - test chi-kwadrat licznosci (test który sprawdza licznosć w grupach)
 - test fischera dla bardzo małych grup jeżeli występuje jakas licznośc poniżej 5!
 - b. dla 2 grup (zmienne zależne):
 - test McNammary
 - c. dla więcej niż 2 grup (test Chi-kwadrat) - uwaga to inny niż chi kwadrat licznosci.

Testy parametryczne a nieparametryczne

Testy parametryczne vs nieparametryczne - ogólnie w parametrycznych porównujemy PARAMETRY (parametrami są średnia, mediana, odchylenie wariancja), więc wszystkie te mówiąc krotko które mają rozkład normalny, w nieparametrycznych porównujemy inne kryteria które parametrami nie są jak np. proporcje, rankingi itd.

Proba zależna vs niezależna :

zależna - badanie grupy w różnych warunkach : przed i po ingerencji np. przed i po zastosowaniu leku, badanie zysków przed i po zastosowaniu jakieś kampanii marketingowej, badanie satysfakcji pracowników przed i po wprowadzeniu pracy zdalnej

niezależna - badanie różnych grup lub tej samej nie podanej ingerencji np. badania kobiet i mężczyzn po zastosowaniu leku, analiza zysków w 2 firmach po zastosowaniu kampanii marketingowej.

Statystyka w data science

- Użyteczne pojęcia w data science z zakresu statystyki matematycznej to:
 - Średnia
 - Medianą
 - Odchylenie standardowe i wariancja
 - Kowariancja
 - Korelacja

Średnia i mediana

- Średnia arytmetyczna – suma liczb podzielona przez ich liczbę. Dla liczb jest to więc wyrażenie. W języku potocznym średnią arytmetyczną określa się po prostu jako średnią.
- Mediana:

Aby obliczyć medianę ze zbioru n obserwacji, sortujemy je w kolejności od najmniejszej do największej i numerujemy od 1 do n . Następnie, jeśli n jest nieparzyste, medianą jest wartość obserwacji w środku (czyli obserwacji numer $\frac{n+1}{2}$). Jeśli natomiast n jest parzyste, wynikiem jest średnia arytmetyczna między dwiema środkowymi obserwacjami, czyli obserwacją numer $\frac{n}{2}$ i obserwacją numer $\frac{n}{2} + 1$.

Wariancja i odchylenie standardowe

Wariancja – miara rozproszenia wyników wokół średniej, możliwa do obliczenia tylko dla zmiennych o ilościowym poziomie pomiaru.

Wariancja jest wyliczana poprzez iloraz zsumowanych kwadratów odchyleń wyników od średniej, przez liczbę wyników pomniejszoną o 1.

Wariancja przybiera wartość od 0 do + nieskończoności. Przy zerowej wartości w danym zbiorze wyników nie ma żadnego zróżnicowania (wszystkie wyniki badanych są takie same). Z kolei wraz ze wzrostem wartości wariancji, zróżnicowanie wyników rośnie.

Wariancja jest kluczowym pojęciem dla testów porównujących wartości średnich (np. test t studenta dla prób niezależnych, analiza wariancji), w których jednym z założeń jest założenie o jednorodności wariancji. Ponadto, na podstawie wariancji i średniej, szacowane są wyniki na poziomie populacji.

Odchylenie standardowe to pierwiastek z wariancji, również mierzy zróżnicowanie zbioru.

Kowariancja

Kowariancja jest to wielkość charakteryzująca wspólne zmiany dwóch zmiennych X i Y. Jest oczekiwana wartością iloczynu odchyлеń wartości zmiennych X i Y od ich wartości oczekiwanych.

Zakładając, że X i Y to para **zmiennych losowych** o **rozkładach normalnych** i średnich μ_x i μ_y oraz standardowych odchyleniach σ_x i σ_y . Kowariancję dwóch **zmiennych** X i Y liczymy ze wzoru

$$\text{cov}(X, Y) = E[X - E(X)][Y - E(Y)]$$

co można też przedstawić w postaci

$$\text{cov}(X, Y) = E(XY) - E(X)E(Y)$$

Korelacja

Zależność korelacyjna pomiędzy cechami X i Y charakteryzuje się tym, że wartościom jednej cechy są przyporządkowane ścisłe określone wartości średnie drugiej cechy.

Correlation
Relationship Between Two Quantities
Such That When One Changes, the Other Does



Przedziały korelacji

Wartość współczynnika korelacji	Interpretacja
-1	idealna korelacja negatywna
od -1 do -0.9	korelacja wysoka
od -0.8 do -0.5	korelacja średnia
od -0.5 do -0.2	korelacja niewielka
od -0.2 do -0.1	korelacja niska
0	brak korelacji !
od 0.1 do 0.2	korelacja niska
od 0.2 do 0.5	korelacja niewielka
od 0.5 do 0.8	korelacja średnia
od 0.8 do 0.9	korelacja bardzo wysoka
1	pełna korelacja pozytywna

Przydatny materiał - korelacja

<https://starthere.pl/korelacja/#dziwna-korelacja>

<https://www.statystyka-zadania.pl/wspolczynnik-korelacji-spearmana/>

<https://statystycznie-istotne.pl/slownik-statystyczny/korelacja-kendalla/>

Zadania

1. Wczytaj plik ex1data2.txt a następnie oblicz dla niej średnią i odchylenie standarowe
2. Ustandaryzuj dane w podanej tablicy
3. Stwórz 2 wektory a następnie policz dla niego statystyki opisowe i korelacje

Sztuczna inteligencja a uczenie maszynowe

AI (Artifical Inteligence)

Maszyny potrafią rozwiązywać zadania zwykle kojarzone z ludzką inteligencją.

ML (Machine Learning)

Wnioskowanie z doświadczenia zamiast wyłącznego bazowania na programowaniu.

DL (Deep Learning)

Machine Learning wykorzystujące warstwy sieci neuronowych.

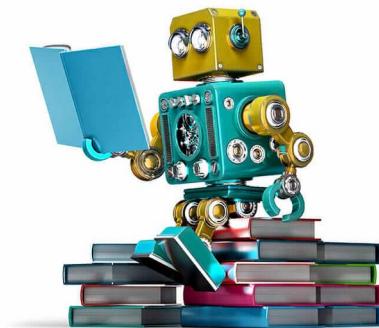
Czym jest uczenie maszynowe

Uczeniem maszynowym nazywamy

dziedzinę nauki programowania

komputerów w sposób umożliwiający

im **uczenie się z danych.**



Czym jest uczenie maszynowe?

Artur Samuel w 1959 roku podał następującą definicję uczenia maszynowego:

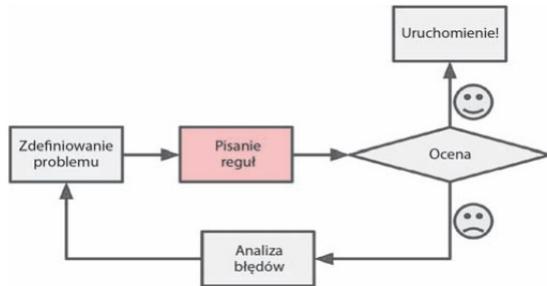
[Uczenie maszynowe to] dziedzina nauki dająca komputerom możliwość uczenia się bez konieczności ich jawnego programowania.

Czym jest uczenie maszynowe?

Podsumowując, by stworzyć model uczenia maszynowego potrzebujemy:

- zdefiniować zadanie, czyli co konkretnie nasz model ma robić, np. oznaczyć wiadomość e-mail jako spam lub oznaczyć jako nie-spam,
- dane, na podstawie których model nauczy się pewnych zależności lub różnic pomiędzy danymi, np. zbiór wiadomości oznaczonych jako spam lub nie-spam,
- funkcję oceny rezultatów, która powie nam, jak skuteczny jest dany model.

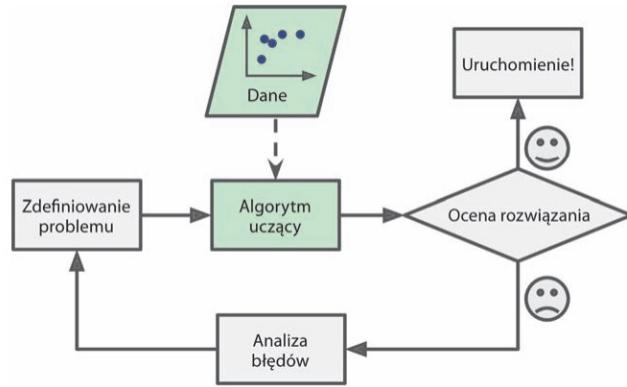
Czym różni się uczenie maszynowe od tradycyjnego programowania?



Załóżmy, że chcemy napisać filtr spamu przy pomocy tradycyjnych technik programistycznych. Musimy więc zastanowić się, jak wygląda klasyczny spam. Bardzo prawdopodobne, że zauważymy, że poszczególne słowa czy całe zwroty często występują w takich wiadomościach. Możemy więc napisać algorytm, który będzie wykrywał występowanie takich szablonów i jeśli pojawi się ich kilka w jednej wiadomości, zaklasyfikuje ją jako spam. Następnie proces ten należy powtórzyć, sprawdzając jak algorytm działa dla nowych wiadomości i dając mu nowe danie.

Czym różni się uczenie maszynowe od tradycyjnego programowania?

sages



Utrzymanie tego modelu jest dużo prostsze – gdy spamerzy nauczą się unikać określonych zwrotów, wystarczy zasilić model powiększoną próbką danych, a on automatycznie zauważyc wzrost częstotliwości pojawiania się nowego zwrotu w spamowych e-mailach i niejako „nauczy się”, że jest to przesłanka ku temu, by otagować wiadomość jako szkodliwą.

Czym różni się uczenie maszynowe od tradycyjnego programowania?

sages

Podsumowując, uczenie maszynowe nadaje się do:

- problemów, których rozwiązanie wymaga mnóstwo ręcznego dostrajania algorytmu lub korzystania z długich list reguł, często jeden algorytm upraszcza rozwiązanie i poprawia jej szybkość,
- złożonych problemów, których nie można rozwiązać tradycyjnymi metodami,
- zmiennych środowisk: algorytm uczenia maszynowego dostosuje się do nowych danych,
- pomagania człowiekowi w analizowaniu skomplikowanych zagadnień i olbrzymich ilości danych

Podział uczenia maszynowego

Uczenie maszynowe możemy podzielić na podstawie rodzaju nadzorowania procesu uczenia.

Wyróżniamy:

Uczenie
nadzorowane

Uczenie
nienadzorowane

Uczenie przez
wzmacianie

Uczenie nadzorowane

W **uczeniu nadzorowanym** (ang. *supervised learning*) dane uczące przekazywane algorytmowi zawierają dołączone rozwiązania problemu, tzw. **etykiety** (ang. *labels*).



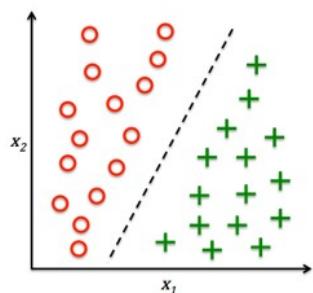
W analizowanym problemie tworzenia filtra spamu, by algorytm mógł się nauczyć różnic pomiędzy wiadomościami prawidłowymi i szkodliwymi, musimy zasilić go zbiorem e-maili z informacją, czy dana wiadomość była spamem, czy nie.

Uczenie nadzorowane

Klasyczne zadania uczenia nadzorowanego to **klasyfikacja i regresja**.

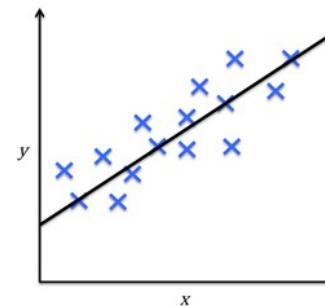


KLASYFIKACJA



W tym zadaniu chodzi o przyporządkowaniu każdego przykładu do konkretnej klasy, np. spam/nie-spam

REGRESJA



W tym zadaniu chodzi o przewidywane wartości numerycznej, takiej jak cena samochodu przy użyciu określonego zbioru cech (przebieg, wiek, marka, itd.)

Problemy uczenia maszynowego

- Niedobór danych – by nauczyć dziecko, czym jest jabłko, wystarczy pokazać ten owoc i powiedzieć "jabłko". Niestety, modele uczenia maszynowego wymagają zapewnienia im dużej ilości danych.
- Niereprezentatywne dane – jeśli chcemy zbudować filtr spamu, najlepiej byłoby, gdybyśmy posiadali zbiór danych z taką samą liczbą wiadomości szkodliwych, co normalnych. Zasilając algorytm 970 przykładami poprawnych wiadomości i 30 e-mailami typu SPAM, ciężko oczekować, że model czegoś się nauczy.

Problemy uczenia maszynowego

- Dane kiepskiej ilości - jeśli dane zawierają mnóstwo błędów, elementów odstających i szumu (np. z powodu niskiej jakości pomiarów), to systemowi będzie znacznie trudniej wykryć wzorce, przez co nie osiągnie optymalnej wydajności. Warto często poświęcić czas na oczyszczenie takich danych.
- Konieczność posiadania wiedzy dziedzinowej - by móc modelować np. predykcję cen mieszkań, musimy mieć świadomość i intuicję, jakie cechy będą na te ceny wpływać.

Gdzie algorytmy machine learning mogą mieć zastosowanie?

- Bankowość i ubezpieczenia
- Analiza obrazów medycznych
- Ukierunkowana reklama
- Optymalizacja czasu podróży
- Sport
- ... i wszystkie pozostałe

Problemy uczenia maszynowego

- Niedobór danych – by nauczyć dziecko, czym jest jabłko, wystarczy pokazać ten owoc i powiedzieć "jabłko". Niestety, modele uczenia maszynowego wymagają zapewnienia im dużej ilości danych.
- Niereprezentatywne dane – jeśli chcemy zbudować filtr spamu, najlepiej byłoby, gdybyśmy posiadali zbiór danych z taką samą liczbą wiadomości szkodliwych, co normalnych. Zasilając algorytm 970 przykładami poprawnych wiadomości i 30 e-mailami typu SPAM, ciężko oczekiwać, że model czegoś się nauczy.

Problemy uczenia maszynowego

- **Twierdzenie o nieistnieniu darmowych obiadów** (ang. *No Free Lunch Theorem - NFL*)

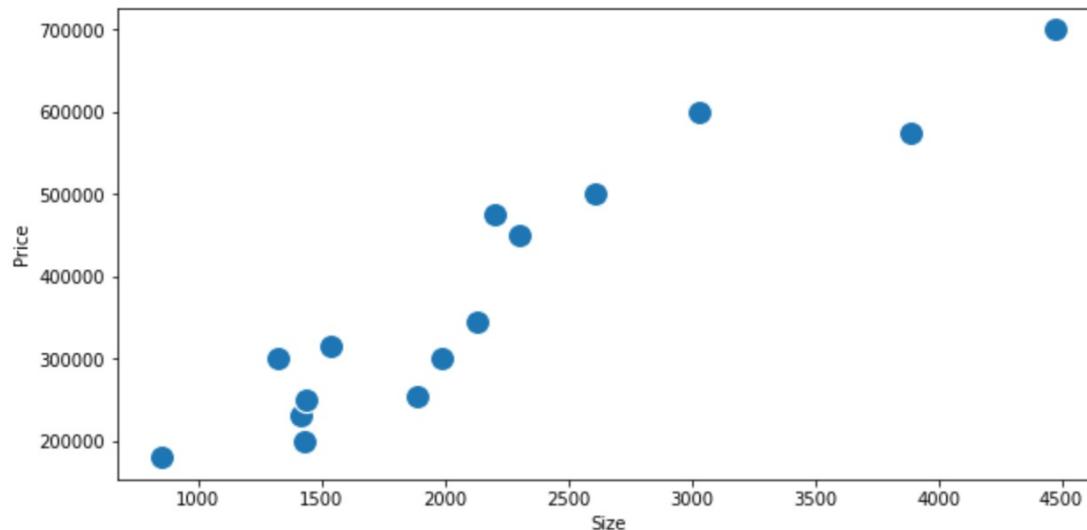
W publikacji z 1996 roku David Wolpert podał tezę, że żaden model nie będzie lepszy od pozostałych - dla pewnych zbiorów danych najlepiej nadaje się model liniowy, natomiast dla innych - sieci neuronowe. Nie istnieje jednak model, który z założenia będzie działał lepiej. Jedynie poprzez ocenę działania każdego modelu możemy przekonać się, który z nich będzie sprawował się najlepiej. W praktyce przyjmujemy rozsądne założenia dotyczące danych i oceniamy działanie tylko kilku rozsądnie dobranych modeli. Na przykład wobec prostych zadań możemy ocenić działanie modeli liniowych różniących się stopniem regularyzacji, a bardziej skomplikowane problemy możemy przetestować przy użyciu sieci neuronowych.

Innymi słowy: There's no „one to rule them all”.



Ceny domów w zależności od ich powierzchni

Poniższy wykres prezentuje, jak wyglądają ceny domów (w dolarach) w Portland (Oregon, USA) w zależności od ich powierzchni (w stopach kwadratowych).



Ceny domów w zależności od ich powierzchni

Tak wyglądają te same dane w postaci tabeli.

Widzimy więc, że za dom o powierzchni 1427 stóp kwadratowych musimy zapłacić 198999 dolarów. Mamy tu dwie zmienne: wielkość oraz cenę. Cena domu zależy od jego wielkości, więc cena będzie **zmienną zależną**, a wielkość - **zmienną niezależną**.

Size	Price
1427	198999
3031	599000
4478	699900
2200	475000
852	179900
1320	299900
1416	232000
1888	255000
1534	314900
2132	345000
1985	299900
2609	499998
2300	449900
3890	573900
1437	249900

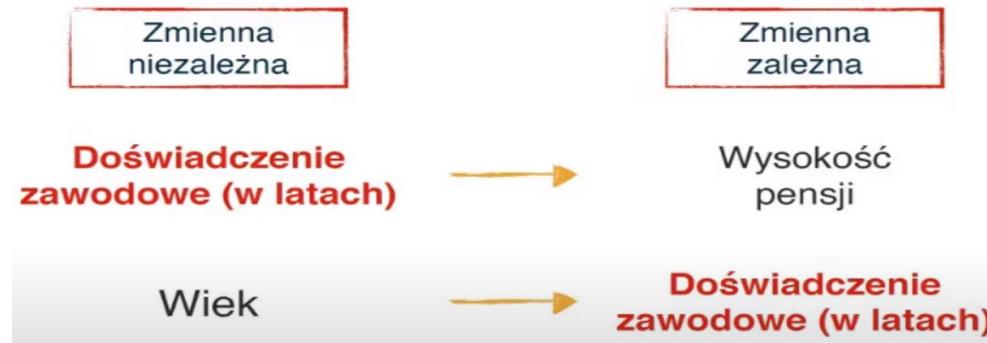
Zmienne niezależne i zależne

Zmienne niezależne to te, które w danym badaniu manipujemy, zmieniamy ich wartości.

Zmienne zależne to te, które opisują wynik doświadczenia, które obserwujemy i mierzymy, sprawdzając, jak zmiana zmiennych niezależnych wpływa na wynik tego doświadczenia.

Dla przykładu: jeśli dziecko wykonuje prace domowe (wynosi śmieci, wiesza pranie, zmywa naczynia), to za każde wykonane zadanie dostaje od rodziców 5 złotych. W tym przypadku zmienną niezależną jest liczba prac domowych, które wykona dziecko, bo to ono ma nad tym kontrolę. Zmienną zależną jest ilość zarobionych pieniędzy, ponieważ ta wartość zależy od tego, ile czynności zostanie wykonanych.

Zmienne niezależne i zależne



W pierwszym przykładzie badamy wpływ doświadczenia zawodowego w latach na wysokość pensji, doświadczenie będzie więc zmienną niezależną, a wysokość pensji - zależną.

W drugim przykładzie badamy, w jaki sposób wiek ludzi wpływa na ich doświadczenie zawodowe, więc wiek będzie zmienną niezależną, natomiast doświadczenie - zależną.

Konwencja oznaczeń

Konwencja jest taka, że zmienną niezależną oznaczamy jako **x**, zależną - jako **y**.

W konwencji nazewnicyzej uczenia maszynowego **x** będzie cechą, a **y** - etykietą. Każdy przykład jest parą wartości (**x, y**).

Size	Price
1427	198999
3031	599000
4478	699900
2200	475000
852	179900
1320	299900
1416	232000
1888	255000
1534	314900
2132	345000
1985	299900
2609	499998
2300	449900
3890	573900
1437	249900

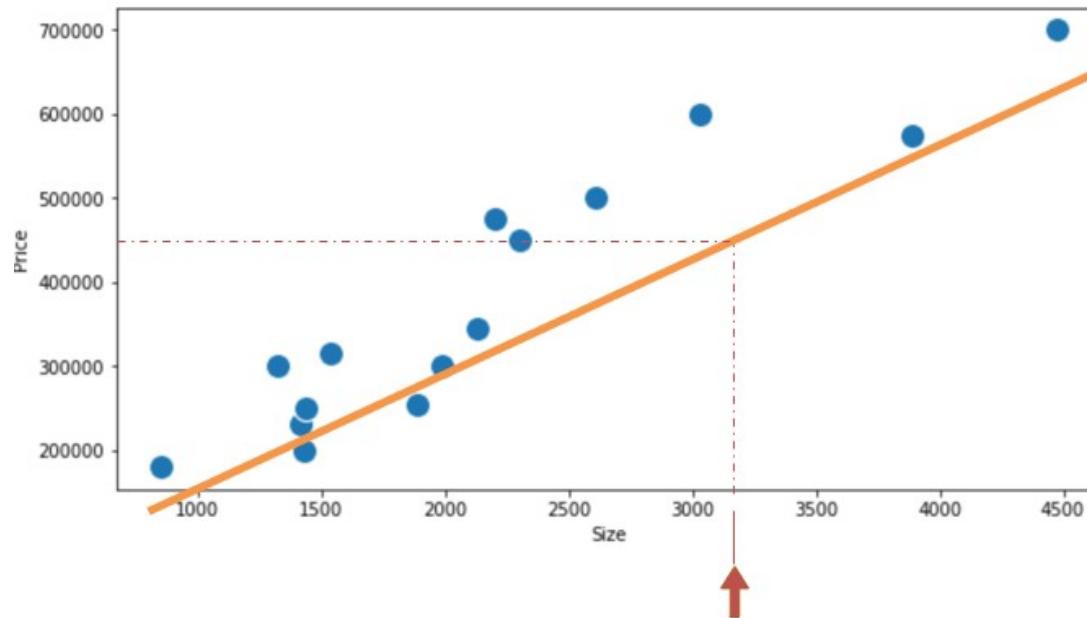
Różne typy zmiennych

Zmienne dzielimy na:

- ilościowe (mierzalne) - np. wzrost, masa, wiek
 - ciągłe, np. wzrost, masa, wiek, temperatura
 - porządkowe, np. wzrost w postaci kategorycznej (niski, średni, wysoki), wykształcenie (podstawowe, średnie, wyższe)
 - skokowe (dyskretne), np. liczba posiadanych dzieci
- jakościowe (niemierzalne) - np. kolor oczu, płeć, grupa krwi

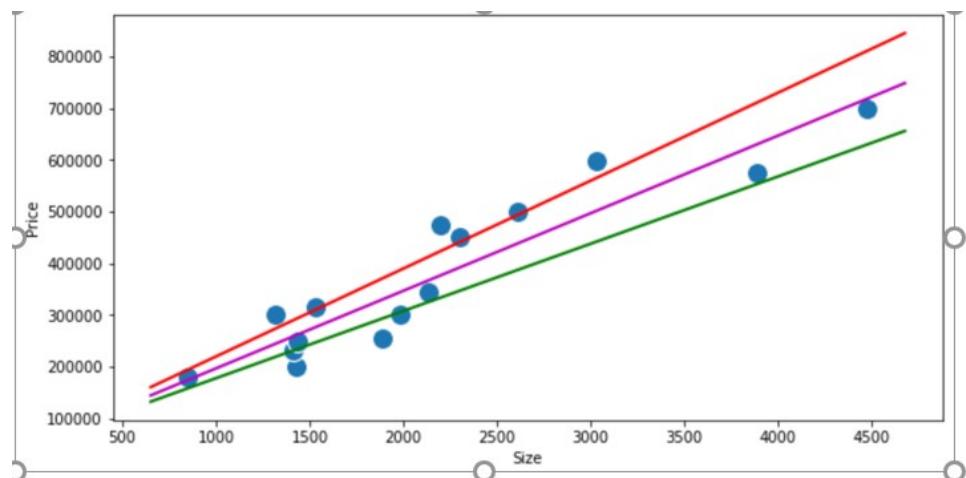
Regresja liniowa

Widzimy, że punkty układają się liniowo, możemy więc poprowadzić linię i odczytać wartość przecięcia.

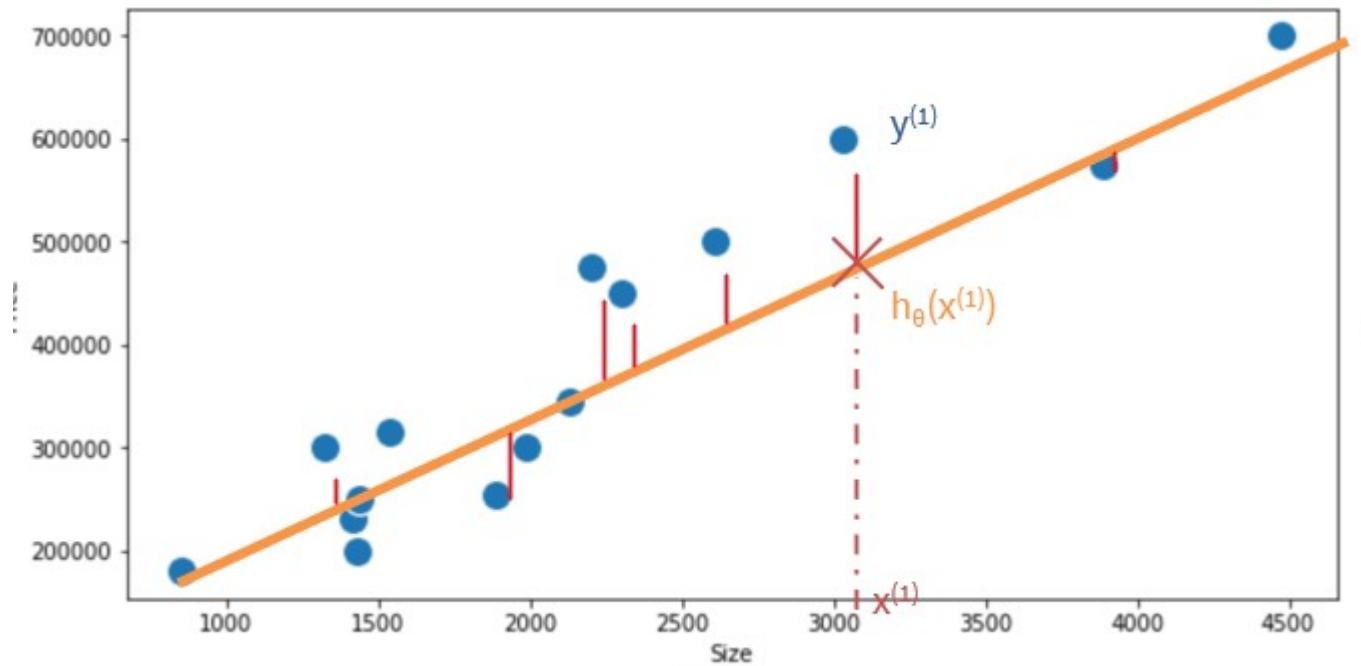


Regresja liniowa

Ale w jaki sposób tę linię poprowadzić?
Przecież możemy zrobić to na wiele sposobów,
skąd mamy wiedzieć, który z nich jest
najlepszy?



Regresja liniowa



Regresja liniowa

Jeśli więc chcemy dowiedzieć się, jak dobrze prosta regresyjna jest dopasowana do danych, musimy dodać do siebie wszystkie reszty. Mogą być one jednak dodatnie lub ujemne, więc prosta suma wszystkich reszt może być równa, nawet jeśli poszczególne reszty nie są bliskie zeru.

Możemy zamiast reszt rozpatrywać kwadraty reszt, które zawsze będą nieujemne. Taką metodę dopasowania linii prostej, by leżała jak najbliżej wszystkich wyników, nazywamy **metodą najmniejszych kwadratów**.

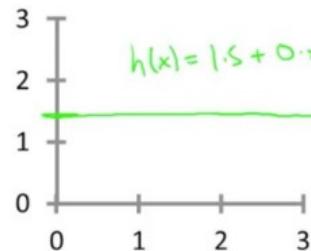
Regresja liniowa

Regresję liniową zapisujemy wzorem:

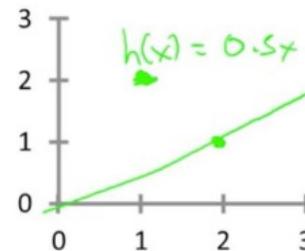
$$h_{\theta}(x) = \theta_0 + \theta_1 x$$

gdzie θ_0 oraz θ_1 nazywamy **współczynnikami regresji** (parametrami modelu). Spójrzmy jak wartości współczynników wpływają na to, jak wygląda wykres prostej regresyjnej:

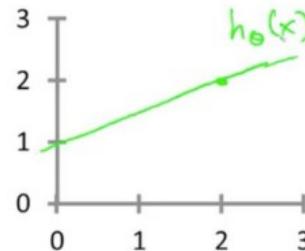
$$\underline{h_{\theta}(x) = \theta_0 + \theta_1 x}$$



$$\begin{aligned} \rightarrow \theta_0 &= 1.5 \\ \rightarrow \theta_1 &= 0 \end{aligned}$$



$$\begin{aligned} \rightarrow \theta_0 &= 0 \\ \rightarrow \theta_1 &= 0.5 \end{aligned}$$



$$\begin{aligned} \rightarrow \theta_0 &= 1 \\ \rightarrow \theta_1 &= 0.5 \end{aligned}$$



Funkcja kosztu

Formalnie możemy zapisać to w następujący sposób:

$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^m (h_\theta(x_i) - y_i)^2$$

sumujemy więc wszystkie kwadraty różnic pomiędzy wartościami wyestymowanymi przez funkcję h a faktycznymi wartościami (ground truth) y , a następnie dzielimy przez liczbę przykładów (m) oraz przez 2. Funkcję J nazywamy **funkcją kosztu**.

Funkcja kosztu

Problemem, który musimy rozwiązać to znalezienie takich parametrów θ , by wartość funkcji J była jak najmniejsza.

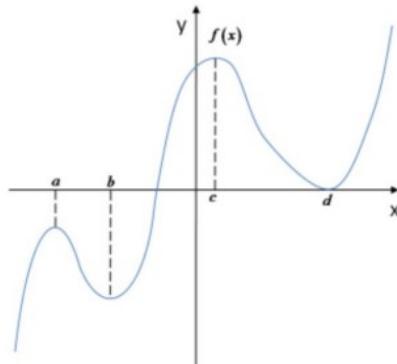
$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^m (h_\theta(x_i) - y_i)^2$$

Musimy więc rozwiązać problem optymalizacyjny.

Ekstrema funkcji

sages

Czym jest ekstremum funkcji? *Extremum* z łaciny oznacza *skrajne*. Są dwa rodzaje ekstremów funkcji: **minimum** i **maksimum**.



Powysza funkcja ma minima w punktach b i d oraz maksima w punktach a i c . Widzimy więc, że nie można tych pojęć mylić z najmniejszą i największą wartością funkcji. Te ekstrema nazywamy lokalnymi.

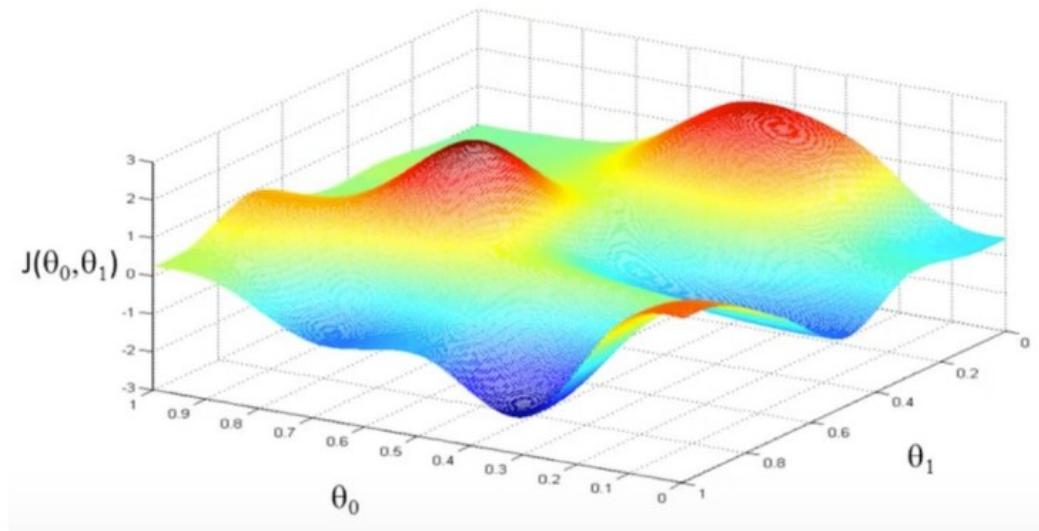
Ekstrema funkcji

Formalnie mówimy, że **ekstremum lokalne** dotyczy pewnego otoczenia (przedziału) punktu, w którym funkcja w żadnym punkcie nie przyjmuje wartości większych (w przypadku maksimum) lub mniejszych (w przypadku minimum).

Wyróżniamy również **ekstremum globalne** i jest to największa lub najmniejsza wartość funkcji.

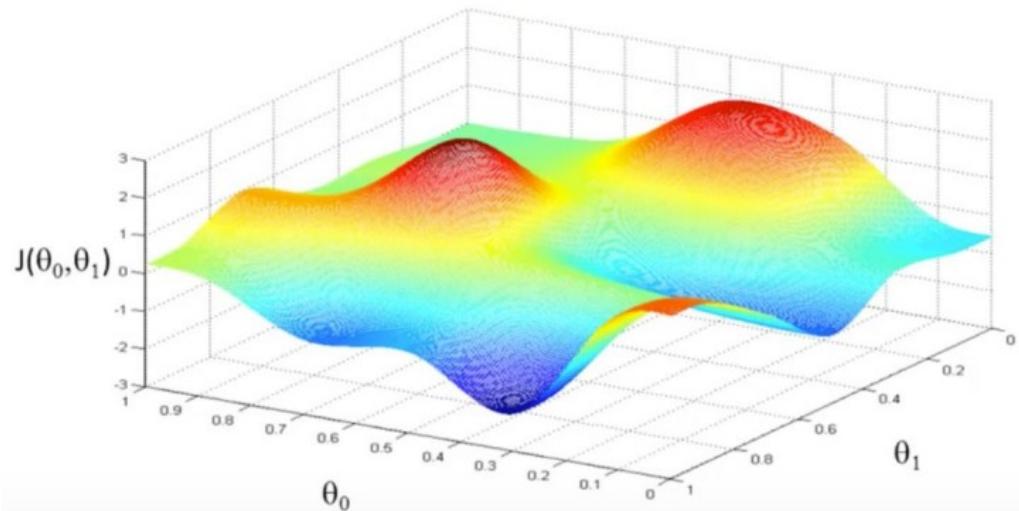
Problem optymalizacyjny

Nasza funkcja J zależy od dwóch parametrów, jej wykres musimy więc narysować w trzech wymiarach. Jak taki wykres może wyglądać?



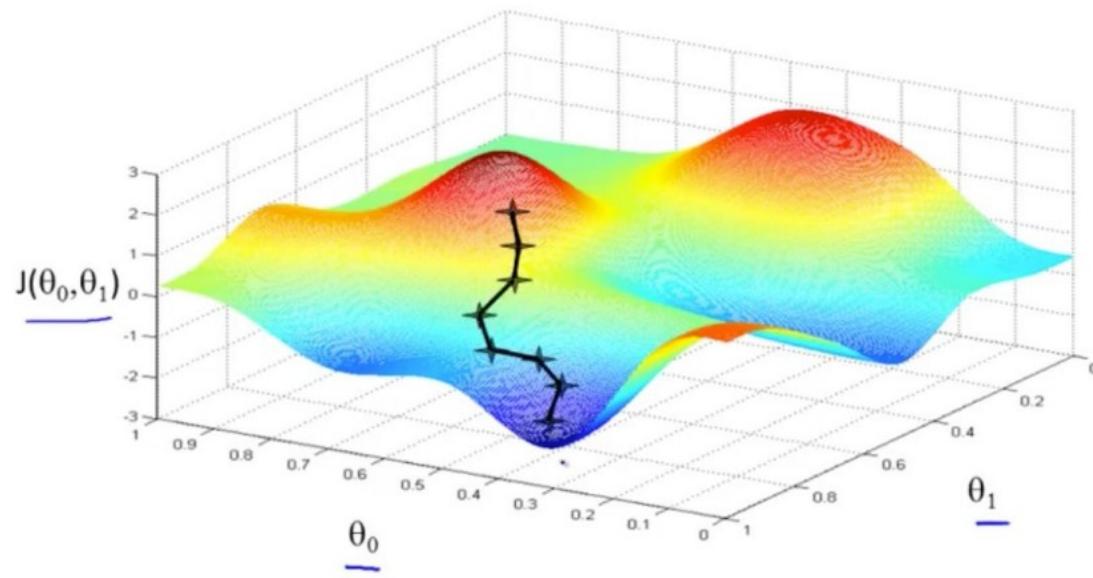
Problem optymalizacyjny

Jak widzimy, funkcja ta ma wiele ekstremów, lokalne maksima zostały zaznaczone na czerwono, lokalne minima - na granatowo. Posiada jednak tylko jedno maksimum i minimum globalne.



Problem optymalizacyjny

To może zacznijmy w dowolnym punkcie (θ_0, θ_1) i przesuwajmy się po płaszczyźnie ku dołowi tak długo, aż nie znajdziemy minimum?



Metoda gradientu prostego

Zaczynanie od dowolnie wybranego punktu i poruszanie się cały czas ku minimum lokalnemu funkcji nazywamy metodą gradientu prostego.

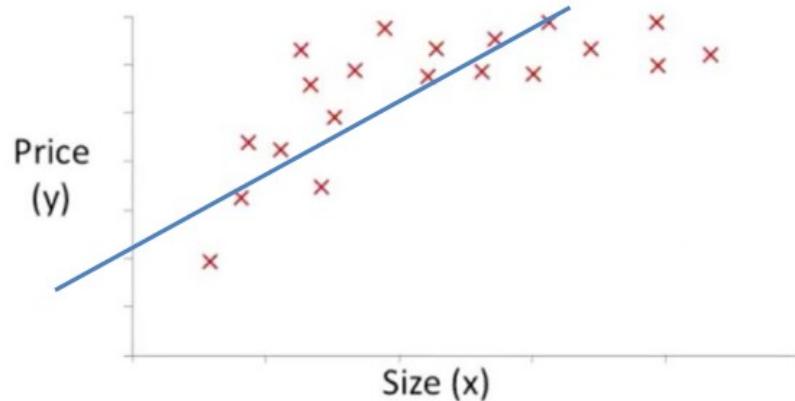
Ale co formalnie znaczy to poruszanie się? Oznacza to, że przy każdym kroku musimy aktualizować nasze wagi w oparciu o wzór:

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1) \quad (\text{for } j = 0 \text{ and } j = 1)$$

gdzie $\frac{\partial}{\partial \theta_j}$ jest pochodną cząstkową funkcji J względem parametru θ_j , a α - dobraną stałą, zwaną **współczynnikiem uczenia** (ang. *learning rate*) - zwykle małą, mniejszą od zera, np. 0,01.

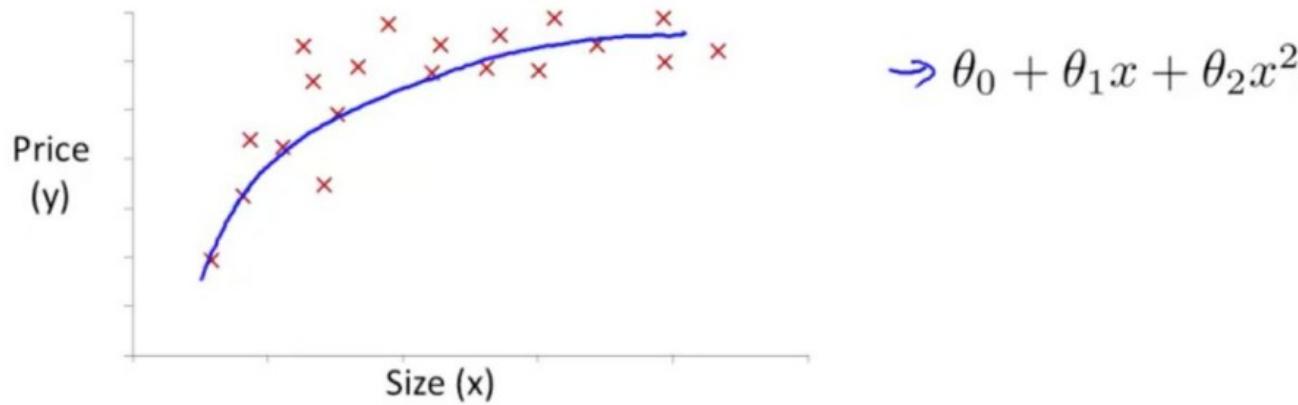
Regresja wielomianowa

Nie zawsze między danymi będzie występowała idealna liniowa korelacja. Jak widzimy, w tym przypadku ciężko jest wyrysować taką prostą, która dobrze by odwzorowywała charakterystykę punktów



Regresja wielomianowa

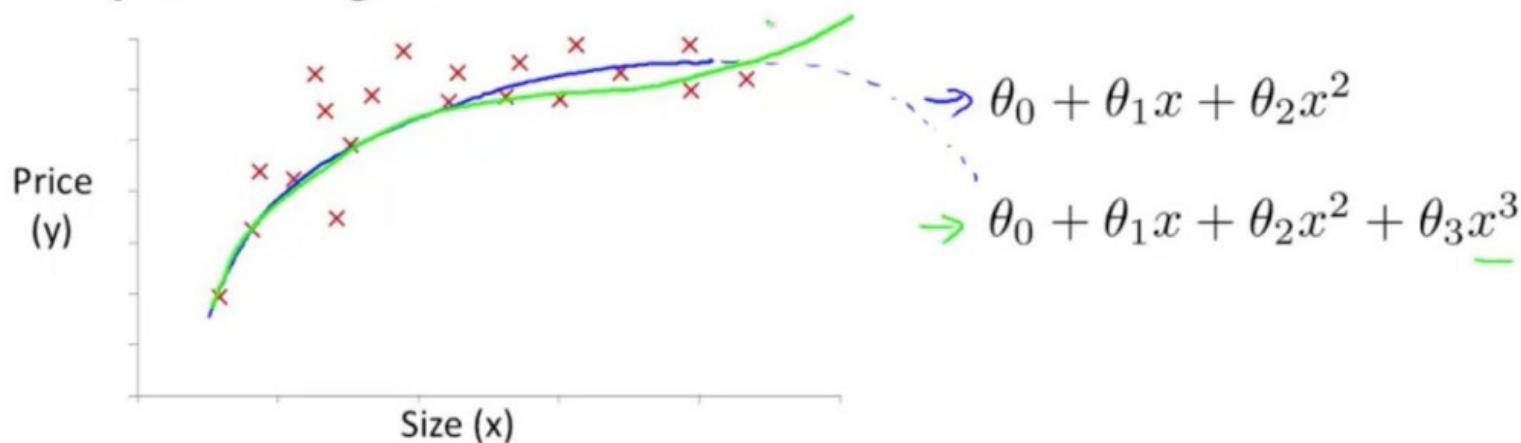
W takim przypadku możemy spróbować wyrysować krzywą. Dla tych danych funkcja kwadratowa pewnie wyglądałaby lepiej:



Regresja wielomianowa

Lepiej więc wyrysować krzywą wielomianu stopnia 3 (linia zielona)

Polynomial regression



Jak oceniać modele?

Wiemy już, jak tworzyć instancje modeli i jak przekazać im dane, z których wyciągają różne zależności, np. pomiędzy ceną domu a jego powierzchnią.

W jaki sposób porównać ze sobą dwa modele? Skąd wiemy, który z nich jest lepszy?

Jak oceniać modele?

Zależy nam, by nasz model poprawnie przewidywał dowolne dane, a nie tylko te, które widział podczas procesu uczenia. To znaczy, że chcemy, by miał dużą zdolność do **generalizacji**, czyli używać wyuczonych zależności do predykcji nowych przykładów.

Możemy więc stworzyć model, zasilić danymi, na których go wytrenujemy, a następnie ręcznie wpisywać kolejne przykłady i sprawdzać, czy zwraca wartości podobne do tych, których byśmy się spodziewali. Takie podejście jest jednak czasochłonne i kosztowne, wymaga zaangażowania w to zadanie konkretnej osoby. Co więcej, pewnie w trakcie realizacji projektu będziemy wielokrotnie musieli powtarzać walidację modeli (np. po zmianie parametru lub dodaniu nowej cechy), co przekłada się na jeszcze większy narzut czasowy i kosztowy.

Jak taki proces zautomatyzować?



Zbiór treningowy i testowy

Lepszym pomysłem jest podział danych na dwa zestawy: **zbiór uczący** (ang. *training set*), **zbiór testowy** (ang. *test set*). Model trenujemy przy pomocy zbioru uczącego, a sprawdzamy jego poprawność (przy użyciu pewnej metryki, ale o tym później) na zbiorze testowym.

Zazwyczaj na zbiór uczący składa się 80% danych, a pozostałe 20% odkładamy jako zbiór testowy.

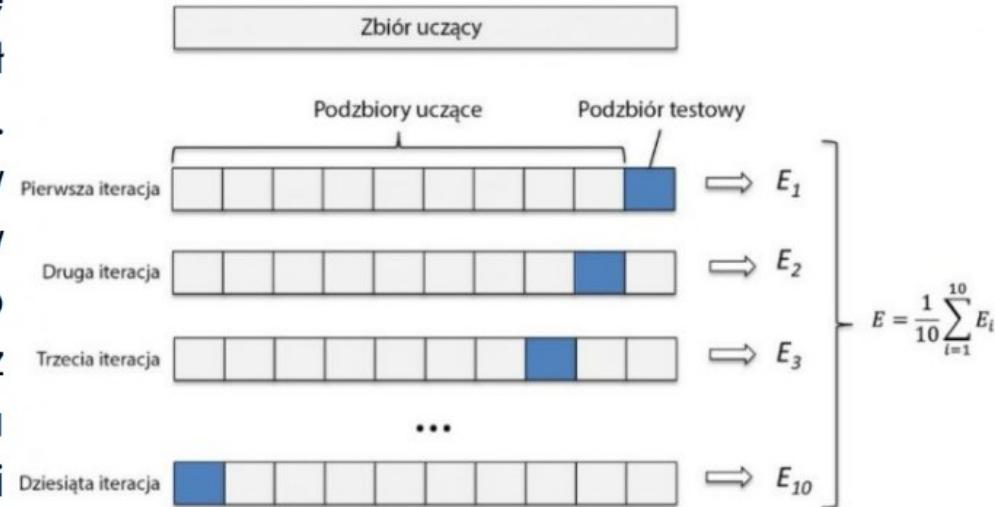
Sprawdzian krzyżowy

By uniknąć poświęcania dużej ilości danych uczących na zbiór walidacyjny, możemy skorzystać z techniki zwanej **sprawdzianem krzyżowym** lub **kroswalidacją** (ang. *cross-validation*).

Zbiór uczący zostaje rozdzielony na wzajemnie uzupełniające się podzbiory - każdy model jest uczyony za pomocą różnych kombinacji tych podzbiorów i oceniany przy użyciu pozostałych, nieużytych podzestawów. Po dobraniu rodzaju i hiperparametrów modelu ostatecznie zostaje on wytrenowany na całym zbiorze uczącym, a błąd oceniamy przy użyciu zbioru testowego.

k-krotna walidacja krzyżowa

Przykład ilustruje 10-krotną walidację krzyżową - zbiór uczący został podzielony na 10 podzbiorów. Następnie wykonano 10 iteracji, w każdej z nich biorąc 9 podzbiorów jako zbiór uczący, a jeden jako testowy. Wyniki uzyskane w każdej z iteracji są uśredniane w celu otrzymania finalnej skuteczności modelu.



Jak ocenić na ile dobra jest regresja?

Możemy skorzystać ze **współczynnika determinacji**, który jest opisową miarą siły liniowego związku między zmiennymi, czyli miarą dopasowania linii regresji do danych.

Przyjmuje wartości z przedziału [0, 1] i wskazuje jaka część zmiennej y jest wyjaśniania przez znaleziony model. Dla przykładu, dla $R^2=0,619$ znaleziony model wyjaśnia około 62% zmienności y.

MAE

MAE (Mean Absolute Error) mierzy średnią różnicę pomiędzy wartościami przewidzianymi a rzeczywistymi.

Najpierw liczymy różnicę między predykcją a wartością rzeczywistą, bierzemy z tego wartość absolutną, sumujemy dla wszystkich przykładów, a następnie dzielimy przez ich liczbę.

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

gdzie y_i oznacza wartość rzeczywistą, a \hat{y}_i – wartość przewidzianą przez model

MAE

Metryka jest łatwa do zrozumienia i policzenia, ale ma też swoje wady. Poprzez wartość absolutną tracimy информацию o kierunku, czy przewidzieliśmy wartość wyższą, czy niższą od oczekiwanej. Również wartości odstające mogą w znaczący sposób wpływać na wynik.

Dla przykładu, budujemy model do przewidywania opóźnień w odlocie samolotów, tak by pasażerowie mogli decydować, kiedy przyjadą na lotnisko. W takim rozważaniu kara za przewidzenie wyższej wartości powinna być wyższa, bo w takim przypadku pasażerowie mogą spóźnić się na samolot, w przeciwnym wypadku jedynie tracą czas oczekując na odlot samolotu. A w przypadku, gdyby wszystkie opóźnienia odlotów były godzinne, a jedno kilkudniowe, wartość MAE będzie bardzo wysoka i nie będzie oddawała rzeczywistości.

MSE

MSE (Mean Squared Error) podobnie jak MAE będzie skupiała się bardziej na dużych błędach.

Najpierw liczymy różnicę między predykcją, a wartością rzeczywistą, podnosimy do kwadratu, sumujemy dla wszystkich przykładów, a następnie dzielimy przez ich liczbę.

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

gdzie y_i oznacza wartość rzeczywistą, a \hat{y}_i - wartość przewidzianą przez model

RMSE

RMSE (Root Mean Squared Error) to pierwiastek kwadratowy z MSE.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

gdzie y_i oznacza wartość rzeczywistą, a \hat{y}_i – wartość przewidzianą przez model

Zdolność modeli do generalizacji

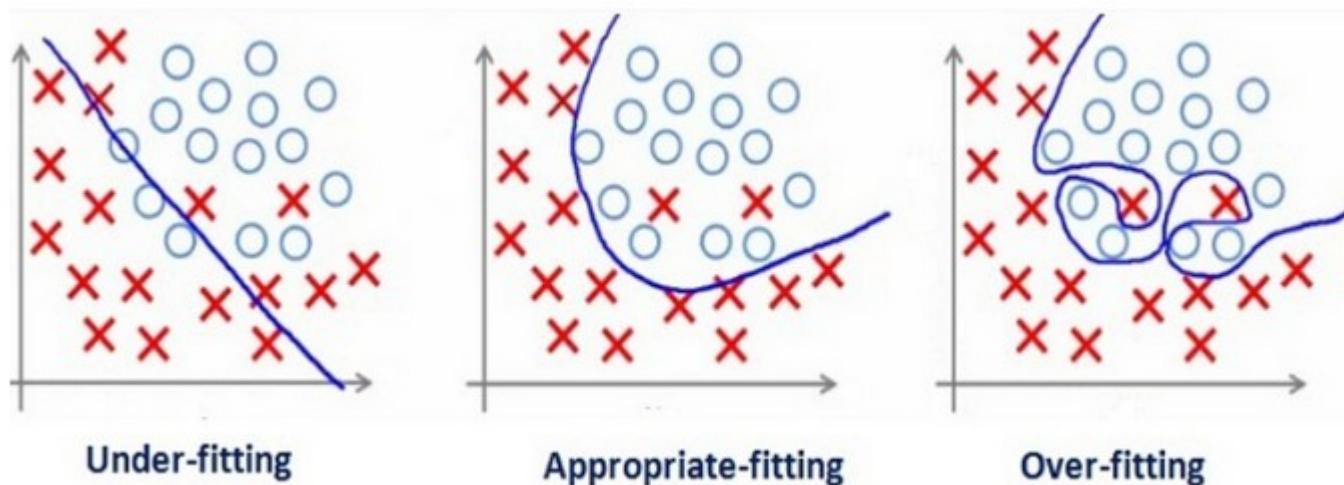
Mówiliśmy, że zależy nam, by modele miały wysoką zdolność do generalizacji. To znaczy, że na podstawie danych treningowych mają się wyuczyć w taki sposób, by zwracać jak najlepsze predykcje, dla danych, których nigdy nie widziały.

Nie zawsze jednak się to udaje i możemy mieć do czynienia z **przetrenowaniem** (ang. *overfitting*) lub **niedotrenowaniem** (*underfitting*).

Niedouczenie i przeuczenie modelu

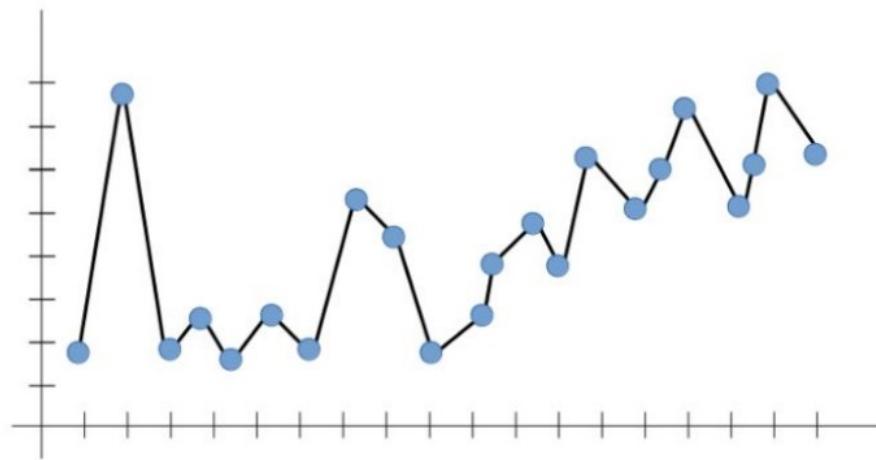
sages

Z drugiej strony, należy również uważać, czy model nie jest niedouczony, to znaczy za bardzo generalizuje dane



Przetrenowanie modelu

Z przetrenowaniem modelu mamy do czynienia w przypadku, gdy wygląda tak:



Widzimy tu zbyt silne dopasowanie do danych uczących - taki model nie będzie dobrze dokonywał predykcji dla danych spoza zbioru treningowego.

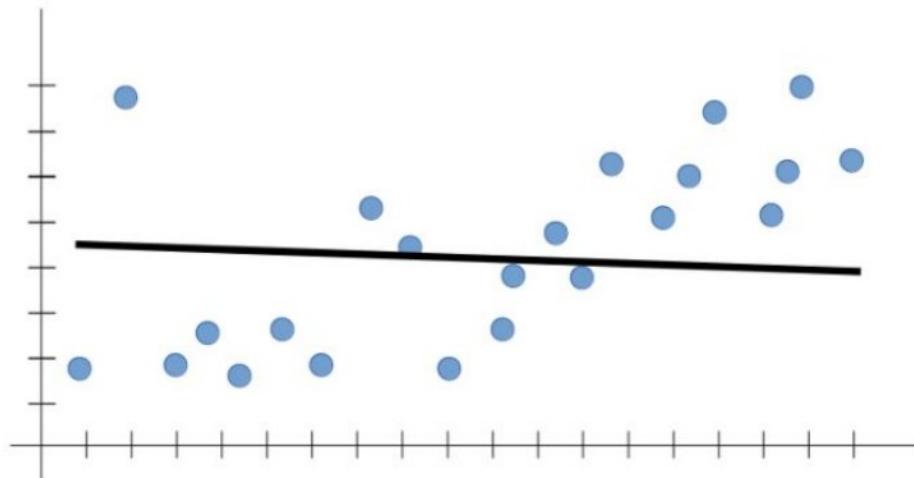
Przetrenowanie modelu

Zjawisko przetrenowania występuje, gdy model jest zbyt skomplikowany w porównaniu do ilości lub zaszumienia danych uczących. W takim przypadku możliwe są następujące rozwiązania:

- uproszczenie modelu poprzez wybór zawierającego mniej parametrów (np. modelu liniowego zamiast wielomianowego), zmniejszenie liczby atrybutów w danych uczących lub ograniczenie modelu,
- pozyskanie większej ilości danych uczących,
- zmniejszanie zaszumienia danych uczących (np. poprzez usunięcie błędnych danych lub elementów odstających),
- użycie walidacji krzyżowej.

Niedotrenowanie modelu

Z taką sytuacją będziemy mieli do czynienia, gdy wytrenowany model będzie wyglądał na przykład tak:



Widzimy, że wyrysowana prosta nie dopasowuje się nawet do zbioru uczącego, więc tym bardziej model nie będzie dobrze prognozował przykładów nigdy wcześniej nie widzianych.

Niedotrenowanie modelu

Sposobami na rozwiązywanie problemu niedotrenowania są:

- wybór bardziej skomplikowanego modelu, który wykorzystuje większą liczbę parametrów,
- dołgczanie większej liczby cech do algorytmu uczącego (feature engineering),
- znalezienie odpowiednich hiperparametrów.