Statystyka cz.2 i model regresji liniowej

O prowadzącym

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 zaobserowanego 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:

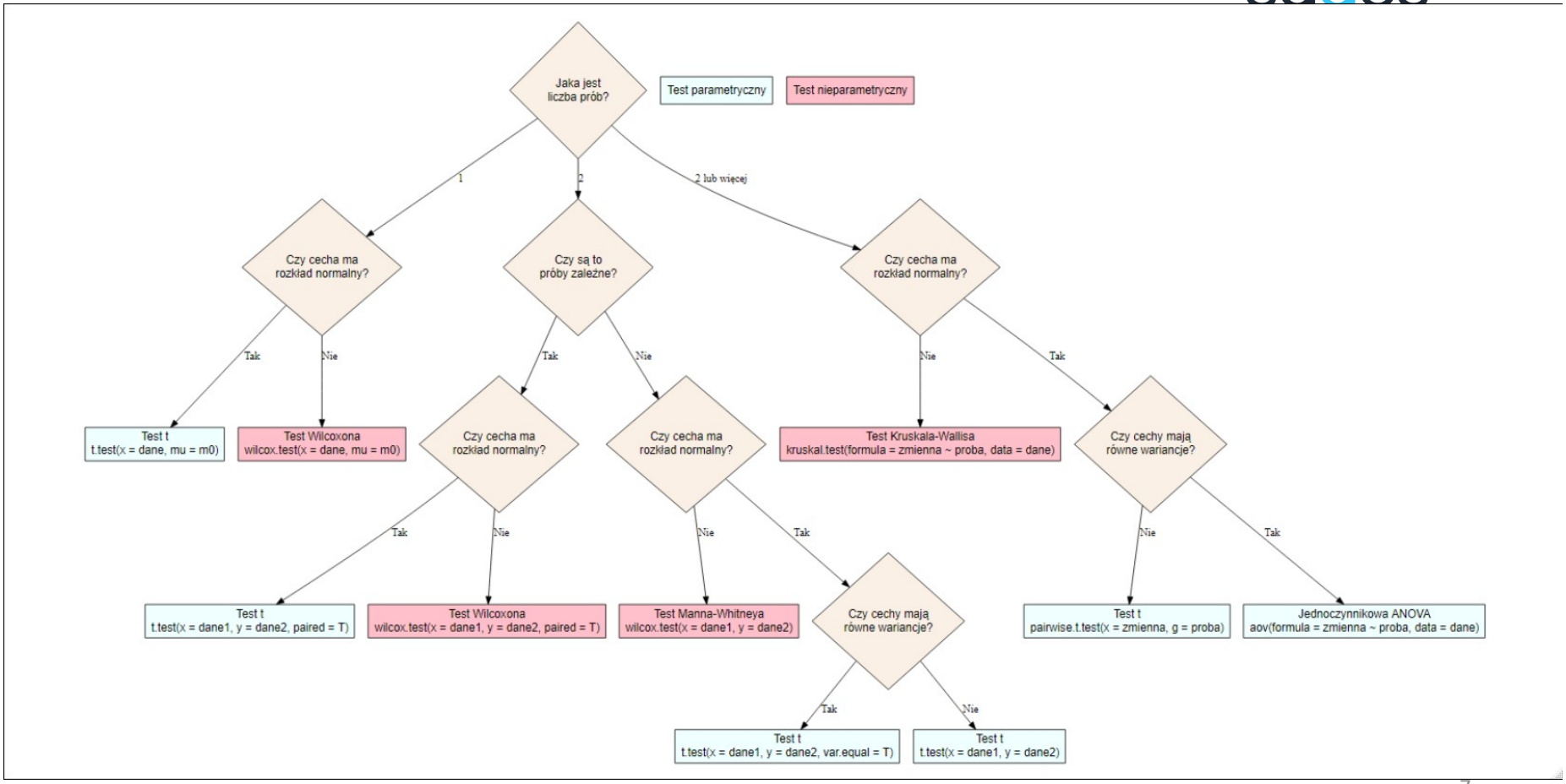
H_0: pomiędzy płcią a przynależnością do związków nie ma zależności

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.

Testowanie średnich



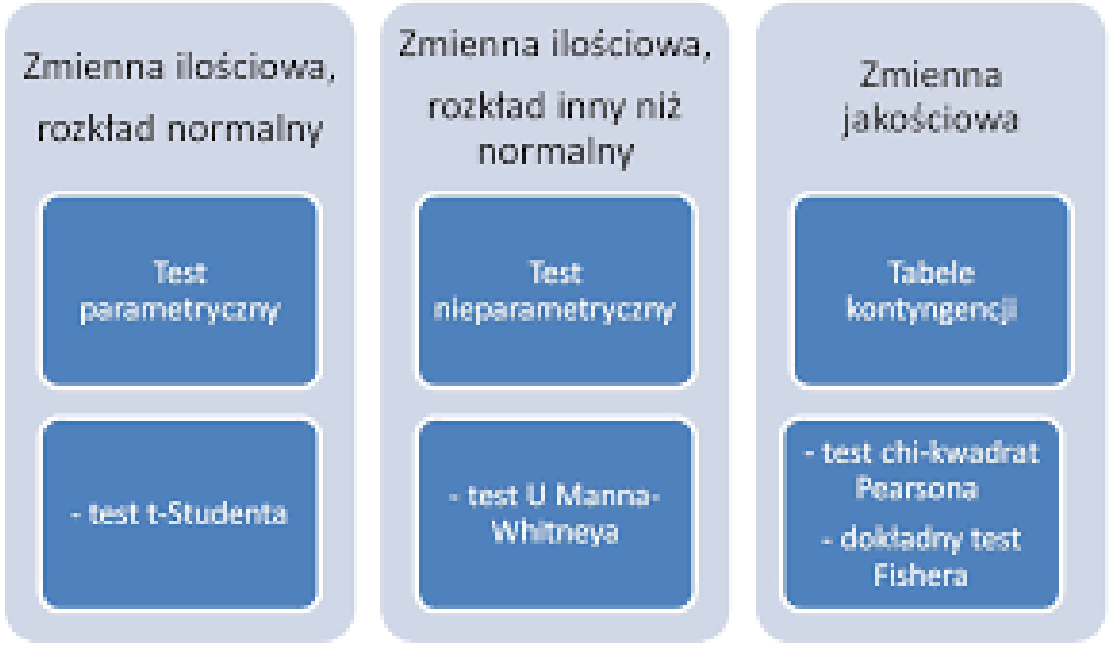
ANOVA - analiza wariancji

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.

Możemy podzielić analizę wariancji na trzy grupy analiz:

- **jednoczynnikowa analiza wariancji**
  - *wpływ jednego czynnika międzygrupowego na zmienną zależną*
- **wieloczynnikowa analiza wariancji**
  - *wpływ kliku czynników międzygrupowych na zmienną zależną*
- **analiza wariancji dla czynników wewnątrzgrupowych**
  - *wpływ czynnika wewnątrzgrupowego na zmienną zależną, tzw. "powtarzane pomiary"*

Porównanie testów



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 kazdym miesiacu)
2. Jeżeli rozkład jest normalny to:
   mamy 1 lub 2 grupy to test t-studenta (1 lub 2 stronny w zaleznosci ile prob (ang samples)) - zmienne sa tu zależne (zaleza nawzajem od siebie) wowczas porownujemy wariancje
   jezeli zmienne są niezależne to porównujemy srednie
   jeżeli mamy więcej niż 2 grupy to analiza ANOVA

3. Jezeli zmienne nie mają rozkładu normalnego:

a. dla 2 grup:
   jezeli zmienne sa niezalezne to porownujemy mediany ze sobą testem rank Wilcoxona
   jezeli zmienne sa zalezne to porównujemy testem znaków Wilcoxona

b. dla więcej niż 2 grup:
   tzw. nieparametryczna ANOVA (tu porownanie median przy parametrycznej byla srednia!!)

4. Jeżeli dane są dyskretne:

a. dla 2 grup (zmienne niezalezne)
   test chi-kwadrat liczności (test ktory sprawdza liczność w grupach)
   test fischera dla bardzo malych grup jeżeli wystepuje jakas liczność poniżej 5!

b. dla 2 grup (zmienne zalezne):
   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 - ogolnie w parametrycznych porównujemy PARAMETRY (parametrami sa srednia, mediana, odchylenie wariancja), wiec wszystkie te mowiac krotko ktore maja rozklad normalny, w nieparametrycznych porownujemy inne kryeteria ktore parametrami nie są jak np. proporcje, rankingi itd.

Proba zależna vs niezależna :

zalezna - badanie grupy w roznych warunkach : przed i po ingerencji np. przed i po zastosowaniu leku, badanie zysków przed i po zastosowaniu jakies kampanii marketingowej, badanie satysfakcji pracownikow przed i po wprowadzeniu pracy zdalnej

niezalezna - badanie roznych grup lub tej samej nie podanej ingerencji np. badania kobiet i mezczyzn po zastosowaniu leku, analiza zyskow w 2 firmach po zastosowaniu kampanii marketingowej.

Statystyka w data science

- Użyteczne pojęcia w data science z zakresu statystyki matematycznej to:
- Średnia
- Mediana
- 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 odchyleń wartości zmiennych X i Y od ich wartości oczekiwanych.
Zakładając, że X i Y to para zmiennych losowych o rozkładach normalnych i średnich $\mu_x|$ i $\mu_y|$ oraz standardowych odchyleniach $\sigma_x|$ i $\sigma_y|$. Kowariancję dwóch zmiennych X i Y liczymy ze wzoru

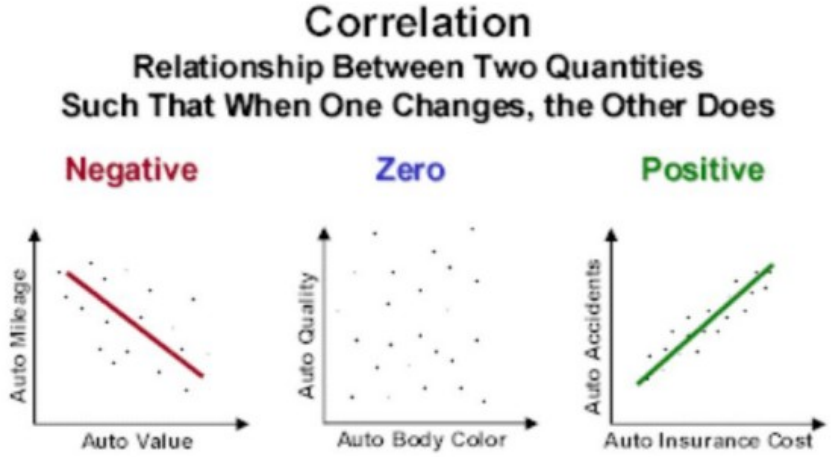
$$cov(X,Y) = E[X - E(X)][Y - E(Y)]|$$

co można też przedstawić w postaci

$$cov(X,Y) = E(XY) - E(X)E(Y)|$$

Korelacja

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



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

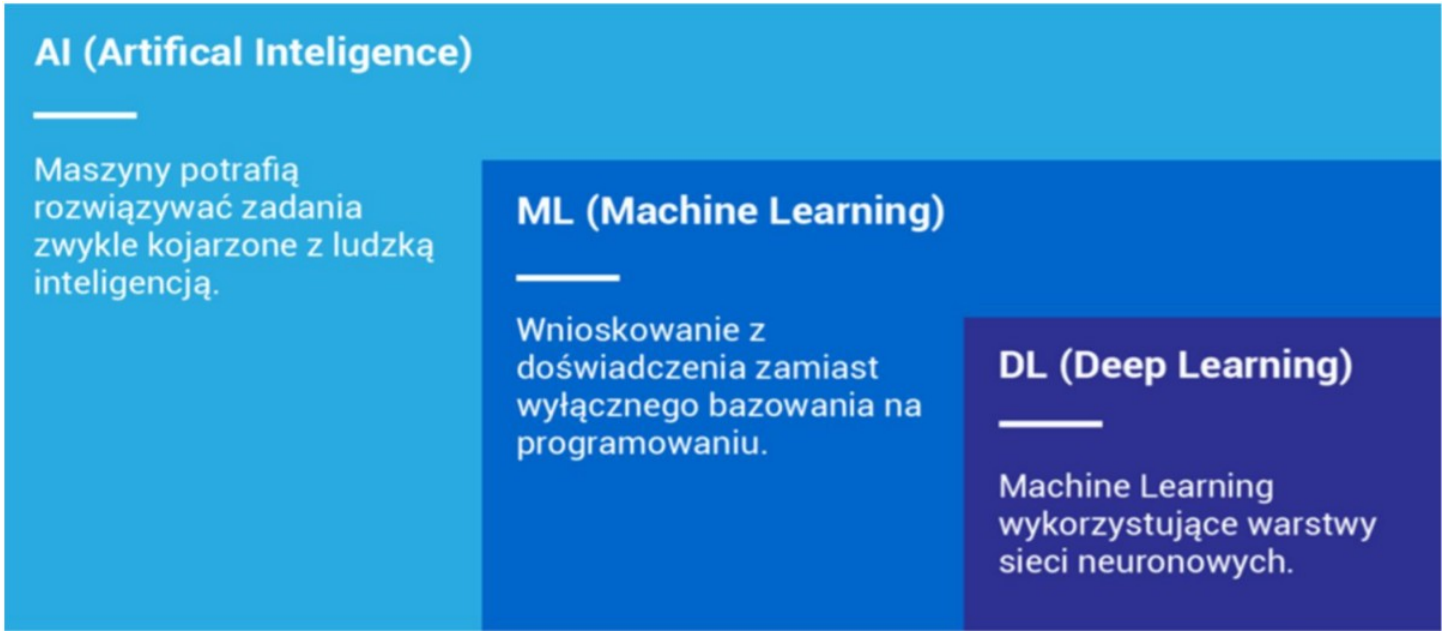
Przydatny materiał - korelacja

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

Zadania

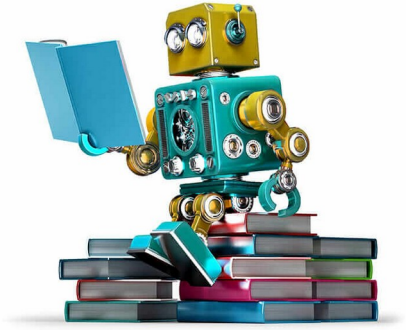
1. Stwórz dowolną tablicę a następnie oblicz dla niej średią i medianę
2. Ustandaryzuj dane w podanej tablicy
3. Stwórz 2 wektory a następnie policz dla niego statystyki opisowe i korelacje

Sztuczna inteligencja a uczenie maszynowe



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:

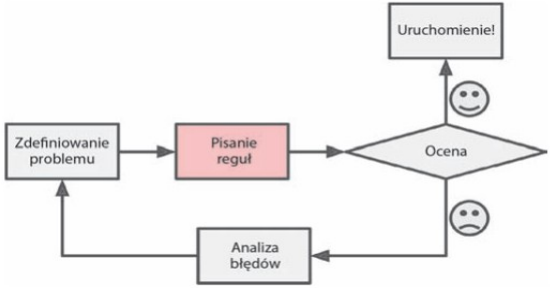
[Uczenie maszynowe to] dziedzina nauki dająca komputerom możliwość uczenia się bez konieczności ich jawnego programowania.

Czym jest uczenie maszynowe?

Podsumowując, by stworzyć model uczenia maszynowego potrzebujemy:

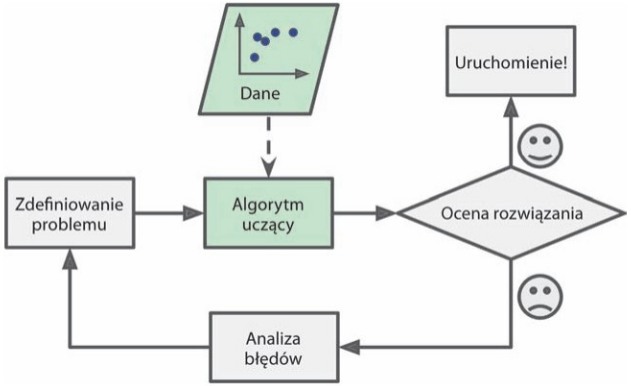
- zdefiniować zadanie, czyli co konkretnie nasz model ma robić, np. oznaczyć wiadomość e-mail jako spam lub oznaczyć jako nie-spam,
- dane, na podstawie których model nauczy się pewnych zależności lub różnic pomiędzy danymi, np. zbiór wiadomości oznaczonych jako spam lub nie-spam,
- funkcję oceny rezultatów,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

Czym różni się uczenie maszynowe od tradycyjnego programowania?



Utrzymanie tego modelu jest dużo prostsze – gdy spamerzy nauczą się unikać określonych zwrotów, wystarczy zasilić model powiększoną próbką danych, a on automatycznie zauważy 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?

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

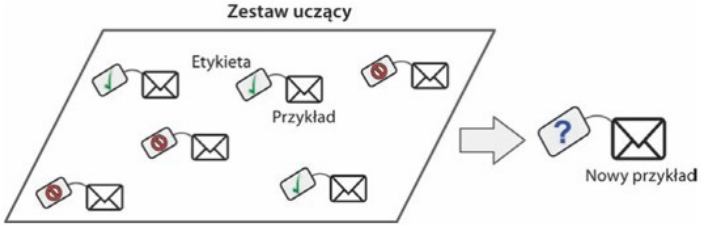
Uczenie maszynowe możemy podzielić na podstawie rodzaju nadzorowania procesu uczenia.

Wyróżniamy:

- Uczenie nadzorowane
- Uczenie nienadzorowane
- Uczenie przez wzmacnianie

Uczenie nadzorowane

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

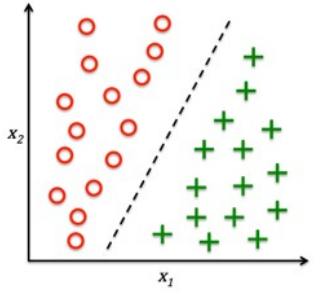


W analizowanym problemie tworzenia filtra spamu, by algorytm mógł się nauczyć różnic pomiędzy wiadomościami prawidłowymi i szkodliwymi, musimy zasilić go zbiorem e-maili z informacją, czy dana wiadomość była spamem, czy nie.

Uczenie nadzorowane

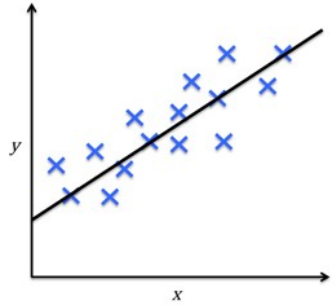
Klasyczne zadania uczenia nadzorowanego to **klasyfikacja** i **regresja**.

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 oczekiwać, 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

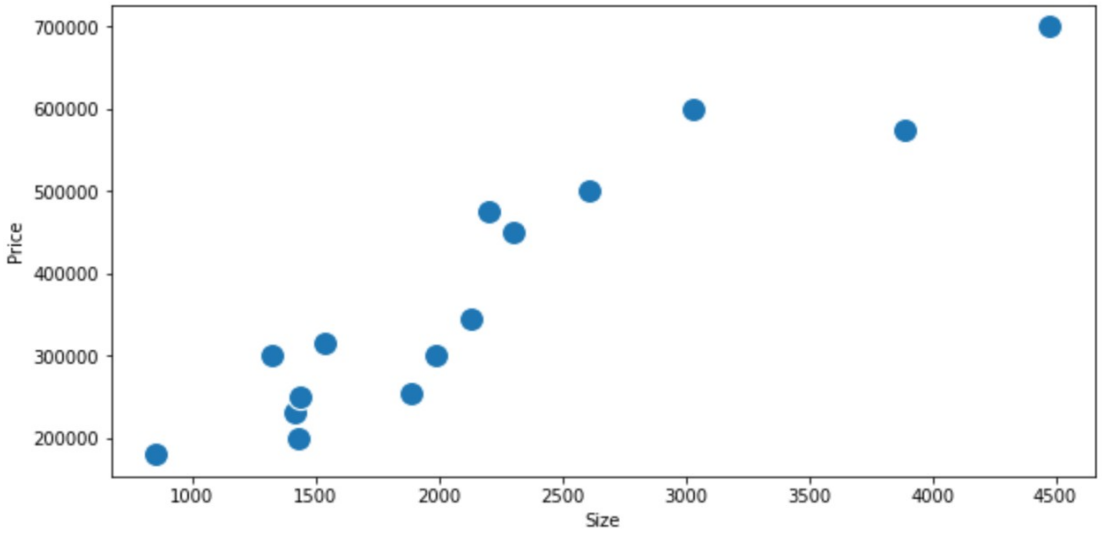
- **Twierdzenie o nieistnieniu darmowych obiadów** (ang. *No Free Lunch Theorem – NFL*)

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.**

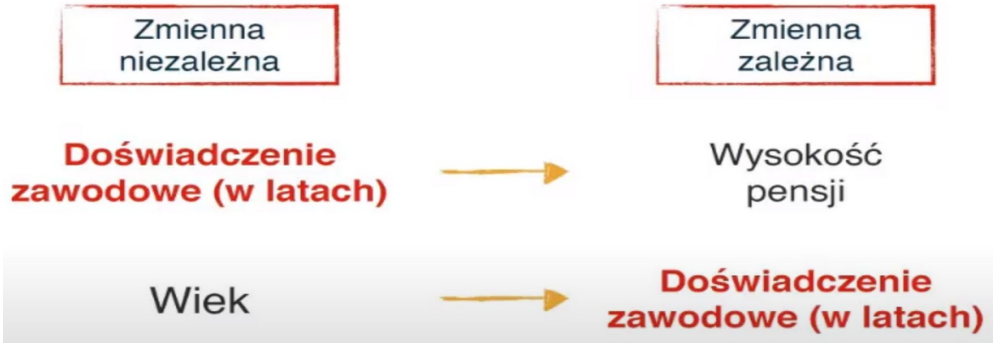
Size	Price
1427	198999
3031	599000
4478	699900
2200	475000
852	179900
1320	299900
1416	232000
1888	255000
1534	314900
2132	345000
1985	299900
2609	499998
2300	449900
3890	573900
1437	249900

Zmienne niezależne i zależne

Zmienne niezależne to te, które w danym badaniu manipulujemy, 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 nazewniczej 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

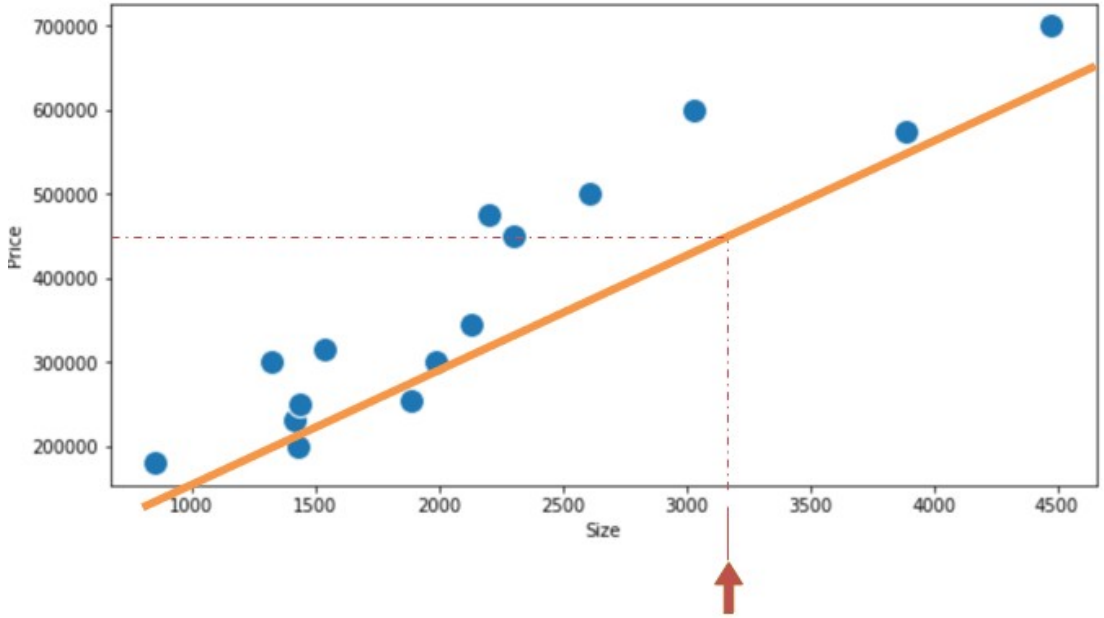
Różne typy zmiennych

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

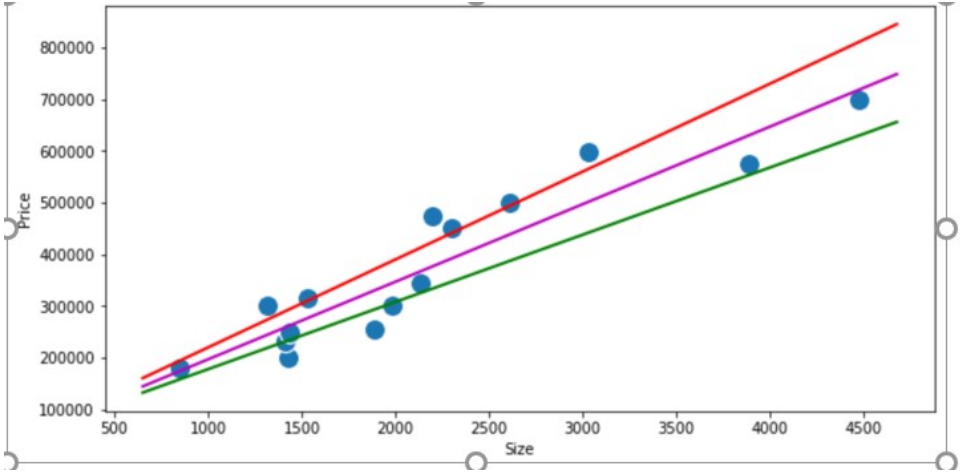
Regresja liniowa

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

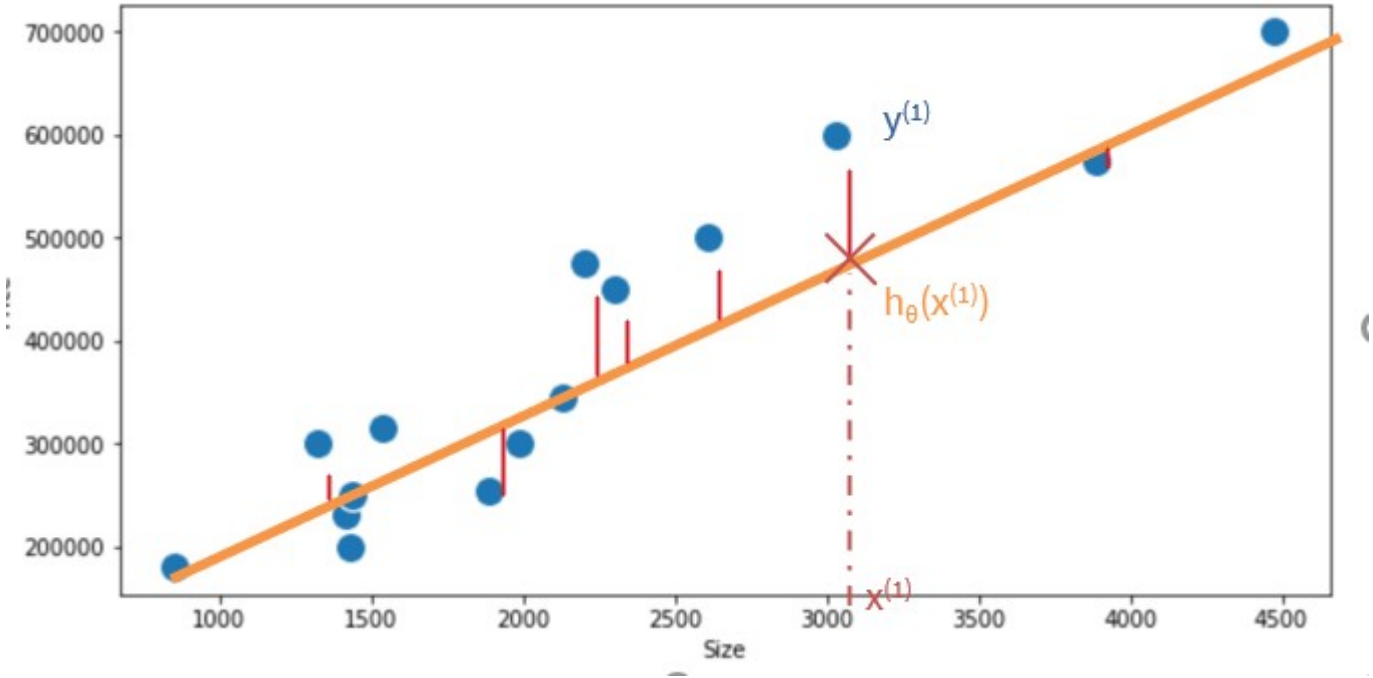


Regresja liniowa

Ale w jaki sposób tę linię poprowadzić? Przecież możemy zrobić to na wiele sposób, 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.**

Metody transformacji danych

- Standardyzacja
- Normalizacja
- Dyskretyzacja

Niedouczenie i przeuczenie modelu

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