Spis treści

[Definicja i zakres Data Science 1](#_Toc164634585)

[Ogólna charakterystyka Data Science 1](#_Toc164634586)

[Rola Data Science w rozwiązywaniu współczesnych problemów 2](#_Toc164634587)

[Zalety i wady Data Science 2](#_Toc164634588)

[Top Data Science skills 2](#_Toc164634589)

[Etapy procesu Data Science 3](#_Toc164634590)

[Krok 1. Zadaj pytania, aby sformułować problem biznesowy 3](#_Toc164634591)

[Krok 2. Zdobądź odpowiednie dane do analizy problemu 3](#_Toc164634592)

[Krok 3. Przeglądaj dane, aby wprowadzić poprawki błędów (eksploracja danych) 3](#_Toc164634593)

[Krok 4. Modeluj dane do szczegółowej analizy 4](#_Toc164634594)

[Krok 5. Przekaż wyniki analizy do prezentacji (wizualizacja danych) 4](#_Toc164634595)

[Podsumowanie codziennej pracy Data Scientist 4](#_Toc164634596)

[Modelowanie w Data Science 5](#_Toc164634597)

[Modelowanie predykcyjne = aproksymacja funkcji 5](#_Toc164634598)

[Klasyfikacyjne Modelowanie predykcyjne 5](#_Toc164634599)

[Regresyjne modelowanie predykcyjne 5](#_Toc164634600)

[Modelowanie wyjaśniające = identyfikacja istotnych zmiennych 6](#_Toc164634601)

[Dane w Data Science 6](#_Toc164634602)

[Uczące, walidacyjne i testowe zestawy danych 6](#_Toc164634603)

[W Machine Learning (bez Deep Learning) 7](#_Toc164634604)

[Zestaw uczący, inaczej treningowy, szkoleniowy (*training data, subset*..) 7](#_Toc164634605)

[Zestaw testujący (test data, subset…) 7](#_Toc164634606)

[W Deep Learning 8](#_Toc164634607)

[Zestaw uczący (train, learning set) 8](#_Toc164634608)

[Zestaw walidacyjny (*validation set)* 8](#_Toc164634609)

[Zestaw testowy (*test set*) 8](#_Toc164634610)

[Skalowanie i normalizacja danych 8](#_Toc164634611)

[Skalowanie 9](#_Toc164634612)

[Standaryzacja 9](#_Toc164634613)

[Normalizacja 9](#_Toc164634614)

[Statystyczne i matematyczne podstawy Data Science 10](#_Toc164634615)

[Statystyka 10](#_Toc164634616)

[Statystyka opisowa 10](#_Toc164634617)

[Rozkład normalny 10](#_Toc164634618)

[Miary Tendencji Centralnej 11](#_Toc164634619)

[Skośność i Kurtoza 11](#_Toc164634620)

[Zmienność 11](#_Toc164634621)

[Wnioskowanie statystyczne 12](#_Toc164634622)

[Twierdzenie centralne graniczne 12](#_Toc164634623)

[Testowanie hipotez 12](#_Toc164634624)

[ANOVA 12](#_Toc164634625)

[Analiza danych ilościowych 13](#_Toc164634626)

[Matematyka 13](#_Toc164634627)

[Algebra liniowa 13](#_Toc164634628)

[Rachunek różniczkowy 13](#_Toc164634629)

[Matematyka dyskretna i logika 14](#_Toc164634630)

[Modele regresji w data science 14](#_Toc164634631)

[Co to jest analiza regresji? 15](#_Toc164634632)

[Ile mamy technik regresji? 15](#_Toc164634633)

[Regresja liniowa 15](#_Toc164634634)

[Wielokrotna regresja liniowa 15](#_Toc164634635)

[Regresja wielokrotna Ważne punkty: 16](#_Toc164634636)

[Regresja logistyczna 16](#_Toc164634637)

[Ważne punkty: 17](#_Toc164634638)

[Regresja wielomianowa 18](#_Toc164634639)

[Ważne punkty: 18](#_Toc164634640)

[Wielokrotna regresja krokowa 18](#_Toc164634641)

[Regresja Ridge 19](#_Toc164634642)

[Ważne punkty: 19](#_Toc164634643)

[Regresja Lasso 19](#_Toc164634644)

[Ważne punkty: 20](#_Toc164634645)

[Inne techniki regresji 20](#_Toc164634646)

[Jak wybrać odpowiedni model regresji? 20](#_Toc164634647)

[Co to jest redukcja wymiarów? 21](#_Toc164634648)

[Analiza głównych składowych (PCA) 21](#_Toc164634649)

[Co to jest zasada analizy składowych (PCA)? 21](#_Toc164634650)

[Kiedy powinno się używać PCA? 22](#_Toc164634651)

[Jak działa zasada analizy składników (PCA)? 22](#_Toc164634652)

[Zanurkujmy do matematyki: 22](#_Toc164634653)

[Liniowa analiza dyskryminacyjna (LDA). 26](#_Toc164634654)

[Co to jest liniowa analiza dyskryminacyjna (LDA)? 26](#_Toc164634655)

[Jak działa liniowa analiza dyskryminacyjna (LDA)? 26](#_Toc164634656)

[Algorytm K- Najbliżsi Sąsiedzi (*KNN*) 27](#_Toc164634657)

[Jak algorytm KNN pracuje? 27](#_Toc164634658)

[Jak wybrać czynnik K? 27](#_Toc164634659)

[Wskaźnik błędu uczenia i wskaźnik błędu walidacji 28](#_Toc164634660)

[Algorytm Maszyny Wektorów Nośnych 28](#_Toc164634661)

[Co to jest Maszyna Wektora Nośnego? 28](#_Toc164634662)

[Hiperpłaszczyzny i wektory nośne 29](#_Toc164634663)

[Funkcje jądra (kernel) 30](#_Toc164634664)

[Twierdzenie Bayes’a – jeden z filarów prognozowania w Data Science 30](#_Toc164634665)

[Wnioskowanie Bayes’a dla Data Science 30](#_Toc164634666)

[Formuła twierdzenia Bayes’a 31](#_Toc164634667)

[Jak działa klasyfikator Naive Bayes? 31](#_Toc164634668)

[Pierwsze podejście (w przypadku jednej funkcji) 31](#_Toc164634669)

[Przykład zastosowania Twierdzenia Bayes’a 33](#_Toc164634670)

[Twierdzenie Naiwnego Bayesa 33](#_Toc164634671)

[Twierdzenie Naiwnego Bayesa c.d. 34](#_Toc164634672)

[Zastosowania twierdzenia Bayesa 35](#_Toc164634673)

[1. Filtrowanie spamu 35](#_Toc164634674)

[2. Analiza sentymentów 35](#_Toc164634675)

[3. Systemy rekomendacji 35](#_Toc164634676)

[4. Bayesowskie sieci neuronowe 35](#_Toc164634677)

[Miary błędów predykcji (regresja) 36](#_Toc164634678)

[Błędy absolutne 36](#_Toc164634679)

[Inne miary błędów 37](#_Toc164634680)

[Błąd R-kwadrat 37](#_Toc164634681)

[Walidacja i walidacja krzyżowa do poprawy wydajności modeli w data science 38](#_Toc164634682)

[Problem 39](#_Toc164634683)

[Co to jest walidacja krzyżowa? 39](#_Toc164634684)

[Walidacja krzyżowa K-fold 39](#_Toc164634685)

[Stratyfikacja K-Fold Cross Validation 40](#_Toc164634686)

[Leave one out cross validation — LOOCV 40](#_Toc164634687)

[Leave-P-Out Cross Validation - uogólnienie LOOCV 41](#_Toc164634688)

[Walidacja krzyżowa szeregów czasowych 41](#_Toc164634689)

# Definicja i zakres Data Science

Obraz zawierający tekst, krąg, Czcionka, zrzut ekranu

Opis wygenerowany automatycznie„Data Science dotyczy wydobywania, przygotowywania, analizy, wizualizacji i utrzymywania informacji.

Jest to dziedzina interdyscyplinarna, która wykorzystuje metody i procesy naukowe do uzyskiwania wglądu w dane. ”

Obraz zawierający tekst, zrzut ekranu, krąg, diagram

Opis wygenerowany automatycznie

# Ogólna charakterystyka Data Science

**Główną cechą Data Science jest znalezienie wzorców nowej wiedzy w danych.**

Wykorzystuje różne techniki statystyczne do analizy i wyciągania wniosków z danych. Od ekstrakcji danych, „czyszczenia” i przetwarzania wstępnego, Data Scientist musi dokładnie zbadać dane. Następnie jest odpowiedzialny za dokonywanie **prognoz, czy wyjaśnień na podstawie danych**.

**Celem Badacza Danych** jest wyciągnięcie wniosków z danych. Dzięki tym wnioskom jest on w stanie pomóc firmom w podejmowaniu mądrzejszych decyzji biznesowych.

Wraz z pojawieniem się nowych technologii nastąpił wykładniczy wzrost **zasobów danych.**

Stworzyło to okazję do analizy i uzyskania **istotnej wiedzy z danych**.

Wymaga to specjalistycznej wiedzy „Data Scientist”, który może korzystać z różnych **narzędzi statystycznych i uczenia maszynowego** w celu zrozumienia i analizy danych.

Badacz danych nie tylko analizuje dane, ale także **wykorzystuje algorytmy uczenia maszynowego do przewidywania przyszłych zdarzeń**.

Dlatego możemy rozumieć Data Science jako dziedzinę zajmującą się **przetwarzaniem danych, analizą i ekstrakcją wiedzy z danych** przy użyciu różnych metod statystycznych i algorytmów komputerowych.

Jest to dziedzina **multidyscyplinarna**, która łączy matematykę, statystykę i informatykę.

# Rola Data Science w rozwiązywaniu współczesnych problemów

Firmy potrzebują danych **do funkcjonowania, rozwoju i ulepszania zarządzania.**  
Badacze danych zajmują się danymi, aby pomóc firmom **w podejmowaniu właściwych decyzji**.

Naukowcy danych zwykle działają jako konsultanci zatrudnieni przez firmy, w których **uczestniczą w różnych procesach decyzyjnych i tworzeniu strategii**.

Podejście oparte na analizie danych jest coraz powszechniej przyjmowane przez firmy.   
Wiedza wynikająca z analizy danych będzie pomocne dla firm, które chcą **analizować siebie i swoje wyniki na rynku.**

Inne niż branże komercyjne, np. **opieki zdrowotnej** wykorzystują również Data Science, np. technologia ta jest bardzo potrzebna do rozpoznania mikroskopijnych guzów i deformacji na wczesnym etapie diagnozy. Ponadto **obserwacje satelitarne środowiska i Ziemi** są obszarem rozwoju zastosowania Data Science.

# Zalety i wady Data Science

Zalety:

1. Pożądana na rynku
2. Wielość specjalności i pozycji
3. Dobre zarobki
4. Wysoki prestiż zawodowy
5. Wszechstronność

Wady (Niedogodności):

1. Jest bardzo rozmytym pojęciowo terminem
2. Mistrzostwo omal niemożliwe
3. Potrzebna obszerna i wielodiscyplinarna wiedza
4. Arbitralne dane mogą dać nieoczekiwane wyniki
5. Problemy z własnością danych

# Top Data Science skills

Obraz zawierający tekst, zrzut ekranu, Reklama internetowa, Strona internetowa

Opis wygenerowany automatycznie

1. Umiejętności praktyczne, takie jak matematyka i umiejętności statystyczne
2. Umiejętności kodowania
3. Umiejętności miękkie, takie jak umiejętności społeczne, wizualizacja danych, prezentacja i komunikacja
4. Umiejętności biznesowe

# Etapy procesu Data Science

## Obraz zawierający tekst, zrzut ekranu, Czcionka, Równolegle Opis wygenerowany automatycznieKrok 1. Zadaj pytania, aby sformułować problem biznesowy

W pierwszym kroku spróbuj dowiedzieć się, jakie są potrzeby firmy i wyodrębnij dane na tej podstawie. Proces nauki danych zaczynasz od zadawania właściwych pytań, aby znaleźć problem.

Weźmy bardzo powszechny problem firmy produkcyjnej - problem ze sprzedażą. Aby przeanalizować problem, musisz zacząć od zadania wielu pytań:

* *Jaki jest rynek docelowy i aktualni klienci firmy?*
* *Jak podchodzisz do rynku docelowego?*
* *Jak obecnie wygląda proces sprzedaży?*
* *Jakie masz informacje na temat rynku docelowego?*
* *Jak możemy zidentyfikować klientów, którzy chętniej kupią nasz produkt?*

Po dyskusji z zespołem marketingowym postanawiasz **skoncentrować się na problemie: „Jak możemy zidentyfikować potencjalnych klientów, którzy chętniej kupią nasz produkt?”**

## Krok 2. Zdobądź odpowiednie dane do analizy problemu

Teraz, gdy wiesz już o swoim problemie biznesowym, czas zebrać dane, które pomogą Ci rozwiązać problem. W wielu przypadkach możesz uzyskać zbiory danych w firmie, np. z raportów lub wcześniej zebrane podczas innych badań.

Jeśli uważasz, że dostępne dane nie są wystarczające, musisz **podjąć kroki w celu zebrania nowych danych**. Możesz nawet uzyskiwać informacje zwrotne od odwiedzających i klientów, wyświetlając lub rozpowszechniając formularz opinii.

Zebrane dane to tak naprawdę „**surowe dane**”, które zawierają błędy i brakujące wartości. Więc zanim przeanalizujesz dane, musisz je wyczyścić (wrangle).

## Krok 3. Przeglądaj dane, aby wprowadzić poprawki błędów (eksploracja danych)

**Eksploracja danych** polega na ich czyszczeniu i porządkowaniu.

Pomimo zebrania wszystkich danych, nie jesteś gotowy do ich użycia, ponieważ często zebrane nieprzetworzone dane prawdopodobnie zawierają **osobliwości**. **Po pierwsze, musisz upewnić się, że dane są czyste i wolne od błędów**. To najważniejszy krok w procesie, który wymaga cierpliwości i skupienia. W tym celu wykorzystywane są różne narzędzia i techniki, takie jak Python, R, SQL itp.

Następnie zaczniesz odpowiadać na następujące pytania:

* *Czy w danych brakuje wartości, np. Czy są klienci bez ich numerów kontaktowych?*
* *Czy są jakieś nieprawidłowe wartości? Jeśli tak, to jak to naprawić?*
* *Czy istnieje wiele zestawów danych? Czy scalanie zestawów danych to dobry wybór? Jeśli tak, to jak je połączyć?*

Po wykryciu brakujących i fałszywych wartości dane są gotowe do analizy. **Uzyskanie błędnych danych jest gorsze niż brak w ogóle danych.**

## Krok 4. Modeluj dane do szczegółowej analizy

Po zbadaniu danych masz wystarczającą ilość informacji, aby stworzyć model, aby odpowiedzieć na pytanie: **„Jak możemy zidentyfikować potencjalnych klientów, którzy są bardziej skłonni do zakupu naszego produktu?”**

W tym kroku analizujesz dane, aby uzyskać z nich informacje. **Analiza danych wymaga zastosowania różnych algorytmów, które wydobędą z nich znaczenie:**

* *Zbuduj model danych, aby odpowiedzieć na pytanie.*
* *Sprawdź poprawność modelu na podstawie zebranych danych.*
* *Wykorzystanie różnych narzędzi wizualizacji do prezentacji danych.*
* *Wykonaj niezbędne algorytmy i analizę statystyczną.*
* *Porównaj wyniki z innymi technikami i źródłami.*

Jednak udzielenie odpowiedzi na te pytania da jedynie wskazówki i hipotezy.

**Modelowanie danych** to sposób na przybliżenie danych za pomocą odpowiedniego równania zrozumiałego dla maszyny. **Powinieneś być w stanie przewidywać na podstawie modelu. Być może będziesz musiał wypróbować kilka modeli, aby znaleźć najlepsze dopasowanie.**

Wracając do problemu sprzedaży, ten model pomoże Ci przewidzieć, którzy klienci są bardziej skłonni do zakupu. Prognozy mogą być specyficzne, np. Kobiety w wieku 16–36 lat mieszkające w Indiach.

## Krok 5. Przekaż wyniki analizy do prezentacji (wizualizacja danych)

**Umiejętności komunikacyjne są ważną częścią pracy informatyka, ale również bardzo niedocenianą.** Będzie to w rzeczywistości bardzo trudna część pracy, ponieważ wiąże się z przedstawieniem wyników badań opinii publicznej i innym członkom zespołu w sposób łatwo zrozumiały dla nich.

Musisz skutecznie komunikować o wynikach wcześniej określonego problemu:

* *Utwórz grafikę wyników badania (zwizualizuj) do prezentacji za pomocą narzędzi - R, Python, Tableau, Excel.*
* *Użyj „opowiadania historii” (narracji), aby dopasować wyniki.*
* *Odpowiedz na różne pytania uzupełniające.*
* *Prezentuj dane w różnych formatach - raporty, strony internetowe.*
* *Odpowiedzi często wywołają więcej pytań, i proces rozpocznie się od nowa.*

## Podsumowanie codziennej pracy Data Scientist

To było spojrzenie na dzień w pracy badacza danych i jego zadania. Konkretne zadania obejmują:

* *Identyfikacja problemów analitycznych związanych z danymi, które oferują duże możliwości dla organizacji.*
* *Zbieranie dużych zestawów danych ustrukturyzowanych i nieustrukturyzowanych z różnych źródeł.*
* *Określanie prawidłowych zestawów danych i zmiennych.*
* *Czyszczenie i usuwanie błędów z danych w celu zapewnienia dokładności i kompletności.*
* *Tworzenie i stosowanie modeli, algorytmów i technik do eksploracji zasobów dużych zbiorów danych.*
* *Analizowanie danych w celu wykrycia ukrytych wzorców i trendów.*
* *Interpretacja danych w celu odkrycia rozwiązań i możliwości oraz podejmowanie na ich podstawie decyzji*
* *Przekazywanie wyników kierownikom i innym osobom przy użyciu wizualizacji i innych środków.*

# Modelowanie w Data Science

## Modelowanie predykcyjne = aproksymacja funkcji

**Modelowanie predykcyjne** to problem polegający na opracowaniu modelu z wykorzystaniem danych historycznych w celu przewidywania nowych danych, na które nie mamy odpowiedzi.

Modelowanie predykcyjne można opisać jako **matematyczny problem aproksymacji funkcji odwzorowania (f) od zmiennych wejściowych (X) do zmiennych wyjściowych (y)**.

**Algorytm modelowania** ma za zadanie znaleźć najlepszą funkcję (f), jaką możemy uzyskać, biorąc pod uwagę dostępny czas i zasoby.

Zasadniczo wszystkie zadania aproksymacji funkcji są podzielone na:

1. zadania klasyfikacyjne
2. regresyjne.

### Klasyfikacyjne Modelowanie predykcyjne

**Modelowanie predykcyjne klasyfikacji jest zadaniem aproksymacji funkcji (f) od zmiennych wejściowych (X) do dyskretnych zmiennych wyjściowych (y).**

Zmienne wyjściowe są często nazywane **klasami** lub **kategoriami**. Funkcja przewiduje klasę lub kategorię dla danej obserwacji. Zadanie z klasyfikacją wymaga, aby dane zaklasyfikować do jednej, dwóch lub więcej klas. Klasyfikacja może zawierać zmienne wejściowe o **wartościach rzeczywistych lub dyskretnych**.

Zadanie z **dwiema klasami** jest często nazywane klasyfikacją **dwuklasową** lub **binarną**.   
Zadanie z **więcej** niż dwiema klasami jest często nazywane klasyfikacją **wieloklasową**.

Modele klasyfikacji często przewidują **wartość ciągłą jako prawdopodobieństwo, że dana wartość należy do określonej klasy.**

Prawdopodobieństwa można interpretować jako **stopień pewności przynależności danej wartości do określonej klasy.** Przewidywane prawdopodobieństwo można przeliczyć na wartość klasy i **wybrać klasę o najwyższym prawdopodobieństwie.**

Istnieje wiele sposobów **oszacowania skuteczności modelu predykcyjnego klasyfikacji**, ale być może najczęstszym jest obliczenie **dokładności (*accuracy*) klasyfikacji**.

**Dokładność (*accuracy*) klasyfikacji to procent poprawnie sklasyfikowanych wartości we wszystkich dokonanych działaniach klasyfikacyjnych.**

### Regresyjne modelowanie predykcyjne

**Modelowanie predykcyjne regresji polega na aproksymacji funkcji (f) ze zmiennych wejściowych (X) do ciągłej zmiennej wyjściowej (y).**

Ciągłe zmienne wyjściowe są często wielkościami, takimi jak **parametry**.   
Zadanie regresji wymaga **predykcji danych ilościowych**.  
Regresja może mieć zmienne **wejściowe** o wartościach rzeczywistych lub dyskretnych.

Zadanie z wieloma zmiennymi wejściowymi jest często nazywany zadaniem **regresji** **wielowymiarowej**.

Zadanie regresji, w którym zmienne wejściowe są uporządkowane według czasu, nazywa się zadaniem **prognozowania szeregów czasowych.**

Ponieważ model predykcyjny regresji przewiduje określoną wielkość (wartość predykcji), **skuteczność modelu należy traktować jako błąd w przewidywaniach.**

Istnieje wiele sposobów na oszacowanie skuteczności modelu predykcyjnego regresji, ale być może najczęstszym jest **obliczenie średniego błędu kwadratu pierwiastkowego,** skróconego akronimem **RMSE**.

Zaletą RMSE jest to, że jednostki wyniku błędu znajdują się w tych samych jednostkach, co przewidywana wartość.

## Modelowanie wyjaśniające = identyfikacja istotnych zmiennych

W modelowaniu objaśniającym jesteśmy zainteresowani **identyfikacją zmiennych, które mają naukowo znaczący i statystycznie istotny związek z wynikiem.**

Podstawowym **celem jest przetestowanie hipotez teoretycznych**, aby położyć nacisk zarówno na znaczące teoretycznie relacje, jak i na ustalenie, czy każda relacja jest istotna statystycznie ( predyktory mają wartości p mniejsze niż 0,05). Modele oparte są na statystycznej analizie regresji, np. **Modelowanie Równań Strukturalnych** (**SEM**).

Niektóre etapy modelowania objaśniającego obejmują:

* dopasowanie potencjalnie ważnych teoretycznie zmiennych (predyktorów),
* sprawdzanie istotności statystycznej,
* ocena efektów,
* prowadzenie diagnostyki dopasowania całego modelu.

# Dane w Data Science

## Uczące, walidacyjne i testowe zestawy danych

**Dlaczego dane są dzielone?**

Tworzenie modelu uczenia maszynowego polega na stworzeniu programu, który jest w stanie wykonać **uogólnień danych wejściowych** na dane predykcyjne, których nigdy wcześniej nie widział. To zadanie wymaga stworzenia modelu - podczas uczenia - na pewnej liczbie danych wejściowych, co prawdopodobnie doprowadzi do wystarczającej dokładności modelu. Obejmuje to wiele kroków, przez które model musi przejść, zanim będzie dostępny do użycia:

**Krok 1:** Wykonanie modelu badającego dane. Uczenie modelu na większej części danych.

**Krok 2:** Sprawdzenie, czy model nie jest źle dopasowany: **nadmiernie** (ovefitting), **za słabo** (underfitting). Uczenie modelu na mniejszej części danych.

**Krok 3:** Wyciągnięcie wniosku na temat wydajności (dokładności) modelu. Sprawdzenie modelu na danych zewnętrznych, nie biorących udziału w uczeniu modelu.

## W Machine Learning (bez Deep Learning)

### Zestaw uczący, inaczej treningowy, szkoleniowy (*training data, subset*..)

**Obraz zawierający tekst, zrzut ekranu, Czcionka, linia

Opis wygenerowany automatycznieDane treningowe to największy (w wielkości) podzbiór oryginalnego zestawu danych, który jest używany do trenowania (dopasowania) modelu uczenia maszynowego.** Po pierwsze, dane treningowe są przekazywane do algorytmów **ML**, co pozwala im nauczyć się przewidywać dane zadanie. Dane treningowe różnią się w zależności od tego, czy używamy algorytmów uczenia **nadzorowanego**, czy **nienadzorowanego**.

W przypadku **uczenia nienadzorowanego** dane treningowe zawierają nieoznakowane punkty danych, tzn. dane wejścia (input) nie są oznaczone odpowiednimi danymi wyjścia (output).. Modele są wymagane, aby znaleźć wzorce z danych treningowych w celu przewidywania.

W przypadku **uczenia nadzorowanego** dane treningowe zawierają etykiety wejścia (input) umożliwiające kontrolę trenowania modelu i przewidywania danych wyjścia (output).

Rodzaj danych treningowych, które dostarczamy do modelu, jest w dużym stopniu odpowiedzialny za dokładność i zdolność przewidywania modelu. Oznacza to, że im lepsza jakość danych treningowych, tym lepsza będzie wydajność modelu. Dane treningowe z reguły są w przybliżeniu większe lub równe 70% wszystkich danych dla projektu uczenia maszynowego.

### Zestaw testujący (test data, subset…)

Obraz zawierający tekst, zrzut ekranu, Czcionka, linia

Opis wygenerowany automatyczniePo wytrenowaniu modelu z zestawem danych treningowych nadszedł czas, aby przetestować model za pomocą **testowego zestawu danych**. Ten zestaw danych ocenia wydajność modelu i zapewnia, że model może być dobrze uogólniony z nowym lub niewidocznym zestawem danych.

**Zestaw danych testowych** jest kolejnym podzbiorem oryginalnych danych, który jest niezależny od zestawu danych treningowych. Ponieważ pochodzi z danych oryginalnych na zasadzie losowego wydzielenia, ma jednak kilka podobnych typów funkcji i rozkładu prawdopodobieństwa klasy. Używa się go jako punktu odniesienia do oceny modelu po zakończeniu szkolenia modelu.

**Dane testowe** dobrze spełniają swoją rolę, gdy zawierają dane dla każdego typu scenariusza dla danego problemu, z którym model musiałby się zmierzyć, gdyby był używany w świecie rzeczywistym. Zwykle zestaw danych testowych stanowi około 20–30% wszystkich oryginalnych danych dla projektu uczenia maszynowego. Nie może być zbyt duży (zbyt podobny do zestawu treningowego), ani zbyt mały (prawdopodobnie za bardzo różny od zbioru uczącego).

Na tym etapie możemy również sprawdzić i porównać **dokładność (accuracy**) testowania z dokładnością treningu, co oznacza, jak dokładny jest nasz model z zestawem danych testowych w porównaniu z zestawem danych treningowych. Jeśli dokładność modelu na danych testowych jest równa lub większa niż na danych treningowych, mówi się, że model ma **nadmierne dopasowanie** (overfitting).Gdy natomiast jest bardzo mała w stosunku do dokładności modelu na danych uczących, model jest **niedopasowany** (underfitting).

## **Obraz zawierający zrzut ekranu, tekst, linia, diagram Opis wygenerowany automatycznie**W Deep Learning

### Zestaw uczący (train, learning set)

Ten zestaw danych odpowiada krokowi 1 w poprzedniej sekcji. Zawiera zestaw danych wejściowych, na których model będzie uczony - poprzez dostosowanie parametrów (tj. wag w kontekście sieci neuronowych).

### Zestaw walidacyjny (*validation set)*

Aby model został nauczony, należy go okresowo oceniać (krok 2) i właśnie do tego służy zestaw sprawdzania poprawności (walidacji). Jest on też uczony tak samo jak zestaw Uczący. Dane miedzy oboma mogą być wymieniane. Obliczając stratę (tj. poziom błędu), jaki daje model na podstawie zestawu walidacyjnego w dowolnym punkcie, możemy wiedzieć, jak dokładny jest model. To jest istota uczenia. Następnie model dostosuje parametry modelu (hiperparametry, np liczba ukrytych warstw w sieci neuronowej) w oparciu o częste wyniki oceny w zestawie walidacyjnym. Zestaw walidacyjny danych można wykorzystać do doskonalenia za pomocą walidacji krzyżowej, regularyzacji modelu oraz do „wcześniejszego zatrzymania”: zatrzymanie uczenia się modelu, gdy zwiększy się błąd w zestawie walidacyjnym danych, co jest oznaką nadmiernego dopasowania (overfitting) do zbioru danych szkolenia. Z reguły zawiera 15-25% danych oryginalnych.

### Zestaw testowy (*test set*)

Dane tego zestawu nie biorą udziału w uczeniu modelu. Są danymi zewnętrznymi. Odpowiada to końcowej ocenie, którą przechodzi model po zakończeniu fazy uczenia (z wykorzystaniem zestawów uczenia i walidacji). Ten krok ma kluczowe znaczenie dla przetestowania uogólnienia modelu (Krok 3). Korzystając z tego zestawu, możemy uzyskać dokładność działania naszego modelu. Testowy zestaw danych to zbiór danych, który jest niezależny od zestawu danych uczących, ale ma taki sam rozkład prawdopodobieństwa, jak zestaw danych uczących, dotyczy tego samego problemu. Zestaw testowy jest zatem zestawem danych używanych tylko do oceny wydajności w pełni określonego (zestawami uczącym i walidacyjnym) modelu.

# Skalowanie i normalizacja danych

Skalowanie danych jest najważniejszą częścią wstępnego przetwarzania danych. Jeśli zobaczymy nasz zestaw danych, wówczas jakiś atrybut (cecha, zmienna) zawiera informacje w wartości liczbowej, niektóre wartości są bardzo wysokie, a niektóre są bardzo niskie, jeśli np. widzimy wiek i szacunkowe wynagrodzenie. Spowoduje to, że w naszym modelu wszystkie wartości chcemy mieć w tej samej skali.

## Skalowanie

Obraz zawierający tekst, diagram, Wykres, linia

Opis wygenerowany automatycznieW **skalowaniu** (zwanym także **skalowaniem Min Max, skalowaniem automatycznym**) dane są przekształcane w taki sposób, że funkcje znajdują się w określonym zakresie, np. [0, 1].

gdzie x’ jest znormalizowaną (zeskalowaną) wartością.

Obraz zawierający tekst, Czcionka, linia, zrzut ekranu

Opis wygenerowany automatycznie

## Standaryzacja

Obraz zawierający Czcionka, linia, zrzut ekranu, tekst

Opis wygenerowany automatycznieStandaryzacja (zwana również normalizacją z-score) przekształca dane w taki sposób, że wynikowy rozkład ma średnią 0 i odchylenie standardowe równe 1.   
gdzie x jest oryginalnym wektorem cech, xmean jest średnią tego wektora cech, a σ jest jego odchyleniem standardowym.

## Normalizacja

Obraz zawierający diagram, Wykres, tekst, design

Opis wygenerowany automatycznieCelem normalizacji jest zmiana twoich danych, aby można je było opisać jako **rozkład normalny**.

**Rozkład normalny (rozkład Gaussa**), znany również jako krzywa dzwonowa, jest rozkładem statystycznym, w którym mniej więcej równe obserwacje spadają powyżej i poniżej średniej, średnia i mediana są takie same, a więcej obserwacji znajduje się bliżej średniej.

# Statystyczne i matematyczne podstawy Data Science

## Statystyka

Data Science to nowsza terminologia. Chociaż istnieje od wielu dziesięcioleci, jego oficjalna nazwa brzmiała „Statystyka”, a Data Scientist był znany jako „Statystyk”.



**, Calculus**

Nauka o danych opiera się na statystykach. Statystyki zasilają narzędzia Data Science i dają im możliwość przetwarzania wszystkich nadchodzących informacji (danych).

Ogólnie statystykę można podzielić na dwie grupy:

* **Statystykę opisową**
* **Wnioskowanie statystyczne**

## Statystyka opisowa

Statystyka opisowa dotyczy opisu danych. Pomaga zrozumieć dane. Zajmuje się ilościowym podsumowaniem danych za pomocą **reprezentacji numerycznych** lub **wykresów**. Niektóre pojęcia potrzebne do zrozumienia statystyki opisowej to: **Rozkład normalny**, **Tendencja centralna**, **Kurtoza** i **Zmienność** .

### Rozkład normalny

Obraz zawierający Wykres, diagram, tekst, stok

Opis wygenerowany automatycznie**Rozkład normalny**, znany również jako **rozkład Gaussa**, jest reprezentacją dużych próbek danych na wykresie. Jest to rozkład wartości zmiennej za pomocą funkcji prawdopodobieństwa.

W rozkładzie normalnym istnieje symetryczna krzywa w kształcie dzwonu, w której obserwacje skupiają się na centralnym szczycie, gdzie reprezentują średnią.

Gdy wartości oddalają się od średniej, zmniejszają się jednakowo zarówno w kierunku lewym, jak i prawym. **W celu przeprowadzenia wnioskowania statystycznego konieczny jest normalny rozkład danych.**

### Miary Tendencji Centralnej

Tendencja centralna lub miara Tendencji ***centralnej identyfikuje centralny punkt w obrębie danej próbki danych***. Istnieją trzy części tendencji centralnej - średnia, mediana i moda (najczęstsza wartość). Podobnie jak omawiany powyżej rozkład normalny, **średnia** występuje w centralnym punkcie przykładowych danych. Jest on znany jako średnia arytmetyczna wszystkich wartości danych. Jest to suma wszystkich wartości danych podzielona przez liczbę wartości danych. Jest napisane jako: https://d2h0cx97tjks2p.cloudfront.net/blogs/wp-content/uploads/sites/2/2019/05/Central-Tendency-Formula.png

Inną miarą tendencji centralnej jest **mediana**. Mediana to środkowa wartość danych ułożonych w porządku rosnącym. Łatwiej jest znaleźć środkową wartość z zestawu liczb nieparzystych, ale aby znaleźć środkową wartość z zestawu liczb parzystych, bierzemy średnią z dwóch średnich wyników i przyjmujemy wartość wyjściową jako medianę.

Trzecią miarą tendencji centralnej jest **moda.** Moda jest najczęściej występującą wartością w danej próbce danych.

### Skośność i Kurtoza

Obraz zawierający tekst, diagram, linia, Wykres

Opis wygenerowany automatycznie**Skośność** *mierzy brak symetrii w próbce danych*. Dane symetryczne występują w postaci rozkładu normalnego, w którym szlak jest równomiernie rozłożony na lewą i prawą stronę średniej. **Skośność wynosi zero dla normalnie rozłożonej próbki danych.** **Dodatnie** pochylenie jest ułatwione przez agregację danych po lewej stronie, natomiast **ujemne** pochylenie występuje, gdy dane są ułożone w stos po prawej stronie.

Z drugiej strony, **Kurtoza** *jest miarą „dopasowania” rozkładu prawdopodobieństwa*. Mierzy, czy dane są **gruboogonowe**, czy **lekkoogoniaste** w stosunku do centralnej lokalizacji rozkładu. W przypadku wysokiej kurtozy zbiory danych są „ciężkie”, podczas gdy w przypadku niskiej kurtozy zbiory danych mają „lekkie ogony”.

### Zmienność

Obraz zawierający tekst, linia, diagram, Wykres

Opis wygenerowany automatycznie**Zmienność** *jest miarą, która określa, jak daleko jest punkt danych od środkowej średniej rozkładu*. Mierzy również wielkość odległości między tymi punktami danych. Miarami zmienności są **zakres**, **wariancja**, **odchylenie standardowe** i **zakres międzykwartylowy**.

**Zakres** jest miarą różnicy między największą a najmniejszą wartością próbki danych.

**Wariancja** *jest różnicą kwadratowych wartości od średniej*, a **odchylenie standardowe** *jest pierwiastkiem kwadratowym wariancji.*

## Wnioskowanie statystyczne

Podczas gdy statystyki opisowe dotyczą opisu danych, wnioskowanie statystyczne dotyczy wnioskowania na podstawie danych. Ogólnie rzecz biorąc, statystyki wnioskowania dotyczą ***wnioskowania na temat populacji za pomocą mniejszej próby.***

### Twierdzenie centralne graniczne

Obraz zawierający tekst, diagram, Czcionka, Jaskrawoniebieski

Opis wygenerowany automatycznieZgodnie z centralnym twierdzeniem granicznym ***średnia próbki jest taka sama jak dla całej populacji***. Oznacza to również, że odchylenie standardowe próbki będzie równe odchyleniu standardowemu populacji. I wreszcie, wraz **ze wzrostem wielkości** **próby**, standardowe błędy będą niższe, co spowoduje bardziej normalną krzywą kształtu. Pomoże to również w dokładniejszym określeniu średniej populacji.

Jedną z koncepcji ważnych w centralnym twierdzeniu o granicy jest „**przedział ufności**”. Przedział ufności ***jest miarą oszacowania średniej populacji***. Przedział ufności otacza średnią próbki. Proces konstruowania przedziału wymaga dodania ***marginesu błędu***. Możesz obliczyć margines błędu, mnożąc błąd standardowy średniej przez wynik **Z-odsetka poziomu ufności.**

## Testowanie hipotez

Testowanie hipotez jest ***miarą testowania założenia***. Służy do wnioskowania o wynikach hipotezy przeprowadzonej na grupie próbek wybranych z większej grupy lub populacji.

Hipoteza, którą musimy przetestować, nazywa się **hipotezą zerową**, a hipoteza, na podstawie której musimy ją przetestować, jest znana jako **hipoteza alternatywna**. Hipoteza zerowa jest często idealnym przypadkiem, który musimy przetestować.

Na podstawie podanych danych i dowodów jesteśmy w stanie przetestować dwie hipotezy i stwierdzić, czy musimy odrzucić hipotezę zerową lub nie musimy odrzucać hipotezy zerowej.

**Hipoteza jest testowana w czterech krokach:**

* Musimy wyjaśnić zarówno hipotezę zerową, jak i alternatywną, aby jedna z nich mogła zostać odrzucona.
* Ocena danych za pomocą planu analizy.
* Oblicz statystyki testowe i fizycznie przeanalizuj przykładowe dane.
* Na koniec interpretujemy wynik i odrzucamy jedną z dwóch hipotez.

### ANOVA

ANOVA to zasadniczo ***testowanie hipotez dla wielu grup***. Służy do sprawdzania, czy grupy mają takie same środki i wariancje. Podczas gdy test t może być żmudny i złożony ze względu na różne czynniki, ANOVA może pełnić tę rolę znacznie lepiej przy minimalnym poziomie błędu. ***Stosując współczynnik F, możemy zmierzyć ANOVA***. Współczynnik F jest zdefiniowany jako ***stosunek średniego kwadratu (między grupami) do średniego kwadratu (wewnętrznie w grupie).***

Różne etapy obliczania ANOVA to:

* Zrozumienie i wygenerowanie dwóch hipotez - zerowej i alternatywnej. W hipotezie zerowej załóż, że średnia wszystkich grup jest taka sama, podczas gdy w hipotezie alternatywnej średnia jest inna.
* Oblicz średnie kwadraty w grupach i średnie kwadraty między grupami.
* Obliczyć współczynnik F, korzystając z powyższych wartości.
* Korzystając z tabeli F, oblicz prawdopodobieństwo.
* Jeśli obliczona wartość F jest większa niż wartość krytyczna F, wówczas odrzucamy naszą hipotezę zerową.

# Analiza danych ilościowych

Dwie popularne techniki analizy danych ilościowych to korelacja i regresja.

**Korelacja** jest statystyczną ***zależnością między dwiema zmiennymi losowymi*** a danymi dwuwymiarowymi. Istnieją trzy rodzaje korelacji - ***korelacja dodatnia, korelacja ujemna i korelacja zerowa***. Korelacja dodatnia oznacza, że ​​istnieje zależność między dwiema zmiennymi, która jest wprostproporcjonalna. W korelacji ujemnej przyrost jednej postaci zmiennej powoduje zmniejszenie innej zmiennej. Podczas gdy w zerowej korelacji nie ma absolutnie żadnej zależności między dwiema zmiennymi.

Innym rodzajem analizy danych ilościowych jest regresja. **Regresja** jest techniką statystyczną służącą **do szacowania zależności między zmiennymi.**

Obraz zawierający diagram, linia, Wykres, zrzut ekranu

Opis wygenerowany automatycznieRegresja może być regresją prostą, regresją wielokrotną opartą na liczbie zmiennych niezależnych. Ponadto, jeśli zastosowana funkcja ma **charakter nieliniowy**, wówczas typ regresji nazywany jest regresją nieliniową.

# Matematyka

## Algebra liniowa

Algebra liniowa odnosi się do badania wektorów i funkcji liniowych. Algebra liniowa jest kluczowa dla wszystkich tematów matematyki, a w przypadku nauki o danych, stanowi rdzeń **uczenia maszynowego**. Obejmuje różne operacje macierzowe, wektorowe i tensorowe w celu wykonania algorytmów uczenia maszynowego. Różne tematy uczenia maszynowego, takie jak widzenie komputerowe i przetwarzanie języka naturalnego, w dużej mierze opierają się na algebrze liniowej.

Wraz z pojawieniem się zaawansowanych algorytmów głębokiego uczenia komputery mogą wykonywać złożone problemy algebry liniowej. Algebra liniowa jest również wykorzystywana w głównej analizie komponentów (**PCA**), która jest tematem **redukcji wymiarów** w uczeniu maszynowym. Co więcej, jest on stosowany w połączeniu ze statystyką poprzez jego implementację w statystyce wielowymiarowej, rozwiązania najmniejszych kwadratów w regresji i macierz kowariancji w rozkładzie Gaussa. Dlatego, aby zostać badaczem danych, trzeba mieć solidne podstawy w algebrze liniowej.

## Rachunek różniczkowy

Rachunek różniczkowy to ***matematyczne badanie ciągłych zmian***. Potrzebujemy rachunku różniczkowego, aby zrozumieć, w jaki sposób ilości rozpraszają się (**pochodne)** i kumulują się (**całkują**) w czasie. W Data Science rachunek różniczkowy jest wykorzystywany w technikach optymalizacji w **Sztucznych Sieciach Neuronowych**. Jedną z takich technik optymalizacji, która wykorzystuje rachunek różniczkowy, jest opadanie gradientu. Algorytm **Spadku wzdłuż gradientu (*Gradient Descent)*** to pomiar zmiany wyniku funkcji ze zmianą wejścia. *Gradient Descent* wykorzystuje „funkcję kosztu”, która mierzy dobro dopasowania danej linii. Aby znaleźć gradient tej linii, używamy częściowego rachunku różniczkowego. Inną ważną formą rachunku różniczkowego stosowaną w Data Science jest rachunek wielowymiarowy.

## Matematyka dyskretna i logika

Jest to często mniej dyskutowany temat w schemacie „Matematyki dla nauki o danych”, ale faktem jest, że cała współczesna nauka danych odbywa się za pomocą systemów obliczeniowych, a matematyka dyskretna stanowi sedno systemów obliczeniowych. Odświeżenie matematyki dyskretnej zapewni badaczowi danych koncepcje kluczowe dla codziennego wykorzystania algorytmów i struktur danych w projekcie analitycznym. Kilka kluczowych tematów do nauczenia się tutaj,

* Zbiory, podzbiory, zbiory mocy
* Funkcje zliczania, kombinatoryka, policzalność
* Podstawowe techniki dowodowe - indukcja, dowód przez sprzeczność
* Podstawy logiki indukcyjnej, dedukcyjnej
* **Podstawowe struktury danych - stosy, kolejki, wykresy, tablice, tabele skrótów, drzewa**
* **Właściwości grafów** - połączone elementy, stopień, koncepcje maksymalnego przepływu / cięcia minimalnego, kolorystyka grafu
* Relacje i równania rekurencyjne
* Wzrost funkcji i **koncepcja notacji O (n)**

W każdej analizie sieci społecznościowej należy znać właściwości grafu i szybkiego algorytmu do wyszukiwania i przechodzenia przez sieć.

Obraz zawierający diagram, rysowanie, tekst, szkic

Opis wygenerowany automatyczniePrzy każdym wyborze algorytmu musisz zrozumieć złożoność czasu i przestrzeni, tj. Jak czas pracy i zapotrzebowanie na miejsce rosną wraz z wielkością danych wejściowych, używając **notacji O (n) (Big-Oh).**

***Notacja dużego O****– notacja przedstawiająca****asymptotyczne tempo wzrostu****, wykorzystywana do zapisywania złożoności obliczeniowej algorytmu. Za pomocą tej notacji zapisywany jest rząd wielkości funkcji wyrażającej liczbę operacji dominujących (w przypadku złożoności czasowej) lub rozmiar wymaganej pamięci (w przypadku złożoności pamięciowej) w zależności od liczby danych wejściowych.*

# Modele regresji w data science

**Regresje liniowe i logistyczne** są zwykle pierwszymi algorytmami, których ludzie uczą się w informatyce. Ze względu na ich popularność wielu analityków nawet myśli, że są jedyną formą regresji. Ci, którzy są nieco bardziej zaangażowani, uważają, że są najważniejsze spośród wszystkich form analizy regresji.

Prawda jest taka, że istnieją niezliczone formy regresji, które można wykonać. Każda forma ma swoje znaczenie i określony warunek, w którym najlepiej nadają się do zastosowania.

Celem jest wypracowanie idei zakresu regresji, zamiast stosowania regresji liniowej / logistycznej do każdego napotkanego problemu uczenia maszynowego, mając nadzieję, że będą pasować!

Rodzaje regresji:

1. Regresja liniowa
2. Regresja logistyczna
3. Regresja wielomianowa
4. Regresja krokowa
5. Regresja Ridge
6. Regresja Lasso
7. Regresja ElasticNet

Jak wybrać odpowiedni model regresji?

## Co to jest analiza regresji?

Obraz zawierający linia, Czcionka, diagram

Opis wygenerowany automatycznie**Analiza regresji jest formą techniki modelowania predykcyjnego**, która bada związek między **zmienną zależną (docelową**) a **zmienną (zmiennymi) niezależnymi (predyktorem**). Technikę tę stosuje się do prognozowania, modelowania szeregów czasowych i znajdowania związku przyczynowego między zmiennymi. **Na przykład** związek między jazdą ryzykowną a liczbą wypadków drogowych przez kierowcę najlepiej zbadać za pomocą regresji.

**Analiza regresji jest ważnym narzędziem do modelowania i analizy danych.**

W tym przypadku dopasowujemy krzywą / linię do punktów danych w taki sposób, aby zminimalizować różnice między odległościami punktów danych od krzywej lub linii.

## Ile mamy technik regresji?

Obraz zawierający tekst, Czcionka, zrzut ekranu, Jaskrawoniebieski

Opis wygenerowany automatycznieIstnieją różne rodzaje technik regresji do prognozowania. Techniki te są głównie oparte na trzech metrykach (**liczba zmiennych niezależnych**, **rodzaj zmiennych zależnych** i **kształt linii regresji**).

## Regresja liniowa

Obraz zawierający tekst, zrzut ekranu, linia, diagram

Opis wygenerowany automatycznieJest to jedna z najbardziej znanych technik modelowania. Regresja liniowa jest zwykle jednym z pierwszych kilku tematów, które ludzie wybierają podczas nauki modelowania predykcyjnego. W tej technice **zmienna zależna** jest **ciągła**, zmienne niezależne mogą być ciągłe lub dyskretne, a charakter linii regresji jest **liniowy**.

Regresja liniowa ustala związek między zmienną zależną **(Y)** a jedną lub większą liczbą zmiennych niezależnych **(X)** przy użyciu najlepiej dopasowanej linii prostej (znanej również jako linia regresji).

Jest to reprezentowane przez równanie **Y = a + b \* X + e**, gdzie **a** to przecięcie, **b** to nachylenie linii, a **e** to błąd. To równanie można wykorzystać do przewidywania wartości zmiennej docelowej na podstawie danych zmiennych predyktorów.

## Wielokrotna regresja liniowa

Różnica między prostą regresją liniową a **wielokrotną regresją liniową** polega na tym, że wielokrotna regresja liniowa ma (> 1) zmienne niezależne, podczas gdy prosta regresja liniowa ma tylko 1 zmienną niezależną.

**Jak uzyskać linię najlepszego dopasowania (wartość a i b)?**

Zadanie to można łatwo wykonać **metodą najmniejszych kwadratów**.

Jest to najczęstsza metoda stosowana do dopasowania linii regresji. Oblicza linię najlepiej dopasowaną do obserwowanych danych, minimalizując sumę kwadratów odchyleń pionowych od każdego punktu danych do linii. Ponieważ odchylenia są najpierw podniesione do kwadratu, po dodaniu nie ma możliwości anulowania wartości dodatnich i ujemnych.

### Regresja wielokrotna Ważne punkty:

* **Musi istnieć liniowy związek między zmiennymi niezależnymi i zależnymi**
* Obraz zawierający tekst, Czcionka, biały, typografia

  Opis wygenerowany automatycznieObraz zawierający linia, diagram, Wykres, stok

  Opis wygenerowany automatycznieRegresja wielokrotna cierpi z powodu **wielo-współliniowości, autokorelacji**, **heteroskedastyczności** (*przynajmniej jedna zmienna losowa z ciągu różni się od innych wariancją*).
* Regresja liniowa jest bardzo wrażliwa na **wartości odstające**. Może to strasznie wpłynąć na linię regresji i ostatecznie prognozowane wartości.
* **Wielo-współliniowość** może zwiększyć wariancję oszacowań współczynników i uczynić oszacowania bardzo wrażliwymi na drobne zmiany w modelu. W rezultacie oszacowania współczynników są niestabilne
* W przypadku wielu zmiennych niezależnych możemy przejść do **selekcji do przodu**, **eliminacji do tyłu** i stopniowego podejścia do wyboru najbardziej znaczących zmiennych niezależnych.

## Regresja logistyczna

Obraz zawierający tekst, linia, Wykres, zrzut ekranu

Opis wygenerowany automatycznie**Regresja logistyczna** jest podstawową techniką klasyfikacji. Należy do grupy **klasyfikatorów liniowych.** Chociaż jest to zasadniczo metoda klasyfikacji binarnej, może być również stosowana do problemów klasyfikacji wieloklasowych.

Regresja logistyczna służy **do znalezienia prawdopodobieństwa zdarzenia** = sukces i zdarzenie = niepowodzenie.

Powinniśmy stosować regresję logistyczną, gdy **zmienna zależna ma charakter binarny** (0/1, prawda / fałsz, tak / nie).

Tutaj wartość Y wynosi od 0 do 1 i można ją przedstawić za pomocą następującego równania:

Y = p / (1-p) = prawdopodobieństwo wystąpienia zdarzenia / prawdopodobieństwo braku zdarzenia

ln (Y) = ln (p / (1-p))

**logit** (p) = ln (p / (1-p)) = b0 + b1X1 + b2X2 + b3X3 .... + bkXk

Powyżej **p jest prawdopodobieństwem obecności cechy będącej przedmiotem zainteresowania**. Pytanie, które powinieneś tutaj zadać, brzmi „dlaczego użyliśmy logarytmu w równaniu?”. XDD

Ponieważ pracujemy tutaj z rozkładem dwumianowym (zmienna zależna), musimy wybrać funkcję łączącą, która najlepiej nadaje się do tego rozkładu. I jest to **funkcja logarytmiczna logit.**

W powyższym równaniu parametry wybiera się tak, **aby zmaksymalizować prawdopodobieństwo zaobserwowania wartości próbki**, a nie zminimalizować sumę błędów kwadratu (jak w regresji zwykłej).

Obraz zawierający tekst, diagram, Wykres, zrzut ekranu

Opis wygenerowany automatycznieBędziesz potrzebował zrozumienia **funkcji esicy (sigmoid**) i **funkcji logarytmu naturalnego**, aby zrozumieć, czym jest regresja logistyczna i jak działa.  
Ten obraz pokazuje funkcję esicy (sigmoid lub krzywą w kształcie litery S) pewnej **zmiennej x**:

**Funkcja sigmoidalna** ma wartości bardzo zbliżone do 0 lub 1 w większości swojej domeny. Fakt ten sprawia, że nadaje się do stosowania w metodach klasyfikacji.

Zwróć uwagę, że logarytm naturalny standardowo jest oznaczony **ln** zamiast **log** (dziesiętny).

**Regresja logistyczna** jest klasyfikatorem liniowym, więc użyjesz funkcji liniowej f(x) = b₀ + b₁x₁ + ⋯ + brxr, zwanej także **logit**. Zmienne b₀, b₁, ..., br są **estymatorami współczynników regresji**, które nazywane są również **przewidywanymi wagami** lub współczynnikami.

Funkcja **regresji logistycznej** p(x) jest funkcją **sigmoidalną** f(x): p(x) = 1 / (1 + exp(−f(x)). W związku z tym często jest bliski 0 lub 1. Funkcja p(x) jest często interpretowana jako przewidywane prawdopodobieństwo, że wynik dla danego x jest równy 1. Zatem 1 − p(x) jest prawdopodobieństwem, że wynik wynosi 0.

**Regresja logistyczna** określa najlepsze **przewidywane wagi** b₀, b₁, ..., br tak, aby funkcja p(x) była jak najbardziej zbliżona do wszystkich rzeczywistych odpowiedzi yi, i = 1, ..., n, gdzie **n** jest liczbą obserwacji. Proces obliczania najlepszych wag przy użyciu dostępnych obserwacji nazywany jest **treningiem modelu** lub **dopasowaniem**.

### Ważne punkty:

Regresja logistyczna nie wymaga liniowej zależności między zmiennymi zależnymi i niezależnymi. Może obsługiwać różne typy relacji, ponieważ stosuje **nieliniową transformację logistyczną** do przewidywanego ilorazu szans

Aby uniknąć **nadmiernego dopasowania i niedopasowania**, powinniśmy uwzględnić wszystkie znaczące zmienne. Dobrym podejściem do zapewnienia tej praktyki jest zastosowanie metody krok po kroku do oszacowania regresji logistycznej

**Wymaga dużych próbek**, ponieważ szacunki maksymalnego prawdopodobieństwa są mniej skuteczne przy małych rozmiarach próby niż zwykłe najmniejsze kwadratowe

Zmienne niezależne nie powinny być ze sobą skorelowane, tzn. Nie mogą występować **wielokrotne współliniowości**. Mamy jednak opcje włączenia efektów interakcji zmiennych kategorialnych do analizy i modelu.

Jeśli wartości zmiennej zależnej są porządkowe, wówczas nazywa się to **regresją logistyczną ordynacyjną**

Jeśli zmienna zależna jest wieloklasowa, wówczas jest znana jako **regresja logistyczna wielomianowa.**

## Regresja wielomianowa

Równanie regresji jest równaniem regresji wielomianowej, jeśli **potęga zmiennej niezależnej jest większa niż 1**.

Obraz zawierający linia, Wykres, diagram, zrzut ekranu

Opis wygenerowany automatyczniePoniższe równanie reprezentuje równanie wielomianowe:

y = a + b \* x ^ 2

*y* = *a* + *b*1*x* + *b*2 *x*2 ⋯ *bn xn*

gdzie n jest stopniem wielomianu, a b jest zbiorem współczynników.

W tej technice regresji, **linia najlepszego dopasowania nie jest linią prostą.** Jest to raczej krzywa, która pasuje do punktów danych.

### Ważne punkty:

Obraz zawierający linia, diagram, Wykres, Czcionka

Opis wygenerowany automatycznieChociaż może występować pokusa, aby dopasować wielomian wyższego stopnia, aby uzyskać niższy błąd, może to jednak prowadzić do **nadmiernego dopasowania (overfitting)**.

Zawsze kreśl relacje, aby zobaczyć dopasowanie i skoncentruj się na upewnieniu się, że krzywa pasuje do charakteru problemu. Oto przykład, w jaki sposób wykres może pomóc:

Zwróć szczególną uwagę na zakrzywienia na końcach i sprawdź, czy te kształty i trendy mają sens.

Wyższe wielomiany mogą przynieść dziwne wyniki przy ekstrapolacji.

## Wielokrotna regresja krokowa

Ta forma regresji jest używana, gdy mamy do czynienia z **wieloma niezależnymi zmiennymi**.

W tej technice selekcji zmiennych niezależnych dokonuje się za pomocą automatycznego procesu, który nie wymaga interwencji człowieka.

Osiąga się to, obserwując wartości statystyczne, takie jak **R-kwadrat, t-statystyki i metryka AIC** **w celu rozpoznania istotnych statystycznie zmiennych**.

Regresja krokowa zasadniczo pasuje do modelu regresji poprzez dodawanie / upuszczanie ko-wariacji pojedynczo w oparciu o określone kryterium.

Niektóre z najczęściej używanych metod regresji krokowej są wymienione poniżej:

* **Standardowa regresja krokowa** robi dwie rzeczy. Dodaje i usuwa predyktory zgodnie z potrzebami dla każdego kroku.
* **Wybór do przodu** rozpoczyna się od najbardziej znaczącego predyktora w modelu i dodaje zmienną dla każdego kroku.
* **Eliminacja wsteczna** rozpoczyna się od wszystkich predyktorów w modelu i usuwa najmniej znaczącą zmienną dla każdego kroku.

Celem tej techniki modelowania jest maksymalizacja mocy predykcyjnej przy minimalnej liczbie zmiennych predykcyjnych. Jest to jedna z metod radzenia sobie z **wyższymi wymiarami** zestawu danych.

## Regresja Ridge

Regresja Ridge'a jest techniką stosowaną, gdy **dane „cierpią” na wielo-współliniowość** (zmienne niezależne są wysoce skorelowane).

W wielo-współliniowości, chociaż oszacowania metodą najmniejszych kwadratów (OLS) są obiektywne, ich wariancje są duże, co odbiega od wartości obserwowanej daleko od wartości prawdziwej.

Dodając pewien stopień odchylenia do oszacowań regresji, regresja Ridge zmniejsza standardowe błędy.

Powyżej widzieliśmy równanie regresji liniowej. Może być prezentowana jako:

***y = a + bx***

To równanie zawiera również błąd. Pełne równanie jest więc następujące:

**y = a + bx + e** (błąd),

[*termin błędu to wartość potrzebna do skorygowania błędu prognozy między wartością obserwowaną a przewidywaną*]

**y = a + b1x1 + b2x2 + .... + e**, dla wielu zmiennych niezależnych.

W równaniu liniowym błędy prognozowania można rozłożyć na dwa podskładniki. Pierwszy wynika z **tendencyjności**, a drugi z **wariancji**. Błąd prognozy może wystąpić z powodu jednego z tych dwóch lub obu składników. Tutaj omówimy błąd spowodowany przez **wariancję**.

Obraz zawierający tekst, Czcionka, biały, design

Opis wygenerowany automatycznieRegresja Ridge rozwiązuje problem wielo-współliniowości poprzez parametr kurczenia λ (lambda). Spójrz na równanie.

W tym równaniu mamy dwa składniki. Pierwszy z nich jest **najmniejszym kwadratem**, a drugi to **lambda sumy β2 (beta-kwadrat)**, gdzie **β** jest współczynnikiem. Jest to dodawane do metody najmniejszych kwadratów, aby zmniejszyć wartość parametru, tak aby miał bardzo niską wariancję.

### Ważne punkty:

* Założenia tej regresji są takie same jak regresja najmniejszych kwadratów, z tym wyjątkiem, że nie należy zakładać normalności
* Zmniejsza wartość współczynników, ale nie osiąga zera, co sugeruje brak funkcji wyboru cech (zmiennych)
* Jest to **metoda regularyzacji**.

## Regresja Lasso

Podobnie jak regresja grzbietu, Lasso (operator najmniejszego bezwzględnego skurczenia i operatora selekcji) również **„karze” bezwzględny rozmiar współczynników regresji**.

Ponadto jest w stanie zmniejszyć zmienność i poprawić dokładność modeli regresji liniowej.

Obraz zawierający tekst, Czcionka, biały, design

Opis wygenerowany automatycznieSpójrz na równanie: Regresja Lasso różni się od regresji ridge tym, że wykorzystuje wartości bezwzględne w funkcji „kary” zamiast kwadratów.

Prowadzi to do „karania” (inaczej równoważnego ograniczania sumy wartości bezwzględnych oszacowań) wartości, co powoduje, że niektóre oszacowania parametrów osiągają dokładnie zero. Im większa zastosowana „kara”, tym bardziej szacunki kurczą się w kierunku zera absolutnego. To powoduje wybór zmiennych spośród podanych n zmiennych.

### Ważne punkty:

* Założenia tej regresji są takie same jak regresja najmniejszych kwadratów, z tym wyjątkiem, że nie należy zakładać normalności
* Zmniejsza współczynniki do zera (dokładnie zero), co z pewnością pomaga w wyborze funkcji
* Jest to **metoda regularyzacji** i wykorzystuje normalizację
* Jeśli grupa predyktorów jest wysoce skorelowana, lasso wybiera tylko **jeden** z nich, a zmniejsza pozostałe do zera

## Inne techniki regresji

**Regresja ElasticNet** jest hybrydą technik regresji Lasso i Ridge'a. ElasticNet jest przydatna, gdy istnieje wiele skorelowanych elementów. Lasso prawdopodobnie wybierze jeden z nich losowo, podczas gdy „elastyczna siatka” prawdopodobnie wybierze dwa.

**Bayesowska regresja liniowa** to podejście do regresji liniowej, w którym analiza statystyczna jest przeprowadzana w kontekście wnioskowania bayesowskiego. Kiedy w modelu regresji występują błędy o rozkładzie normalnym i przy założeniu określonej formy wcześniejszego rozkładu (a priori), dostępne są wyraźne wyniki dla późniejszych (a posteriori) rozkładów prawdopodobieństwa parametrów modelu.

## Jak wybrać odpowiedni model regresji?

W ramach wielu typów modeli regresji ważne jest, aby wybrać najlepiej dopasowaną technikę w oparciu o **rodzaj zmiennych niezależnych** i **zależnych**, **wymiarowość danych** i **inne** istotne cechy danych. Poniżej znajdują się kluczowe czynniki, które powinieneś przećwiczyć, aby wybrać odpowiedni model regresji:

**Eksploracja danych** jest nieuniknioną częścią budowania modelu predykcyjnego. To powinien być Twój pierwszy krok przed wyborem odpowiedniego modelu, na przykład określenie zależności i wpływu zmiennych

Aby porównać poprawność dopasowania dla różnych modeli, możemy analizować różne metryki, takie jak **istotność statystyczna parametrów**, **R-kwadrat**, **Skorygowany** **r-kwadrat**, **błąd**.

**Walidacja krzyżowa** jest najlepszym sposobem oceny modeli używanych do prognozowania. Tutaj dzielisz swój zestaw danych na dwie grupy (trenuj i weryfikuj). Prosta średnia kwadratowa różnica między zaobserwowanymi a przewidywanymi wartościami daje miarę dokładności prognozowania.

Jeśli twój zestaw danych zawiera wiele mylących zmiennych, nie powinieneś wybierać metody automatycznego wyboru modelu, ponieważ nie chcesz jednocześnie umieszczać ich w modelu.

Będzie to również zależeć od twojego celu. Może się zdarzyć, że łatwiej jest wdrożyć mniej wydajny model w porównaniu z modelem o znaczącej statystyce.

**Metody regularyzacji** regresji (Lasso, Ridge i ElasticNet) działają dobrze w przypadku wysokiej wymiarowości i wielokoliniowości między zmiennymi w zbiorze danych.

# Co to jest redukcja wymiarów?

W uczeniu maszynowym i statystyce **redukcja wymiarów** to proces zmniejszania liczby rozważanych zmiennych losowych poprzez uzyskanie zestawu zmiennych głównych.

**Można go podzielić na eliminację jednych zmiennych i wybór drugich i ekstrakcję zmiennych.**

Zajmiemy się dwoma głównymi algorytmami w redukcji wymiarów

1. Analiza głównych składowych (PCA)
2. Liniowa analiza dyskryminacyjna (LDA)

Obraz zawierający tekst, zrzut ekranu, diagram

Opis wygenerowany automatycznie**Problem:** jak przedstawić obraz dwuwymiarowy (mapę) powiązań korelacyjnych między wieloma zmiennymi ?

Y = a + bx 2 wymiary

Y = a + b1x + b2x 3 wymiary

Y = a + b1x + b2x + b3x 4 wymiary

Y = a + b1x + b2x + b3x +…+bnx n wymiarów

## Analiza głównych składowych (PCA)

### Obraz zawierający linia, zrzut ekranu, diagram, Wykres Opis wygenerowany automatycznieCo to jest zasada analizy składowych (PCA)?

Obraz zawierający linia, diagram, Grafika, krąg

Opis wygenerowany automatycznie Obraz zawierający linia, diagram

Opis wygenerowany automatycznie

Jeśli wcześniej pracowałeś z wieloma zmiennymi, wiesz, że może to stanowić problem. Czy rozumiesz związek między każdą zmienną? Czy masz tak wiele zmiennych, że istnieje ryzyko „nadpisania” modelu danymi?

**Rozwiązanie:** z technicznego punktu widzenia **możesz „zmniejszyć” ilość wymiarów przestrzeni obiektów**. Zmniejszając wymiar przestrzeni obiektów, masz mniej zależności między zmiennymi do rozważenia i mniej prawdopodobne jest „przegrzanie” modelu.

Zmniejszenie wymiaru przestrzeni cech nazywa się „redukcją wielowymiarowości”. Istnieje wiele sposobów osiągnięcia redukcji wielowymiarowości, ale większość technik należy do 1 z dwóch klas.

* **Eliminacja cech (zmiennych)**
* **Ekstrakcja cech (zmiennych)**

**Eliminacja cech**: zmniejszamy przestrzeń cech poprzez ich eliminację. Zaletami metody eliminacji cech są prostota i łatwość utrzymania.

**Ekstrakcja cech**: PCA to technika ekstrakcji cech. Więc łączy nasze zmienne wejściowe w określony sposób, a następnie możemy upuścić „najmniej ważne” zmienne, zachowując jednocześnie najcenniejsze części wszystkich zmiennych.

### Kiedy powinno się używać PCA?

1. Gdy chcesz zmniejszyć liczbę zmiennych, ale nie możesz zidentyfikować zmiennych, które można całkowicie usunąć z rozważań.
2. Gdy chcesz mieć pewność, że twoje zmienne są od siebie niezależne?
3. Gdy zmienna niezależna jest mniej interpretowalną?

### Jak działa zasada analizy składników (PCA)?

Mamy zamiar obliczyć macierz, która podsumowuje, w jaki sposób wszystkie nasze zmienne odnoszą się do siebie.

Następnie podzielimy tę macierz na dwa osobne elementy: **kierunek** i **wielkość**. Możemy wtedy zrozumieć kierunek naszych danych i ich wielkość.

Obraz zawierający diagram, linia, Wykres

Opis wygenerowany automatycznie🡪 wykres pokazuje dwa główne kierunki tych danych: kierunek zielony i kierunek czerwony.

Najważniejszy jest kierunek czerwony. Biorąc pod uwagę sposób ułożenia kropek, możesz zobaczyć, dlaczego czerwony kierunek wydaje się ważniejszy niż zielony kierunek. Czerwony kierunek najlepiej pasowałoby do linii najlepszego dopasowania danych.

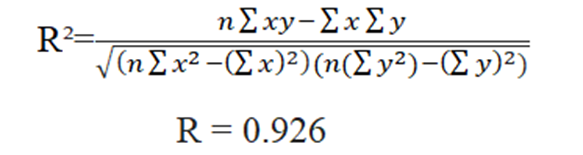
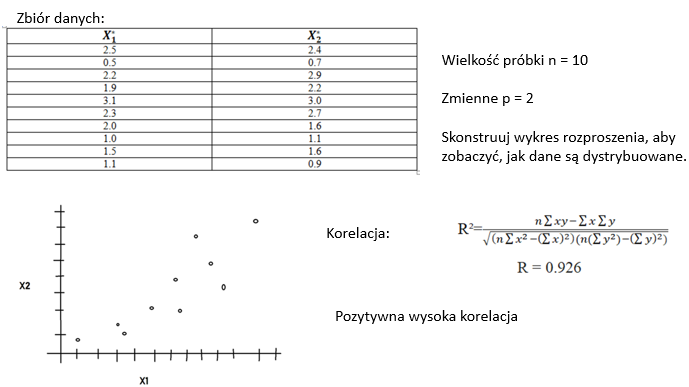
Obraz zawierający tekst, linia, Wykres, diagram

Opis wygenerowany automatycznie🡪 wykresprzekształca nasze oryginalne dane wyrównując je kierunkami czerwonym i zielonym. Wykres zawiera te same dane wyodrębnienia jak powyżej, ale został przekształcony tak, aby oś X i Y były teraz kierunkiem.

Co zrobić, aby linia najlepszego dopasowania wyglądałaby jak wyżej :

1. Oblicz macierz kowariancji X punktów danych.
2. Oblicz **wektory własne** (eigenvectors) i odpowiadające im **wartości własne** (eigevalues) .
3. Sortuj wektory własne według ich wartości w porządku malejącym.
4. Wybierz pierwsze k wektorów własnych, a będą to nowe k wymiary.
5. Przekształć oryginalne n-wymiarowe punkty danych w k wymiarowe.

### Zanurkujmy do matematyki:



Średnia z naszych zmiennych

Obraz zawierający paragon, linia, tekst, diagram

Opis wygenerowany automatycznie

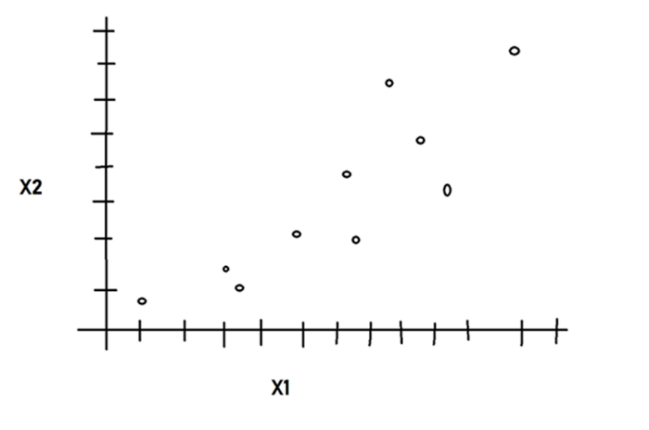
#### Krok 1: utwórz macierz X

Obraz zawierający tekst, numer, krzyżówka

Opis wygenerowany automatycznie· Odejmij średnią z odpowiedniego komponentu danych, aby zaktualizować zestaw danych.

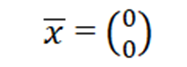
· Zrekonstruuj wykres rozproszenia.

· Zapisz „skorygowane” dane jako macierz X.  
Uwaga: ten „skorygowany” zestaw danych będzie miał średnią 0.



Teraz napisz „skorygowane” dane jako Macierz X

Uwaga: skorygowany „zbiór danych będzie miał średnią zero:



· Zrekonstruuj wykres rozproszenia.

#### Obraz zawierający tekst, Czcionka, zrzut ekranu, pismo odręczne Opis wygenerowany automatycznieKrok 2: macierz wariancji-kowariancji C

Oblicz przykładową **macierz wariancji-kowariancji C**.

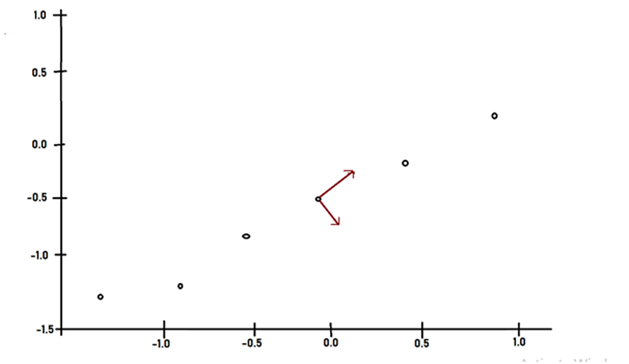
#### Krok 3: wartości własne (eigenvalues)

Obliczyć wartości własne (eigenvalues) lambda

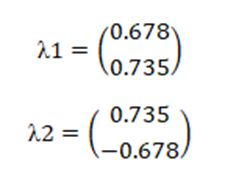
Obraz zawierający diagram, linia, Wykres

Opis wygenerowany automatycznieObraz zawierający tekst, Czcionka, biały, typografia

Opis wygenerowany automatycznieZ macierzy C uporządkuj odpowiednie pary od najwyższych do najniższych wartości własnych. Powstanie macierz A



**Teraz wektory własne (eigenvalue)**



W Eigen Vector1 idź w prawo, a 0,735 kierunków w górę

W Eigen Vector2 idź w prawo, a -0,678 kierunków w górę

Można to udowodnić:

Całkowita wariancja próbki = suma wartości własnej (eigenvalue)

= 1,28 +0,0490 = 1,33

Dzięki temu procesowi będziemy w stanie dokładnie określić linie charakteryzujące dane. Pierwszy wektor własny przejdzie przez środek punktów danych, jakby to były linie najlepszego dopasowania.

Drugi wektor własny da nam inny mniej ważny wzorzec danych. Oznacza to, że wszystkie punkty danych są zgodne z główną linią. Ale czy z jakiegoś powodu są one z boku głównej linii?

#### Obraz zawierający diagram, linia, Wykres Opis wygenerowany automatycznieKrok 4: wektory własne (eigenvectors)

Obraz zawierający tekst, Czcionka, biały, Grafika

Opis wygenerowany automatycznieWybierz komponenty i utwórz **macierz wektorów własnych V**. Po uporządkowaniu **wektorów własnych według wartości własnej** daje to składniki w kolejności ich **istotności. Zatem wektor własny o najwyższej wartości własnej jest główną składową.** Składniki o mniejszym znaczeniu można zignorować. Aby zmniejszyć wymiary zestawu danych.

Wybierz oba komponenty

Obraz zawierający Czcionka, biały, Grafika, logo

Opis wygenerowany automatycznieOdrzuć więc mniej znaczący komponent (bierzemy PC1, który stanowi 96% przechwytywania informacji)

#### Krok 5: Przekształcenie danych

Podziel nowy zestaw danych, biorąc wektor własny V

Y = XV

Przekształciliśmy nasze dane i wyrazimy to w formie wzorca między nimi, gdzie wzorce są liniami funkcji, które najściślej opisują związek między danymi.

Obraz zawierający Czcionka, tekst, biały, design

Opis wygenerowany automatycznieObraz zawierający tekst, diagram, linia

Opis wygenerowany automatycznieObraz zawierający tekst, zrzut ekranu, Czcionka, numer

Opis wygenerowany automatycznieObraz zawierający diagram, linia

Opis wygenerowany automatycznieObraz zawierający tekst, Czcionka, biały, algebra

Opis wygenerowany automatycznie

Teraz ponownie należy odrzuć mniej znaczący komponent

Teraz:

Obraz zawierający tekst, diagram, linia, numer

Opis wygenerowany automatycznieObraz zawierający tekst, linia, Czcionka, numer

Opis wygenerowany automatycznieObraz zawierający linia, diagram

Opis wygenerowany automatycznie

Obraz zawierający tekst, Czcionka, typografia, biały

Opis wygenerowany automatycznie**W tym przypadku PCA zmniejsza jeden wymiar**

**Tak pracuje redukcja wymiaru w PCA**

## Obraz zawierający krąg, diagram, zrzut ekranu, linia Opis wygenerowany automatycznieLiniowa analiza dyskryminacyjna (LDA).

### Co to jest liniowa analiza dyskryminacyjna (LDA)?

LDA jest rodzajem kombinacji liniowej, matematycznym procesem wykorzystującym różne elementy danych i stosującym funkcję, aby oddzielnie analizować wiele **klas obiektów lub elementów.**

Liniowa analiza dyskryminacyjna może być przydatna w takich obszarach, jak rozpoznawanie obrazów i analiza predykcyjna w marketingu.

**LDA pomaga w danych dla więcej niż dwóch klas, gdy regresja logistyczna nie jest wystarczająca.**

Liniowa analiza dyskryminacyjna przyjmuje wartość średnią dla każdej klasy i uwzględnia warianty w celu przewidywania przy założeniu rozkładu Gaussa.

Maksymalizacja osi składowych dla separacji klas.

### Jak działa liniowa analiza dyskryminacyjna (LDA)?

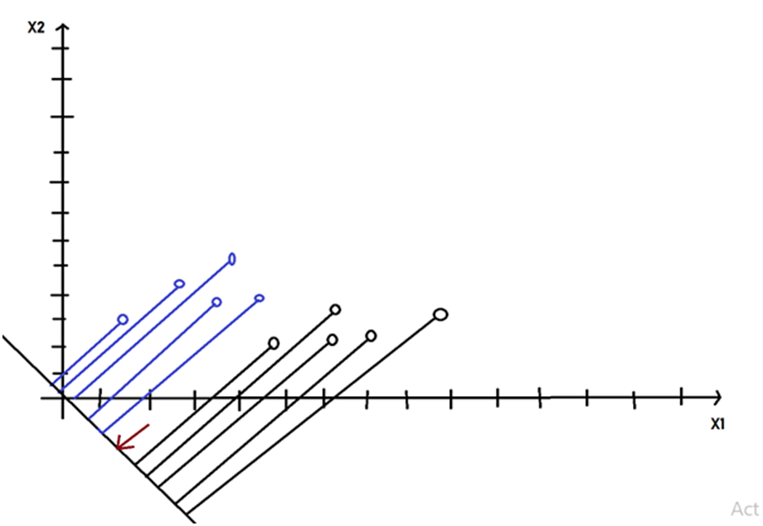
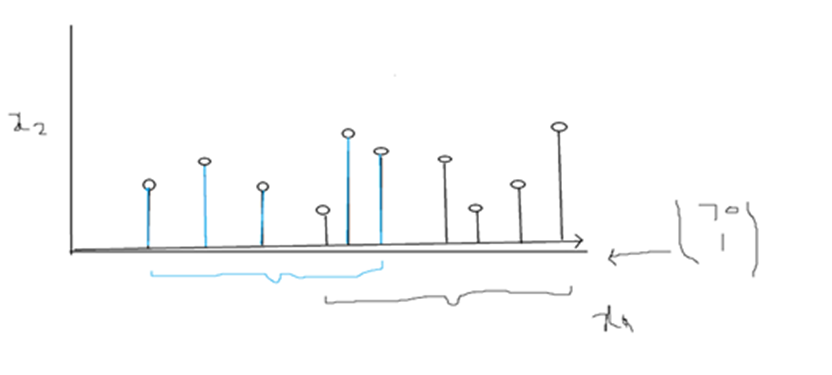
Pierwsze ogólne kroki do przeprowadzenia liniowej analizy dyskryminacyjnej:

1. Oblicz wektor średniej wymiarowej d dla różnych klas z zestawu danych.
2. Oblicz macierz rozproszenia (pomiędzy klasami i wewnątrz klasowej macierzy rozproszenia)
3. Posortuj wektor własny, zmniejszając wartość własną i wybierz wektor własny o największej wartości własnej z macierzy wymiarowej d \* k w (gdzie każda kolumna reprezentuje wektor własny)
4. Zastosowano macierz wektorów własnych d \* k do przekształcenia próbki w nową podprzestrzeń.

Można to podsumować mnożąc macierz:

Y = X \* W

(gdzie X jest macierzą wymiarów n \* d reprezentującą n próbek, a LDA przekształca n \* k próbek wymiarowych w nowej podprzestrzeni.

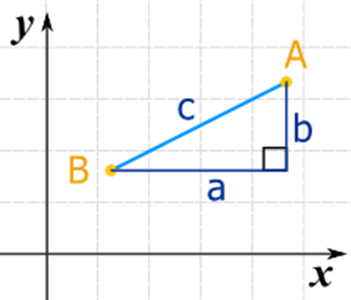
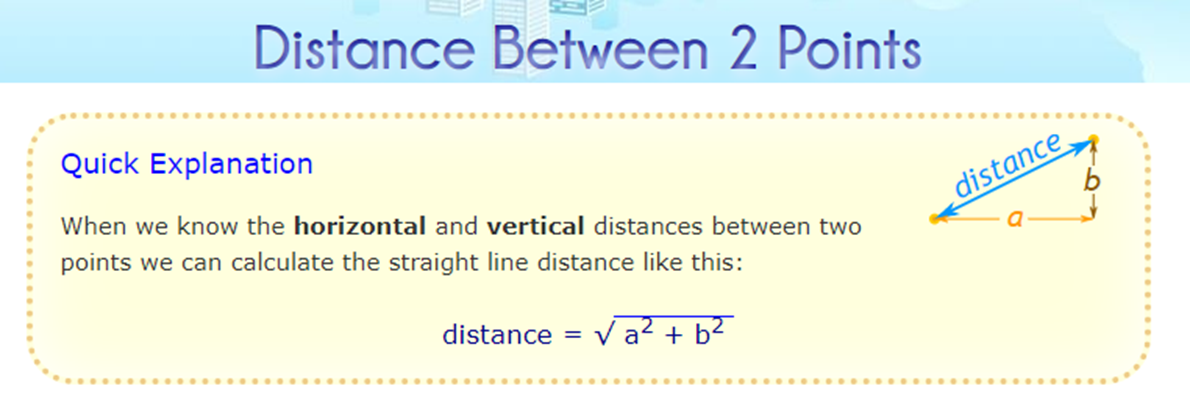


# Algorytm K- Najbliżsi Sąsiedzi (*KNN*)

Obraz zawierający krąg, diagram, design

Opis wygenerowany automatycznie**KNN** jest nieparametrycznym algorytmem uczenia nadzorowanego. Jego celem jest **wykorzystanie zestawu (próby) danych, w którym punkty danych są podzielone na kilka klas** *(kwadraty, trójkąty)***, aby przewidzieć klasyfikację nowego punktu** *(kółko).*

Algorytm KNN zakłada, że podobne rzeczy istnieją w bliskiej odległości. Innymi słowy, podobne rzeczy są blisko siebie.



## Jak algorytm KNN pracuje?

Obraz zawierający linia, zrzut ekranu, diagram

Opis wygenerowany automatycznieWeźmy prosty przypadek, aby zrozumieć ten algorytm. Poniżej przedstawiono rozkład czerwonych kół (RC) i zielonych kwadratów (GS):

Zamierzasz poznać klasę niebieskiej gwiazdy (BS). BS może być RC lub GS i nic więcej. „K” to algorytm KNN najbliżsi sąsiedzi, od których chcemy dostać głos.

Obraz zawierający diagram, zrzut ekranu, linia, krąg

Opis wygenerowany automatyczniePowiedzmy, że K = 3. Zrobimy teraz okrąg z BS o tak dużym promieniu, aby zawierał tylko trzy punkty danych na płaszczyźnie. Aby uzyskać więcej informacji, zapoznaj się z poniższym diagramem:

Trzy najbliższe punkty BS to wszystkie RC. Dlatego przy wysokim poziomie ufności możemy powiedzieć, że BS powinna należeć do klasy RC. Tutaj wybór stał się bardzo oczywisty, ponieważ wszystkie trzy głosy od najbliższego sąsiada przeszły do RC.

Wybór parametru K jest bardzo istotny w tym algorytmie. Następnie zrozumiemy, jakie czynniki należy wziąć pod uwagę, aby dojść do najlepszego K.

## Jak wybrać czynnik K?

Najpierw spróbujmy zrozumieć, co dokładnie wpływa na K w algorytmie. Jeśli zobaczymy ostatni przykład, biorąc pod uwagę, że wszystkie 6 obserwacji uczenia pozostają stałe, przy danej wartości K możemy wyznaczyć granice każdej klasy. Granice te oddzielą RC od GS.

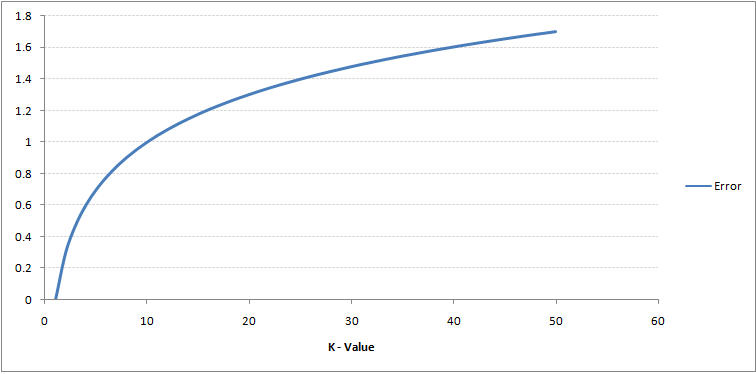
Obraz zawierający zrzut ekranu, Grafika

Opis wygenerowany automatycznieW ten sam sposób spróbujmy zobaczyć wpływ wartości „K” na granice klas. Poniżej przedstawiono różne granice oddzielające dwie klasy o różnych wartościach K.

Jeśli uważnie się przyjrzysz, zobaczysz, że granica staje się gładsza wraz ze wzrostem wartości K. Gdy K rośnie do nieskończoności, ostatecznie staje się cała niebieska lub cała czerwona, w zależności od całkowitej większości.

**Wskaźnik błędu uczenia i wskaźnik błędu walidacji** to dwa parametry, do których musimy uzyskać dostęp przy różnej wartości K. Poniżej przedstawiono krzywą wskaźnika błędu szkolenia ze zmienną wartością K:

## Wskaźnik błędu uczenia i wskaźnik błędu walidacji

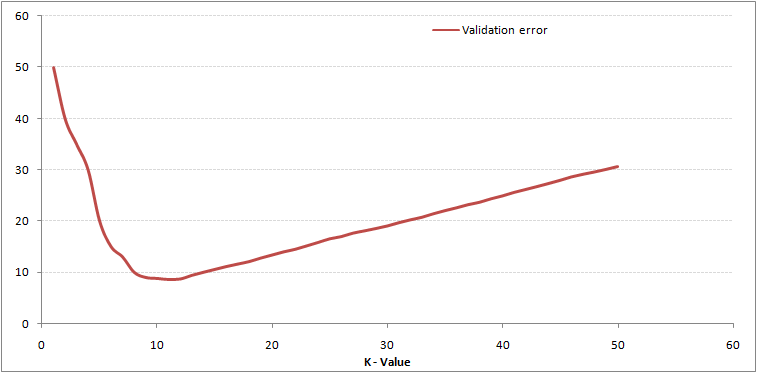


Próbka danych ucząca

Poziom błędu przy K = 1 wynosi zawsze zero dla próbki uczącej. Wynika to z faktu, że najbliższy punkt do dowolnego punktu danych uczących jest on sam. Dlatego przewidywanie jest zawsze dokładne z K = 1.

Jeśli krzywa błędu walidacji byłaby podobna, nasz wybór K byłby 1. Poniżej krzywa błędu walidacji o zmiennej wartości K:

Dzięki temu sprawa jest bardziej przejrzysta. Przy K = 1 przekraczaliśmy granicę dopasowania (50%). Odtąd poziom błędu początkowo zmniejsza się i osiąga minimum. Po punkcie minimalnym zwiększa się wraz ze wzrostem K.



Próbka danych walidacyjna

**Aby uzyskać optymalną wartość K, trzeba podzielić początkowy zestaw danych na dane uczące i walidacyjne. Teraz można wykreślić krzywą błędu walidacji,** aby uzyskać optymalną wartość K. Tej wartości K należy użyć do prognozy.

# Algorytm Maszyny Wektorów Nośnych

**(Support Vector Machines (SVM))**

Maszyna wektorów nośnych to kolejny, po algorytmach regresji liniowej i regresji logistycznej, prosty algorytm, który każdy ekspert uczenia maszynowego powinien mieć w swoim arsenale. Maszyna wektora nośnego jest bardzo preferowana przez wielu, ponieważ zapewnia znaczną dokładność przy mniejszej mocy obliczeniowej.

Obsługa maszyny wektorowej, w skrócie SVM, może być używana zarówno do zadań regresji, jak i klasyfikacji. Ale jest szeroko stosowany w celach klasyfikacji.

## Co to jest Maszyna Wektora Nośnego?

Obraz zawierający linia, diagram, Wykres

Opis wygenerowany automatycznie**Algorytm maszyny wektora nośnego ma na celu znalezienie hiperpłaszczyzny w przestrzeni N-wymiarowej (N - liczba elementów), która wyraźnie klasyfikuje punkty danych.**

Aby oddzielić dwie klasy punktów danych, można wybrać wiele możliwych hiperpłaszczyzn. Naszym celem jest znalezienie płaszczyzny, która ma maksymalny margines, tj. Maksymalną odległość między punktami danych obu klas.

Maksymalizacja odległości marginesu zapewnia pewne wzmocnienie, aby przyszłe punkty danych mogły być klasyfikowane z większą pewnością.

## Hiperpłaszczyzny i wektory nośne

Obraz zawierający diagram, tekst, linia

Opis wygenerowany automatycznie

Hiperpłaszczyzny w przestrzeni funkcji 2D i 3D

Hiperpłaszczyzny to granice decyzyjne, które pomagają klasyfikować punkty danych. Punkty danych przypadające po obu stronach hiperpłaszczyzny można przypisać do różnych klas.

Obraz zawierający linia, diagram, zrzut ekranu, Wykres

Opis wygenerowany automatyczniePonadto wymiar hiperpłaszczyzny zależy od liczby elementów. Jeśli liczba obiektów wejściowych wynosi 2, wówczas hiperpłaszczyzna jest tylko linią. Jeśli liczba elementów wejściowych wynosi 2, ale w wymiarze 3D, wówczas hiperpłaszczyzna staje się płaszczyzną dwuwymiarową. Trudno sobie wyobrazić, gdy liczba funkcji wynosi 3 i więcej.

**Wektory nośne** to punkty danych, które są bliżej hiperpłaszczyzny i wpływają na pozycję i orientację hiperpłaszczyzny. Korzystając z tych wektorów pomocniczych, maksymalizujemy margines klasyfikatora.

Obraz zawierający diagram, tekst, zrzut ekranu, linia

Opis wygenerowany automatycznieUsunięcie wektorów nośnych zmieni położenie hiperpłaszczyzny. Są to punkty, które pomagają nam budować naszą maszynę SVM.

Klasyfikacja z próbkami (punktami) naruszającymi granice marginesu klasyfikatora – typowa sytuacja w badaniach.

Aby ocenić precyzję modelu klasyfikacji SVM, używamy miar **dokładności / wydajności i miar błędów klasyfikacji w oparciu o obliczenia macierzy pomyłek.**

**Modele SVM można dostroić (zoptymalizować) za pomocą walidacji krzyżowej.**

## Funkcje jądra (kernel)

Jedną z mocnych stron algorytmu SVM jest to, że **może poznać granice decyzyjne między klasami, których nie można rozdzielić liniowo**. Dodatkowy wymiar nazywa się ***jądrem (kernel).***

Algorytm SVM dodaje dodatkowy wymiar, aby liniowo oddzielić dane. Klasy w oryginalnych danych nie można rozdzielić liniowo. **Algorytm SVM dodaje dodatkowy wymiar**, który w dwuwymiarowej przestrzeni cech można zilustrować jako „rozciąganie” danych do trzeciego wymiaru.

Obraz zawierający wzór, krąg, diagram, design

Opis wygenerowany automatycznieObraz zawierający tekst, diagram, zrzut ekranu, design

Opis wygenerowany automatycznieTen dodatkowy wymiar pozwala na liniowe rozdzielenie danych. Kiedy ta hiperpłaszczyzna jest rzutowana z powrotem na pierwotne dwa wymiary, pojawia się jako **zakrzywiona granica decyzji funkcji jądra (kernel) (promieniowa).**

Jest wiele funkcji jądra (kernel), które udostępnia algorytm SVM

# Twierdzenie Bayes’a – jeden z filarów prognozowania w Data Science

## Wnioskowanie Bayes’a dla Data Science

Obraz zawierający tekst, diagram, Czcionka, linia

Opis wygenerowany automatycznie**A posteriori** (łac. „z następstwa”) – w filozofii termin będący przeciwieństwem wyrażenia ***a priori***, oznaczający tyle co: "po fakcie" tudzież "w następstwie faktu". Odnosi się on do poznania powstałego na doświadczeniu – rozumowania o czymś drogą indukcji ("od szczegółu do ogółu"), wskutek percepcji jakiegoś faktu

**A priori**[łac., ‘z góry’, ‘uprzedzając fakty’], *filoz. wcześniejszy, przed faktem, przed poznaniem faktu.*

**Twierdzenie Bayesa** jest podstawową zasadą prawdopodobieństwa. **Jest to określenie warunkowego prawdopodobieństwa zdarzenia.** To prawdopodobieństwo warunkowe jest znane jako hipoteza. Hipoteza ta jest obliczana na podstawie wcześniejszych dowodów lub wiedzy. **To prawdopodobieństwo warunkowe jest prawdopodobieństwem wystąpienia zdarzenia, biorąc pod uwagę, że miało miejsce już inne zdarzenie**.

## Formuła twierdzenia Bayes’a

Obraz zawierający tekst, Czcionka, biały, typografia

Opis wygenerowany automatycznieWzór na twierdzenie Bayesa obejmuje prawdopodobieństwo a posteriori ***p* (*H* | *E*)** jako iloczyn warunkowego prawdopodobieństwa hipotezy ***p* (*E* | *H*),** pomnożone przez prawdopodobieństwo a priori hipotezy ***p* (*H***) i podzielone przez prawdopodobieństwo dowodów ***p* (*E*)**.

Przyjrzyjmy się teraz szczegółowo każdemu terminowi formuły Twierdzenia Bayesa -

***p* (*H* | *E*)** - Jest to określane jako prawdopodobieństwo a posteriori. Posteriori zasadniczo oznacza czerpanie teorii z podanych dowodów. Oznacza warunkowe prawdopodobieństwo ***H*** (hipoteza), biorąc pod uwagę dowody ***E***.

***p* (*E* | *H*)** - Ten element twierdzenia Bayesa oznacza prawdopodobieństwo warunkowe. Jest to warunkowe prawdopodobieństwo wystąpienia dowodów, biorąc pod uwagę hipotezę. Oblicza prawdopodobieństwo dowodów, przy założeniu, że przyjęta hipoteza jest prawdziwa.

***p* (*H***) - Jest to określane jako a priori prawdopodobieństwo. Oznacza to pierwotne prawdopodobieństwo, że hipoteza H jest prawdziwa przed implementacją twierdzenia Bayesa. Oznacza to, że prawdopodobieństwo to nie wymaga udziału danych ani dowodów.

***p* (*E*)** - Jest to prawdopodobieństwo wystąpienia dowodów niezależnie od hipotezy.

## Jak działa klasyfikator Naive Bayes?

Zobaczmy przykład działania Naive Bayes. Podany przykład warunków pogodowych i uprawiania sportu. Musisz obliczyć prawdopodobieństwo uprawiania sportu: czy gracze będą grać, czy nie, na podstawie warunków pogodowych.

### Pierwsze podejście (w przypadku jednej funkcji)

Naiwny klasyfikator Bayesa oblicza prawdopodobieństwo zdarzenia w następujących krokach:

**Krok 1:** Oblicz a priori (wcześniejsze) prawdopodobieństwo dla danej etykiety klasy

**Krok 2:** Znajdź możliwość określenia prawdopodobieństwa dla każdego atrybutu dla każdej klasy

**Krok 3:** Umieść te wartości we wzorze Bayesa i oblicz prawdopodobieństwo a posteriori (późniejsze).

**Krok 4:** Sprawdź, która klasa ma większe prawdopodobieństwo, zakładając, że dane wejściowe należą do klasy z wyższym prawdopodobieństwem.

Aby uprościć wcześniejsze i późniejsze obliczanie prawdopodobieństwa, można użyć dwóch **tabel częstotliwości i prawdopodobieństwa.** Obie te tabele pomogą ci obliczyć prawdopodobieństwo wcześniejsze a-priori i późniejsze a-posteriori.

Obraz zawierający tekst, zrzut ekranu, numer, Czcionka

Opis wygenerowany automatycznieTabela częstotliwości zawiera występowanie etykiet dla wszystkich funkcji. Istnieją dwie tabele możliwości. Tabela możliwości (*Likelihood*) 1 pokazuje wcześniejsze prawdopodobieństwa zdażeń, a tabela możliwości (*Likelihood*) 2 pokazuje prawdopodobieństwo późniejsze.

#### Prawdopodobieństwo grania:

P (Tak | Pochmurno) = P (Pochmurno | Tak) P (Tak) / P (Pochmurno) ..................... (1)

Oblicz wcześniejsze a-priori prawdopodobieństwa:

P (pochmurno) = 4/14 = 0,29

P (tak) = 9/14 = 0,64

Oblicz prawdopodobieństwo a-posteriori:

P (Zachmurzenie | Tak) = 4/9 = 0,44

Umieść prawdopodobieństwa a-priori i a-posteriori w równaniu (1)

P (tak | zachmurzenie) = 0,44 \* 0,64 / 0,29 = 0,98 (wyższy)

#### Prawdopodobieństwo braku gry:

P (Nie | Pochmurno) = P (Pochmurno | Nie) P (Nie) / P (Pochmurno) ..................... (2)

Oblicz wcześniejsze a-priori prawdopodobieństwa:

P (pochmurno) = 4/14 = 0,29

P (nie) = 5/14 = 0,36

Oblicz prawdopodobieństwo a-posteriori:

P (Zachmurzenie | Nie) = 0/9 = 0

Umieść prawdopodobieństwa a-priori i a-posteriori w równaniu (2)

P (No | pochmurno) = 0 \* 0,36 / 0,29 = 0

## Przykład zastosowania Twierdzenia Bayes’a

Załóżmy, że pogoda w dzień jest pochmurna. Teraz musisz wiedzieć, czy dzisiaj będzie padać, biorąc pod uwagę zachmurzenie dnia. Dlatego należy obliczyć prawdopodobieństwo opadów deszczu, biorąc pod uwagę oznaki zachmurzenia.

To znaczy, **P** **(Deszcz | Chmury) = P(H | E),** gdzie stwierdzenie, czy dzisiaj będzie padać, to Hipoteza **(H)**, a Zachmurzenie jest Dowodem **(E)**. To jest część prawdopodobieństwa a posteriori naszego równania.

Załóżmy teraz, że wiemy, że w 60% przypadków opady deszczu jest spowodowane pochmurną pogodą. Dlatego istnieje prawdopodobieństwo, że będzie pochmurnie, biorąc pod uwagę deszcz, czyli **P (Chmury | Deszcz) = P (E | H).** Jest to prawdopodobieństwo wstecz, gdzie **E** jest dowodem na obserwację chmur, biorąc pod uwagę prawdopodobieństwo opadów, co pierwotnie jest naszą hipotezą. Teraz spośród wszystkich dni 75% dni w miesiącu jest pochmurno. Jest to prawdopodobieństwo zachmurzenia lub **P (Chmury).** Ponadto, ponieważ jest to deszczowy miesiąc w roku, pada zwykle przez 15 dni z 30 dni. Oznacza to, że prawdopodobieństwo hipotezy o opadach deszczu lub **P (H)** wynosi **P (Deszcz) = 15/30 = 0,5 lub 50%.** Teraz obliczmy prawdopodobieństwo deszczu, biorąc pod uwagę pochmurną pogodę.

**P (deszcz | chmura) = (P (chmura | deszcz) \* P (deszcz)) / (P (chmura)) = (0,6 \* 0,5) / (0,75) = 0,4**

Dlatego dowiadujemy się, że ze względu na pochmurną pogodę istnieje 40% szans na opady deszczu.

### Twierdzenie Naiwnego Bayesa

Naive Bayes to potężny algorytm uczenia nadzorowanego, który jest wykorzystywany do klasyfikacji. Klasyfikator Naive Bayesa jest rozszerzeniem wyżej omówionego standardowego twierdzenia Bayesa. W Naive Bayes obliczamy prawdopodobieństwo wniesione przez każdy czynnik. Najczęściej używamy go do operacji klasyfikacji tekstowej, takich jak filtrowanie spamu. Przykład pokazuje, w jaki sposób Naive Bayes oblicza prawdopodobieństwo udziału wszystkich czynników.

#### Przykłady zastosowania twierdzenia Naive Bayesa

***Przykład (1)***

Załóżmy, że jako specjalista od danych masz za zadanie opracować filtr antyspamowy. Otrzymujesz listę słów kluczowych spamu, takich jak

* *Darmowy*
* *Zniżka*
* *Pełen zwrot kosztów*
* *Pilne*
* *Utrata masy ciała*

Jednak firma, z którą współpracujesz, jest firmą finansującą produkty. Dlatego część słownictwa występującego w wiadomościach spamowych jest wykorzystywana w wiadomościach e-mail Twojej firmy. Niektóre z tych słów to:

* *Ważny*
* *Darmowy*
* *Pilne*
* *Zapasy*
* *Klienci*

Istnieje również prawdopodobieństwo użycia słów w wiadomościach spamowych i firmowych.

**E-mail ze spamem**

* *Bezpłatnie* (0,3)
* *Zniżka* (0,15)
* *Pełny zwrot* (0,1)
* *Pilne* (0,2)
* *Utrata masy ciała* (0,25)

**E-mail firmy**

* *Ważne* (0,5)
* *Bezpłatnie* (0,25)
* *Pilny* (0,1)
* *Zapasy* (0,5)
* *Klienci* (0,1)

Załóżmy, że otrzymałeś komunikat:

„*Bezpłatne próby dla programu odchudzania. Zostań członkami ze zniżką*. ”

Czy ta wiadomość to spam lub firmowy adres e-mail? Obliczanie prawdopodobieństwa wystąpienia składników w zdaniu - za darmo (0,4) + utrata masy ciała (0,25) + zniżka (0,15) = 0,8 lub 80%

Natomiast obliczenie prawdopodobieństwa, że jest to wiadomość e-mail od Twojej firmy = Bezpłatnie (0,25) = 0,25 lub 25%.

Dlatego prawdopodobieństwo, że poczta jest spamem jest znacznie wyższe niż firmowa wiadomość e-mail.

Obraz zawierający tekst, zrzut ekranu, numer, Czcionka

Opis wygenerowany automatycznie***Przykład (2)***

Używanie Naive Bayes do przewidywania stanu „play” przy użyciu zmiennej „pogoda”.

Na przykładzie 4, jaki jest wynik, jeśli pogoda = „słonecznie”?

Aby ustalić wynik, graj = „tak” lub „nie”, biorąc pod uwagę wartość zmiennej pogody = „słonecznie”,

  oblicz P (tak | słoneczny) i P (nie | słoneczny) i wybierz wynik z większym prawdopodobieństwem.

-> P (tak | słoneczny) = (P (słoneczny | tak) \* P (tak)) / P (słoneczny) = (3/9 \* 9/14) / (5/14) = 0,60

-> P (nie | słoneczny) = (P (słoneczny | nie) \* P (nie)) / P (słoneczny) = (2/5 \* 5/14) / (5/14) = 0,40

Tak więc, jeśli pogoda = „słonecznie”, wynikiem jest play = „tak”.

## Twierdzenie Naiwnego Bayesa c.d.

Naiwny klasyfikator Bayesa wybiera wynik o najwyższym prawdopodobieństwie, który w powyższym przypadku był cechą spamu.

Naiwny Bayes jest określany jako „naiwny”, ponieważ zakłada, że funkcje są od siebie niezależne. Funkcje w naszym przykładzie to słowa wejściowe obecne w zdaniu.

Obraz zawierający tekst, Czcionka, pismo odręczne, biały

Opis wygenerowany automatycznieTwierdzenie Bayesa:

Warunkowa niezależność wszystkich funkcji daje nam powyższą formułę.

Częstotliwość występowania cech od ***x*1** do ***x*d** oblicza się na podstawie ich stosunku do klasy ***c*j**. Wraz z wcześniejszym prawdopodobieństwem i prawdopodobieństwem wystąpienia zdarzenia obliczamy prawdopodobieństwo późniejsze, dzięki któremu jesteśmy w stanie znaleźć prawdopodobieństwo obiektu należącego do określonej klasy.

Wraz z tym twierdzeniem Bayesa naukowcy danych używają różnych narzędzi.

## Zastosowania twierdzenia Bayesa

### 1. Filtrowanie spamu

Pierwszym i najważniejszym zastosowaniem Naive Bayes jest jego zdolność do klasyfikowania tekstów, aw szczególności wiadomości e-mail zawierających spam od wiadomości niebędących spamem. Jest to jedna z najstarszych metod filtrowania spamu, z filtrowaniem spamu Naive Bayes datowanym na 1998 rok. Naive Bayes bierze dwie klasy - Spam i Ham i odpowiednio klasyfikuje dane.

### 2. Analiza sentymentów

Jest to część przetwarzania języka naturalnego, która analizuje, czy dane są pozytywne, negatywne czy neutralne. Inną terminologią do analizy sentymentów jest eksploracja opinii. Korzystając z Naive Bayes, możemy sklasyfikować, czy tekst jest pozytywny czy negatywny, lub określić, do której klasy należy uczucia danej osoby.

### 3. Systemy rekomendacji

Korzystając z Naive Bayes, możemy budować systemy rekomendacji. System rekomendacji mierzy prawdopodobieństwo, że osoba obejrzy film lub nie, biorąc pod uwagę wcześniejsze oglądanie. Jest również używany w połączeniu z filtrowaniem grupowym w celu filtrowania informacji dla użytkowników.

### 4. Bayesowskie sieci neuronowe

Niedawno Twierdzenie Bayesa zostało rozszerzone na **Deep Learning**, gdzie służy do projektowania potężnych sieci Bayesian. Jest on następnie wykorzystywany w złożonych zadaniach uczenia maszynowego, takich jak prognozowanie zapasów, rozpoznawanie twarzy itp. Jest to obecnie modny temat i zrewolucjonizował dziedzinę głębokiego uczenia się.

# Miary błędów predykcji (regresja)

**Absolutny błąd średni**

**(*Mean Absolute Error, MAE*)**



Gdzie:

n –granica przewidywań, mean(•) – średnia.

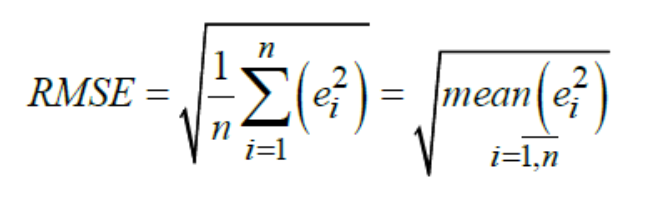
**Średni błąd kwadratowy**

**(Mean Square Error, MSE)**



**Średni błąd kwadratowy pierwiastkowy**

**(Root Mean Square Error, RMSE)**



## Błędy absolutne

**Absolutny Błąd Predykcji**

***(Absolute Prediction Error, APE)***

Obraz zawierający Czcionka, pismo odręczne, biały, kaligrafia

Opis wygenerowany automatycznie

Gdzie :

*yt*  - wartość obserwowana w czasie *t*,

ƒ*t* (m)– wartość predykcji w czasie *t*, otrzymana po   
użyciu modelu predykcyjnego *m*.

W kolejnych formułach (*m*) będzie pomimięta

**RMSE** jest najpopularniejszym wskaźnikiem oceny stosowanym **w problemach z regresją**. Wynika to z założenia, że ​​błąd ma normalny rozkład. Oto kluczowe kwestie do rozważenia w RMSE:

* Moc „pierwiastka kwadratowego” pozwala tej metodzie pokazać duże odchylenia wartości liczbowych.
* Charakter „kwadratowy” tej miary pomaga uzyskać bardziej wiarygodne wyniki, co zapobiega anulowaniu dodatnich i ujemnych wartości błędów. Innymi słowy, ta miara trafnie pokazuje prawdopodobną wielkość błędu.
* Unika stosowania bezwzględnych wartości błędów, co jest wysoce niepożądane w obliczeniach matematycznych.
* Gdy mamy więcej danych, rekonstruowanie rozkładu błędów za pomocą RMSE jest uważane za bardziej niezawodne.
* Obraz zawierający tekst, Czcionka, linia, zrzut ekranu

  Opis wygenerowany automatycznieNa wartości RMSE duży wpływ mają wartości odstające. Dlatego przed użyciem tej metryki upewnij się, że usunąłeś wartości odstające z zestawu danych.

Formuła RMSE jest podawana przez:

Ogólnie rzecz biorąc, **absolutne miary błędów mają następujące wady**.

* Główną wadą jest zależność od skali. Dlatego jeśli zadanie prognozy obejmuje obiekty o różnych skalach lub wielkościach, wówczas nie można zastosować miar błędu bezwzględnego do porównań.
* Kolejną wadą jest duży wpływ wartości odstających w danych na ocenę wyników prognozy. Tak więc, jeśli dane zawierają wartości odstające o maksymalnej wartości (jest to częsty przypadek w rzeczywistych zadaniach), wówczas bezwzględne miary błędów podają wartości zaniżone.
* RMSE, MSE mają niską niezawodność: wyniki mogą być różne w zależności od różnych części danych.

**Miary oparte na błędach procentowych**

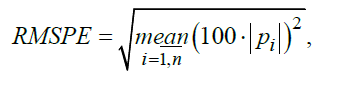
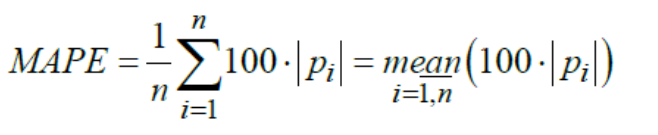
Błędy procentowe są obliczane na podstawie wartości *Pt*

**Średni absolutny błąd procentowy**

**(*Mean Absolute Percentage Error, MAPE*)**

**Średni kwadratowy pierwiastkowy błąd procentowy**

**(*Root Mean Square Percentage Error, RMSPE*)**



Zaleta:

Błędy danych wyrażonych w różnych skalach   
i jednostkach są porównywalne.

Niedociągnięcia:

Problem z dzieleniem przez zero, gdy wartość   
rzeczywista jest równa zero.

Problem niesymetryczny - wartości błędów różnią   
się gdy przewidywana wartość jest większa lub   
mniejsza niż aktualna.

Wartości odstające mają znaczący wpływ na wynik.

## Inne miary błędów

**Logarytmiczny błąd średni kwadratowy pierwiastkowy**

Obraz zawierający tekst, Czcionka, zrzut ekranu, linia

Opis wygenerowany automatycznieW przypadku **błędu logarytmicznego średniego kwadratowego pierwiastkowego**, bierzemy log prognoz i wartości rzeczywistych. RMSLE jest zwykle używany, gdy nie chcemy eliminować ogromnych różnic w wartościach przewidywanych i rzeczywistych, gdy zarówno wartości prognozowane, jak i prawdziwe są ogromnymi liczbami.

* Jeśli zarówno przewidywane, jak i rzeczywiste wartości są małe: RMSE i RMSLE są takie same.
* Jeśli wartość przewidywana lub rzeczywista jest duża: RMSE> RMSLE
* Jeśli zarówno wartości prognozowane, jak i rzeczywiste są duże: RMSE> RMSLE (RMSLE staje się prawie nieistotny).

### Błąd R-kwadrat

Obraz zawierający tekst, zrzut ekranu, Czcionka, biały

Opis wygenerowany automatycznieDowiedzieliśmy się, że gdy RMSE spada, wydajność modelu poprawi się. Ale same te wartości nie są intuicyjne. Nie mamy żadnego odniesienia (reperu) do porównania.

Tutaj możemy użyć metryki R-Squared. Wzór na R-Squared jest następujący:

MSE (model): średni błąd kwadratowy prognoz w stosunku do rzeczywistych wartości.

MSE (wyjściowy*, baseline):* średni błąd kwadratowy średniej prognozy w stosunku do rzeczywistych wartości

Innymi słowy, jak dobry jest nasz model regresji w porównaniu z bardzo prostym modelem, który po prostu przewiduje średnią wartość zestawu uczącego jako prognozy.

# Walidacja i walidacja krzyżowa do poprawy wydajności modeli w data science

**Podejście klasyczne z próbą testową (Metoda Hold-out)**

Jest to klasyczny „**najprostszy rodzaj walidacji**”. Ta metoda jest często klasyfikowana jako rodzaj „prostej walidacji, a nie formy walidacji krzyżowej”.

W tej metodzie losowo dzielimy **nasze dane na dwa: zestaw uczący i testowy tj. Zestaw hold-out**. Następnie uczymy model w zbiorze danych uczących i oceniamy model w zbiorze danych Test.

Obraz zawierający tekst, zrzut ekranu, linia, Czcionka

Opis wygenerowany automatycznieTechniki oceny modelu zastosowane w zbiorze danych sprawdzania poprawności do obliczenia błędu zależą od rodzaju problemu, nad którym pracujemy. Takich jak MSE stosowane w przypadku problemów z regresją, podczas gdy różne miary ( np. dokładności) zapewniające wskaźnik błędnej klasyfikacji pomagają znaleźć błąd w problemach z klasyfikacją.

Zazwyczaj zestaw danych uczących jest większy niż zestaw danych testowych. Typowe współczynniki stosowane do podziału zestawu danych obejmują 60:40, 80:20 itd. **Ta metoda jest używana tylko wtedy, gdy mamy tylko jeden model do oceny i nie ma dostrajania hiper-parametrów.**

Obraz zawierający tekst, zrzut ekranu, linia, Czcionka

Opis wygenerowany automatycznieTa metoda nie jest skuteczna w porównywaniu do wielu modeli i dostrajania ich hiperparametrów, co prowadzi nas do kolejnej bardzo popularnej formy metody hold-out, która obejmuje **podział danych na nie dwa, ale trzy oddzielne zestawy.**

Tutaj dzielimy zestaw danych na zestaw uczący i walidacyjny, **dzieląc w ten sposób oryginalny zestaw danych na zbiór treningowy, walidacyjny i testowy.**

Typowe kroki do wykonania walidacji modelu tutaj obejmują uczenie modelu lub często wielu modeli na zestawie uczącym. Zestaw walidacyjny, który jest zestawem hold-out z zestawu treningowego, tj. **Część zestawu uczącego odłożona na bok, jest następnie wykorzystywana do optymalizacji hiperparametrów modeli i oceny modelu.**

Zatem **zestaw walidacyjny służy do dostrajania różnych hiperparametrów i wybierania algorytmu o najlepszej wydajności**. Aby jednak w pełni ustalić, czy wybrany algorytm jest poprawny, dostosujemy model do zestawu danych uczących. Dzieje się tak dlatego, że jak dostrajamy hiper-parametry w oparciu o zestaw walidacyjny, w końcu nieco nadmiernie dopasowujemy (over-fitting) nasz model w oparciu o ten zestaw walidacyjny.

Tak więc **dokładność, którą otrzymujemy z zestawu sprawdzania poprawności, nie jest uważana za ostateczną, a inny zestaw danych, który jest testowym zestawem danych, służy do oceny końcowego wybranego modelu, a znaleziony tutaj błąd jest uważany za błąd uogólnienia**.

## Problem

Obraz zawierający tekst, zrzut ekranu, linia, Czcionka

Opis wygenerowany automatycznie**Jaki jest problem z tym podejściem?**

Ze względu na zmienność próby między zestawem uczącym a testowym nasz model zapewnia lepsze przewidywanie danych uczących, ale nie generalizuje danych testowych. Prowadzi to do niskiego poziomu błędu uczenia, ale wysokiego poziomu błędu testu.

Obraz zawierający tekst, zrzut ekranu, linia, Czcionka

Opis wygenerowany automatycznieKiedy dzielimy zestaw danych na zestaw uczący, walidacyjny i testowy, używamy tylko podzbioru danych i wiemy, że kiedy uczymy model na mniejszej liczbie obserwacji, model nie będzie działał dobrze i przeszacowujemy poziom błędu testu, tak aby model pasował do całego zestawu danych.

Aby rozwiązać ten problem, stosujemy podejście zwane **Walidacją krzyżową (*cross-validation)***

## Co to jest walidacja krzyżowa?

Obraz zawierający tekst, zrzut ekranu, Czcionka, Prostokąt

Opis wygenerowany automatycznie**Walidacja krzyżowa jest techniką statystyczną, która obejmuje podział danych na podzbiory, uczenie danych w podzbiorze i wykorzystanie tego drugiego podzestawu do oceny wydajności modelu.**

Aby zmniejszyć zmienność, przeprowadzamy wiele rund weryfikacji krzyżowej z różnymi podzbiorami z tych samych danych. Łączymy wyniki walidacji z tych wielu rund, aby uzyskać oszacowanie wydajności predykcyjnej modelu.

Krzyżowa weryfikacja zapewni nam dokładniejsze oszacowanie wydajności modelu

## Walidacja krzyżowa K-fold

Obraz zawierający zrzut ekranu, Prostokąt, linia, Równolegle

Opis wygenerowany automatycznieTa technika polega na **losowym dzieleniu zestawu danych na k grup** lub fałdów (*folds*) o w przybliżeniu jednakowej wielkości. Pierwszy fałd jest przeznaczony do testowania, a model jest uczony na fałdach k-1.

**Proces powtarza się K razy i za każdym razem do walidacji wykorzystuje się inny zwój lub inną grupę punktów danych.**

10-krotna walidacja krzyżowa. Pomarańczowy blok jest zakładką używaną do testowania

Obraz zawierający Czcionka, tekst, zrzut ekranu, numer

Opis wygenerowany automatycznieGdy powtarzamy ten proces k razy, otrzymujemy k razy średni błąd kwadratu (MSE). MSE\_1, MSE\_2,… MSE\_K, więc **błąd k-Fold Cross Validation jest obliczany na podstawie średniej MSE z K fałd (foldów)**.

Błąd K-fold cross validation error

Zazwyczaj wartość K w K-fold wynosi 5 lub 10. Gdy K wynosi 10, odnosi się to również do 10-krotnej walidacji krzyżowej

**Zalety walidacji krzyżowej np. 10-krotnej walidacji krzyżowej**

* Czas obliczeń jest skrócony, ponieważ powtórzyliśmy ten proces tylko 10 razy, gdy wartość k wynosi 10.
* Zmniejszone odchylenie
* Każdy punkt danych musi zostać testowany dokładnie raz i jest wykorzystywany do uczenia k-1 razy
* Wariancja uzyskanego oszacowania zmniejsza się wraz ze wzrostem k

**Wady walidacji krzyżowej K np. 10-krotnej walidacji krzyżowej**

* Algorytm uczący jest obciążony obliczeniowo, ponieważ algorytm musi być powtórzony od zera k razy.

## Stratyfikacja K-Fold Cross Validation

W niektórych przypadkach może występować duży brak równowagi w zmiennych zależnych. Na przykład w zestawie danych dotyczącym ceny domów może istnieć duża liczba domów o wysokiej cenie. Lub w przypadku klasyfikacji może być kilka razy więcej próbek ujemnych niż próbek dodatnich.

W przypadku takich problemów dokonuje się niewielkich zmian w technice walidacji krzyżowej K Fold, tak że każda fałda zawiera w przybliżeniu taki sam procent próbek każdej klasy docelowej, co kompletny zestaw. Lub w przypadku problemów z prognozowaniem średnia wartość odpowiedzi wynosi tyle samo we wszystkich zakładkach. Ta odmiana jest również znana jako stratyfikowane K-Fold.

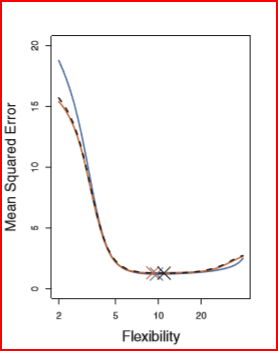
## Leave one out cross validation — LOOCV

Obraz zawierający tekst, zrzut ekranu, Czcionka, numer

Opis wygenerowany automatycznieW LOOCV dzielimy zestaw danych na dwie części. W jednej części mamy **jedną obserwację, która jest naszymi danymi testowymi, a w drugiej części mamy wszystkie inne obserwacje** z zestawu danych tworzących nasze dane uczące.

Jeśli mamy zestaw danych z n obserwacjami, to dane uczące zawierają obserwację n-1, a dane testowe zawierają 1 obserwację.

Ten proces jest iterowany dla każdego punktu danych, jak pokazano poniżej. Powtórzenie tego procesu n razy generuje n razy średni błąd kwadratu (MSE).  
LOOCV jest wariantem K-fold gdzie k=n.



Niebieska linia to prawdziwy błąd testu, czarna przerywana linia to błąd testu LOOCV, a pomarańczowa to 10-krotny błąd testu CV

**Zalety LOOCV**

* Znacznie mniej stronniczości, ponieważ wykorzystaliśmy cały zestaw danych do uczenia w porównaniu z podejściem zestawu walidacyjnego, w którym używamy tylko podzestawu (60% w naszym przykładzie powyżej) danych do uczenia.
* Brak losowości w danych uczących / testowych, ponieważ wielokrotne wykonywanie LOOCV przyniesie takie same wyniki

**Wady LOOCV**

* MSE będzie się różnić, ponieważ dane testowe wykorzystują jedną obserwację. Może to wprowadzać dużą zmienność. Jeśli punkt danych jest wartością odstającą, zmienność będzie znacznie wyższa.
* Wykonanie jest kosztowne, ponieważ model musi zostać uruchomiony aż n razy.

## Leave-P-Out Cross Validation - uogólnienie LOOCV

Podejście to pozostawia punkty danych p poza danymi uczącymi, tj. Jeśli w pierwotnej próbce znajduje się n punktów danych, wówczas próbki n-p są wykorzystywane do uczenia modelu, a punkty p są używane jako zestaw walidacyjny lub testowy. Czynność tę powtarza się dla wszystkich kombinacji, w których można w ten sposób oddzielić pierwotną próbkę, a następnie błąd uśrednia się dla wszystkich prób, aby uzyskać ogólną wydajność modelu.

Ta metoda jest wyczerpująca w tym sensie, że musi nauczyć i zweryfikować model dla wszystkich możliwych kombinacji, a dla umiarkowanie dużego p może stać się niewykonalna obliczeniowo.

## Walidacja krzyżowa szeregów czasowych

Losowe dzielenie danych szeregów czasowych nie może być stosowane, ponieważ jest niedopuszczalne aby dane związane z czasem zostaną pomieszane. Dlatego potrzebujemy innego podejścia do przeprowadzania weryfikacji krzyżowej.

W celu walidacji krzyżowej szeregów czasowych stosujemy **łączenie w przód** również określane jako **rolling-origin**. Początek, na którym opiera się prognoza, zmienia się w czasie.

W szeregach czasowych walidacja krzyżowa każdej jednostki czasu stanowi dane testowe i zakładamy, że dane z poprzedniej jednostki czasowej są zestawem uczącym.

D1, D2, D3 itd. Są danymi każdej jednostki czasowej (np. dnia), a dni zaznaczone na niebiesko służą do uczenia, a dni zaznaczone na żółto do testu.

Obraz zawierający tekst, zrzut ekranu, numer, kwadrat

Opis wygenerowany automatycznieRozpoczynamy uczenie modelu z minimalną liczbą obserwacji i wykorzystujemy dane z następnego dnia do testowania modelu, ciągle poruszamy się po zbiorze danych.

Zapewnia to, że uwzględniamy aspekt szeregów czasowych danych do prognozowania.