

WPROWADZENIE

Współczesna ***analiza danych*** jest rozwijana w obrębie dwóch obszarów nauki ***informatyki*** i ***statystyki***.

Statystyka dostarcza głównie modele oraz opisy własności teoretycznych modeli analizy danych.

Informatyka dostarcza głównie mechanizmy gromadzenia danych oraz algorytmy obliczeniowe związane z przetwarzaniem i analizą danych.

Obecnie jednak granice pomiędzy informatyką i statystyką się zacierają, w ramach wspólnej **nauki o danych** (*ang.* data science).

Główny ***cel analizy*** danych polega na opracowaniu modelu, który pozwala efektywnie interpretować i wykorzystywać dane posiadane w danej chwili oraz dane, które można otrzymać w przyszłości.

PRZYKŁADY WPROWADZAJĄCE

- wycena nieruchomości;
- filtrowanie spamu, kategoryzacja wiadomości:

Powiadomienie Powiadomienie

Przełączamy wszystkich użytkowników naszych kont e-mail na nowy, ulepszony interfejs Zimbra. Już niedługo będziesz miał dostęp do swojej poczty e-mail tutaj, więc oszczędzaj sobie stresu i ukończ proces, logując się na nowej stronie internetowej: **Kliknij tutaj, aby się zalogować**

Dzięki

Szczęśliwego Nowego Roku. Dlaczego milczałam?

Dziś wracam z mojej podróży i milczysz przez pocztę, którą do ciebie wysłałem od zeszłego tygodnia, proszę daj mi znać powód, dla którego milczałeś. Wyobrażałem sobie, dlaczego nie odpowiedziałeś mi bardzo ważna Proszę Szanowny, potrzebuję twojej szczerości? Czy mogę ufać, że przekażecie sumę 12.500.000.00 milionów USD, na moje konto, jeśli to możliwe, wróć do mnie po więcej szczegółów, czekam na twoją odpowiedź i uprzejmie daj mi znać, aby nie milczeć?

Pan Tony Siruno.

- rozpoznawanie znaków;
- analiza marketingowa rynku;
- inne.

SCHEMAT ANALIZY DANYCH (1)

Gromadzenie danych obejmujące zarówno określenie czynności związane z planowaniem eksperymentu lub ustaleniem sposobu pobierania danych, określenie modelu danych jak również zagadnienia związane z przechowywaniem danych.

Przetwarzanie wstępne obejmujące czynności wstępne pomiędzy wczytaniem danych z bazy danych lub pliku a właściwą analizą.

Analizę i wnioskowanie obejmuje czynności, które prowadzą do wyciągnięcia wniosków praktycznych z istniejących danych. Czynności te mogą obejmować przykładowo zadania: klasyfikacji, regresji, wyszukiwania wzorców i inne.

Czynności te często używane są niezależnie.

SCHEMAT ANALIZY DANYCH (2)

Gromadzenia danych może być *nieplanowanym* oraz *planowanym*.

W systemach informatycznych gromadzenie danych jest podporządkowane głównie funkcjonalności systemu, są takie dane gromadzone w sposób ciągły, a analizy danych wykonywane są zwykle w sposób doraźny (*nieplanowane*).

Systemy, w których sposób gromadzenia danych jest ściśle podporządkowany zadaniom analizy danych (*planowane*).

Informatycy lub statystycy nie zajmują się zwykle samym gromadzeniem danych, a wykorzystują oni najczęściej dane dostępne. Dane muszą zostać odpowiednio wyczyszczone i przygotowane.

RODZAJE ZADAŃ

Najważniejsze komponenty typowego zadania analizy danych:

- rodzaj zadania analizy danych;
- rodzaj użytego modelu;
- funkcja oceny modelu lub miary jakości modelu;
- rodzaj metody dopasowującej model do danych;
- sposób zarządzania danymi.

Podstawowe zadania analizy danych:

- identyfikacja rozkładu;
- klasteryzacja (grupowanie) danych;
- klasyfikacja;
- regresja.

IDENTYFIKACJA ROZKŁADU

Identyfikacja rozkładu łącznego wszystkich zmiennych polega na dopasowaniu ciągłego, dyskretnego lub mieszanego rozkładu do skończonej próby.

Mając rozkład łączny wszystkich zmiennych możemy sprowadzić większość zadań wnioskowania z danych do zwykłych zadań z rachunku prawdopodobieństwa.

Zadanie identyfikacji rozkładu nie może być jednoznacznie rozwiązane bez dodatkowych założeń.

Ponadto dokładny proces wnioskowania jest w ogólności zadaniem złożonym obliczeniowo.

Zadanie identyfikacji rozkładu wymaga zwykle założeń odnośnie **postaci funkcyjnej rozkładu** danych oraz z zadaniem identyfikacji rozkładu związane są ściśle **zadania klasteryzacji danych**.

KLASTERYZACJA

Klasteryzacja – zadanie podziału próbki obiektów na podzbiory, zwane klastrami (grupami, klasami) w taki sposób, aby każda grupa zawierała podobne obiekty, a obiekty różnych grup znacznie się różniły pomiędzy sobą.

Klasteryzacja jest procedurą opisową, ona nie robi żadnych wniosków statystycznych, ale daje możliwość wprowadzić wstępną analizę i zbadać "strukturę danych".

Na dzień dzisiejszy opracowano zostało ponad 100 różnych algorytmów klasteryzacji. Należy podkreślić, że w rezultacie wykorzystania różnych metod klasteryzacji mogą być otrzymane klastry różnych form (przy tym łączne i rozłączne), czyli mogą być otrzymane różne rezultaty, co jest właściwością działania tego lub owego algorytmu. Daną właściwość należy uwzględniać przy wyborze metody klasteryzacji.

KLASYFIKACJA

Klasyfikacja –automatyczne określenie przynależności jakiegoś reprezentanta do danej klasy na podstawie wcześniej zgromadzonych informacji.

Dla wykonania klasyfikacji za pomocą metod matematycznych należy mieć formalny opis obiektu (takim opisem w naszym przypadku jest baza danych). Każdy zapis bazy danych niesie w sobie informację o pewnej właściwości obiektu.

REGRESJA

Regresja to metoda statystyczna służąca określeniu związku pomiędzy różnymi wielkościami i przewidywaniu (predykcji) nieznanymi wartościami jednych wielkości na podstawie znanych wartości innych.

Regresja jest sposobem aproksymacji danych, czyli ich przybliżania.

Model regresyjny jest formalnym opisem stochastycznej zależności różnego rodzaju zjawisk od czynników je kształtujących, wyrażonym w formie odpowiedniego równania matematycznego.

KRÓTKO O DANYCH

Ogólnie **dane** przedstawiają sobą fakty, teksty, wykresy, obrazy, dźwięki, analogowe lub liczbowe wideo-segmenty itd. Dane, które są otrzymane w rezultacie pomiarów, eksperymentów, operacji arytmetycznych i logicznych, muszą być przedstawione w postaci, nadającej się do przechowywania, przekazywania i obróbki.

Dane pierwotne – oryginalne informacje zbierane w ściśle określonym celu. Zaletą jest ich wysoka skuteczność wynikająca z określenia celu odnoszącego się bezpośrednio do problemu badania natomiast wadą jest wysoki koszt i czasochłonność.

Dane wtórne – dane, które już istnieją, zostały zebrane i opracowane w innym celu. Zaletą jest ich niski koszt, a wadą ograniczona dokładność i dostosowanie do potrzeb, co wynika z odmiennego celu, dla którego były gromadzone.

DANE JAKOŚCIOWE I ILOŚCIOWE

Znaczna ilość informacji naukowej jest zapisywana w postaci liczb, co pozwala manipulować takimi danymi z użyciem statystycznych metod matematycznych. Takie dane są **ilościowe**. Głównym problemem zbierania danych ilościowych jest opracowanie precyzyjnych narzędzi w postaci pytań ankietowych, skali albo testów.

Istnieje ważna informacja, którą nie można zredukować do postaci liczb. Takie dane nazywa się **jakościowe**. Werbalne pojęcia i relacje między nimi są mniej dokładne niż liczby i odpowiednie łączy. Sprawia to, że badania jakościowe są bardziej zależne od właściwości określenia znaczenia słów, od opracowania pojęć i określenia relacji między nimi. W przeciwieństwie do badań ilościowych, w badaniach jakościowych nie istnieje powszechnie akceptowanej analizy danych.

MACIERZ DANYCH

Formalne dane to para $(\mathbf{U}, \mathbf{A}')$:

- $\mathbf{U} = \{\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_n\}$ - zbiór obiektów (przykładów, próbek, rekordów, przypadków, pomiarów itd.);
- $\mathbf{A}' = \{\mathbf{a}'_1, \mathbf{a}'_2, \dots, \mathbf{a}'_m\}$ - zbiór atrybutów (zmiennych, kolumn, cech itd.).

Zmienne mogą występować w postaci **ciągłej** lub **dyskretnej**. Dla postaci dyskretnej wyróżniają podtypy:

- zmienne binarne: $\{0, 1\}, \{-1, 1\}, \{true, false\}, \{tak, nie\}$;
- zmienne z naturalnym porządkiem: np., $\{\text{mało, średnio, dużo}\}$, wykształcenie, stan cywilny itd.;
- zmienne nominalne: dyskretne nieuporządkowane – kategorie, np., kolor oczu, zawód.

PRZYKŁAD

ID klienta	Kod pocztowy	Płeć	Dochód	Wiek	Stan cywilny	Kwota transakcji
1001	10048	M	75 000	C	M	5000
1002	J2S7K7	K	-40 000	40	W	4000
1003	90210		10 000 000	45	S	7000
1004	6269	M	50 000	0	S	1000
1005	55101	K	99 999	30	R	3000

BRAKI DANYCH

Braki danych – brak wartości wybranych zmiennych dla pewnej liczby obiektów.

Popularna metoda radzenia sobie z brakującymi wartościami jest po prostu pomijanie podczas analizy zapisów z brakującymi wartościami. Jednak może to być niebezpieczne, ponieważ brakujące wartości mogą tworzyć wzorce, a proste usunięcie zapisów z brakującymi wartościami doprowadziłoby do analizy obciążonego podzbioru danych.

Metody uzupełnienia braków za pomocą zastąpienia brakującej wartości na :

- pewną stałą, określoną przez analityka;
- wartością średnią lub modalną;
- wartością wygenerowaną losowo z obserwowanego rozkładu zmiennej.

PUNKTY ODSTAJĄCE

Punkty odstające (oddalone) są skrajnymi wartościami, które znajdują się blisko granic zakresu danych lub są sprzeczne z ogólnym trendem pozostałych danych.

Metody identyfikacji punktów oddalonych:

- graficzna: histogramy, wykresy rozrzutu, wykresy pudełkowe;
- standaryzacja;
- rozstęp międzykwartyłowy (IQR), który oblicza się jako różnica trzeciego i pierwszego kwartyli i może być zinterpretowane jako środkowe 50% danych. Wartość danych jest punktem oddalonym, jeżeli jest położona przynajmniej o 1,5 IQR poniżej kwartyla pierwszego lub jest położona przynajmniej o 1,5 IQR powyżej kwartyla trzeciego.

SKALOWANIE

- zachowująca zero:

$$X_{norm} = \frac{X}{|X|_{\max}};$$

- skalowanie:

$$X_{norm} = \frac{X - X_{\min}}{X_{\max} - X_{\min}};$$

- standaryzacja zmiennej:

$$X_{norm} = \frac{X - \bar{X}}{S}.$$

DYSKRETYZACJA (1)

Dyskretyzacja – zastąpienie zmiennej ciągłej na zmienną dyskretną (przyjmującą skończoną liczbę wartości) niosącym zbliżoną informację do oryginału.

Cele dyskretyzacji mogą być rozmaite, np.: uproszczenie danych w zamian za częściową utratę informacji, szczególnie, jeśli zmienna przyjmuje bardzo dużo różnych wartości; wychwycenie bardziej zgrubionych wzorców; podział danych na podzbiory; wykorzystanie algorytmów, działających tylko na danych kategoriowych; zmniejszenie „rozdzielczości” zmiennej itd.

Sposoby dyskretyzacji też są różne i zależą od konkretnego zadania, na ogół dyskretyzacja wykonuje się metodą przedziałową (przedziały mogą być równomierne, zależne od wpływu na zmienną decyzyjną, o określonych wartościach brzegowych itp.).

DYSKRETYZACJA (2)

Punkt odcięcia – (*ang.* cut point) – mianem tym określamy wszystkie wartości rzeczywiste w przestrzeni ciągłych wartości, które dzielą zakres danych na przedziały.

W chwili obecnej wyróżniamy następujące metody podziału procesu dyskretyzacji: metody lokalne i globalne, metody prymitywne i zaawansowane, metody statyczne i dynamiczne, metody kontrolowane i niekontrolowane oraz metody łączące i dzielące.

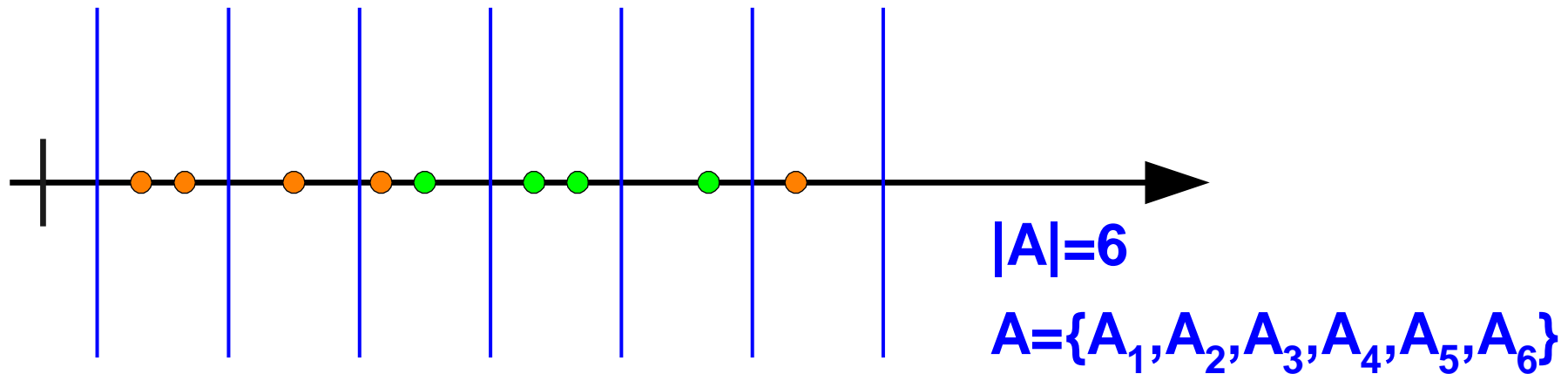
PRZEBIEG PROCESU DYSKRETYZACJI

Proces dyskretyzacji składa się z czterech głównych etapów, w skład których wchodzi:

- Sortowanie – malejąco lub rosnąco.
- Szacowanie – wybór punktów odcięcia, przedziałów sąsiadujących; oszacowanie miary zadowalającej.
- Dzielenie/Łączenie – dzielenie lub łączenie przedziałów.
- Zatrzymanie – określa kiedy proces dyskretyzacji zostanie zatrzymany.

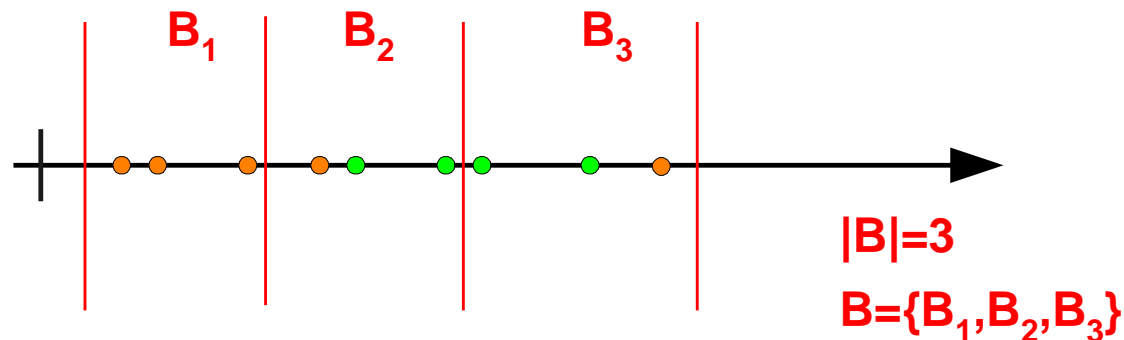
METODY DYSKRETYZACJI (1)

Metoda równej szerokości jest metodą niekontrolowaną. Nie korzysta ona również z informacji o klasach. Dzieli ona zakres wartości, na określoną wcześniej liczbę przedziałów.



METODY DYSKRETYZACJI (2)

Metoda równej częstości – zakres wartości jest dzielony na z góry ustaloną liczbę przedziałów. Nie są to jednak przedziały o równej szerokości. Granice przedziałów dobierane są w taki sposób by możliwie każdemu z nich odpowiadała taka sama liczba atrybutów. Załóżmy, że zbiór danych zawiera n przypadków, wybrana została liczba podprzedziałów równa k – proces dyskretyzacji metodą równej częstości doprowadzi do podziału na przedziały tak, by w każdym z nich znalazła się liczba przypadków równa n/k .



METODY DYSKRETYZACJI (3)

Metoda dyskretyzacji zstępującej jest kontrolowaną oraz dynamiczną metoda dyskretyzacji, wykorzystującą informację o klasie. Dla danego zbioru przykładów X oraz atrybutu X_j wykonuje się podział dyskretyzowanego atrybutu na podprzedziały za pomocą wartości progowych. Zgodnie z istotą podejścia zstępującego początkowo przyjmuje się cały zakres wartości jako jedyny przedział. Umieszczenie pierwszej wartości progowej dzieli go na dwa podprzedziały, z których każdy może być następnie podzielony na kolejne dwa podprzedziały itd. Dla takiego ogólnego algorytmu zstępującej dyskretyzacji najwygodniejsze jest sformułowanie rekurencyjne. W postaci miary podziału występują tu miara teorii informacji, czyli ważona entropia oraz w postaci kryterium stopu przyrost informacji.

BINARYZACJA

Binaryzacja – zastąpienie zmiennej dyskretnej, przyjmującej m wartości, m zmiennymi binarnymi, tzw. indykatorami: dla każdej zmiennej tylko jeden indykator wynosi 1, pozostałe – 0.

color		color=green	color=red	color=blue
green	→	1	0	0
red		0	1	0
red		0	1	0
green		1	0	0
blue		0	0	1

Ogólnie rozpatrujemy przestrzeń zmiennych (atrybutów) jako punkty (wektory) w przestrzeni wielowymiarowej, gdzie każda zmienna może reprezentować inny wymiar.

WYSOKA LICZBA POMIARÓW

Problem wysokiej liczby wymiarów - „**przekleństwo wymiarowości**”:

- coraz większa minimalna liczba przypadków niezbędna, aby uchwycić jakiejkolwiek zależności w danych;
- coraz większa liczba kombinacji zmiennych (i kombinacji wartości tych zmiennych);
- coraz większy promień odległości musi być wzięty pod uwagę, aby objąć ustaloną część przestrzeni.

Rozpowszechnionymi **metodami redukcji** liczby wymiarów są:

- usuwanie niektórych zmiennych (usuwane w pierwszej kolejności są zmienne, które mają niską wartość informacyjną, co można sprawdzać za pomocą różnych miar, np., miar korelacji);
- analiza składowych głównych (PCA – ang. principal component analysis).

WYGŁADZANIE DANYCH

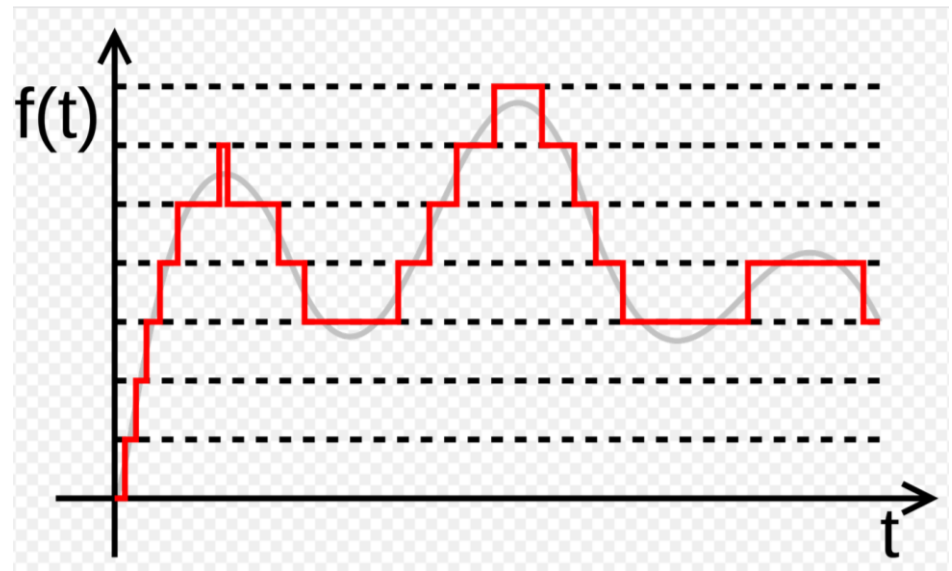
Wygładzanie danych – specjalne działanie uśrednienia za pomocą wielomianów, które zapewnia otrzymanie poprawionej wartości \bar{x}_i na podstawie wartości x_i oraz wartości znajdujących się obok ($\dots, x_{i-1}, x_i, x_{i+1}, \dots$), które są znane z pewnym błędem.

Często w analizie danych przy obróbce pojawia się potrzeba wygładzania danych. Najczęściej jest wykorzystywane przy **szeregach czasowych**, czyli przy zbiorze obserwacji statystycznych charakteryzujących zmiany poziomu zjawiska w czasie. Rozpatrzmy dokładniej ten problem później 😊.

KWANTYZACJA

Kwantyzacja - jest procesem redukcji zbioru możliwych wartości jakie może przyjmować zmienna reprezentująca kwantyzowane źródło. W szczególności kwantyzacja może polegać na aproksymacji zmiennej ciągłej przez zmienną dyskretną przyjmującą wartości ze skończonego zbioru dopuszczalnych wartości.

Inaczej mówiąc jest to reprezentacja dużego (w szczególności nieskończonego) zbioru wartości przez mniejszy, skończony zbiór wartości.



PRZYKŁAD

Źródło generuje liczby z zakresu $[-100, 100]$. Reprezentacja liczby: najbliższa liczba całkowita.

$$\{100,2 \ 100,3 \ 278,1 \ 300 \ 314,15926, \ 370,5\} \Rightarrow \\ \Rightarrow \{100 \ 100 \ 300 \ 300 \ 300 \ 400\}$$

Podstawowa cecha: tracimy informację o dokładnych wartościach danych wejściowych.

➤ **skalarna** – kwantowane są niezależne pojedyncze wartości; może być równomierna i nierównomierna;

➤ **wektorowa** – kwantowane są jednocześnie kilka wartości (co najmniej 2).

KWANTYZACJA SKALARNA

- dzielimy zbiór X na skończoną (przeliczalną) liczbę zbiorów (w przypadku jednowymiarowym odcinki postaci $[x_i, x_{i+1}]$);
- z każdego odcinka wybieramy reprezentantów – stanowią one tzw. alfabet kodowy;
- kwantyzacja polega na zastąpieniu punktu reprezentantem.

Przy kwantyzacji skalarnej należy uwzględniać rodzaj danych, czyli ich rozkład.

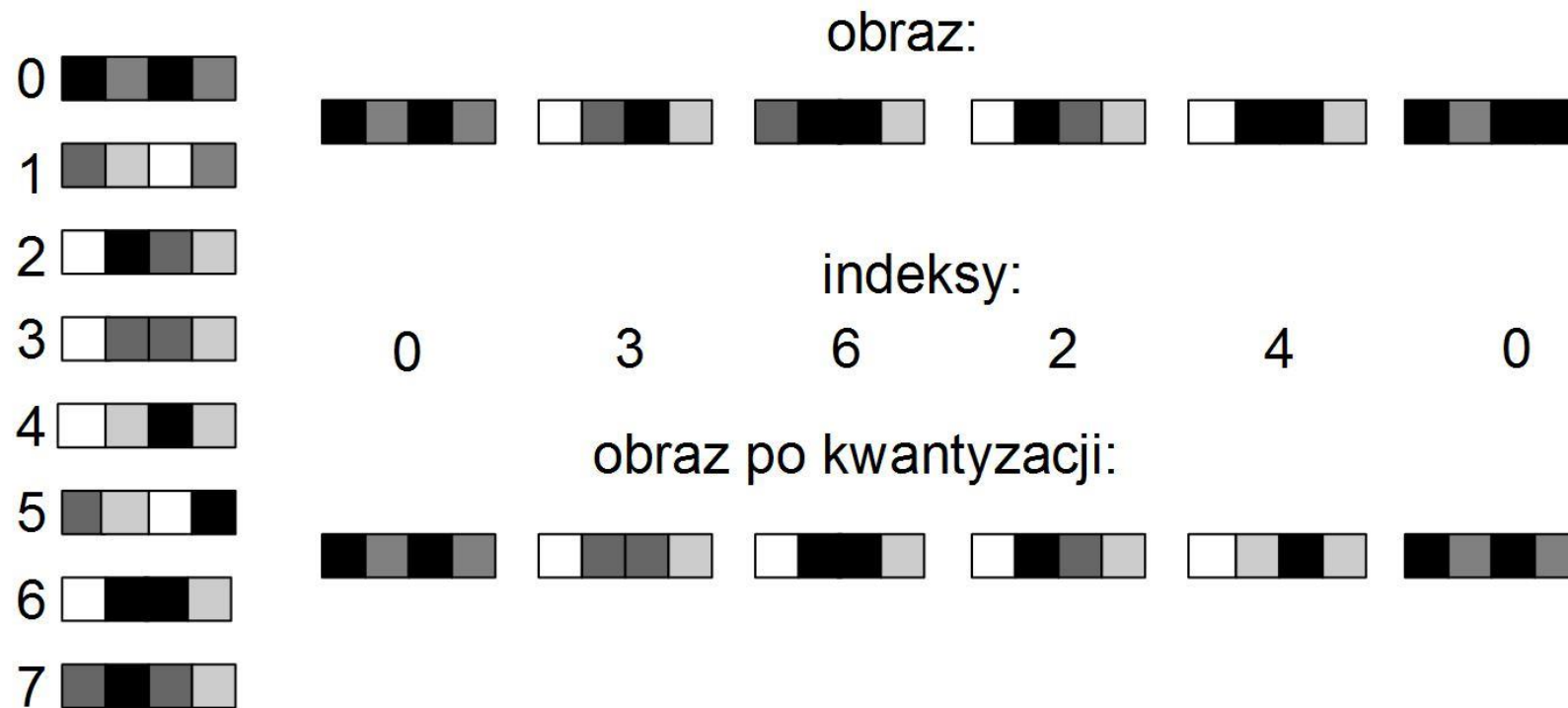
Przykładowo, jeśli w pokazanym wcześniej przykładzie 90% wartości pojawiłoby się w przedziale od -1 do 1, to zaokrąglenie do najbliższej liczby całkowitej nie było by właściwe. Wtedy dla przedziału od -1 do 1 lepiej było by zaokrąglać z dokładnością do 1/10 (o wiele mniejsza strata informacji).

KWANTYZACJA WEKTOROWA

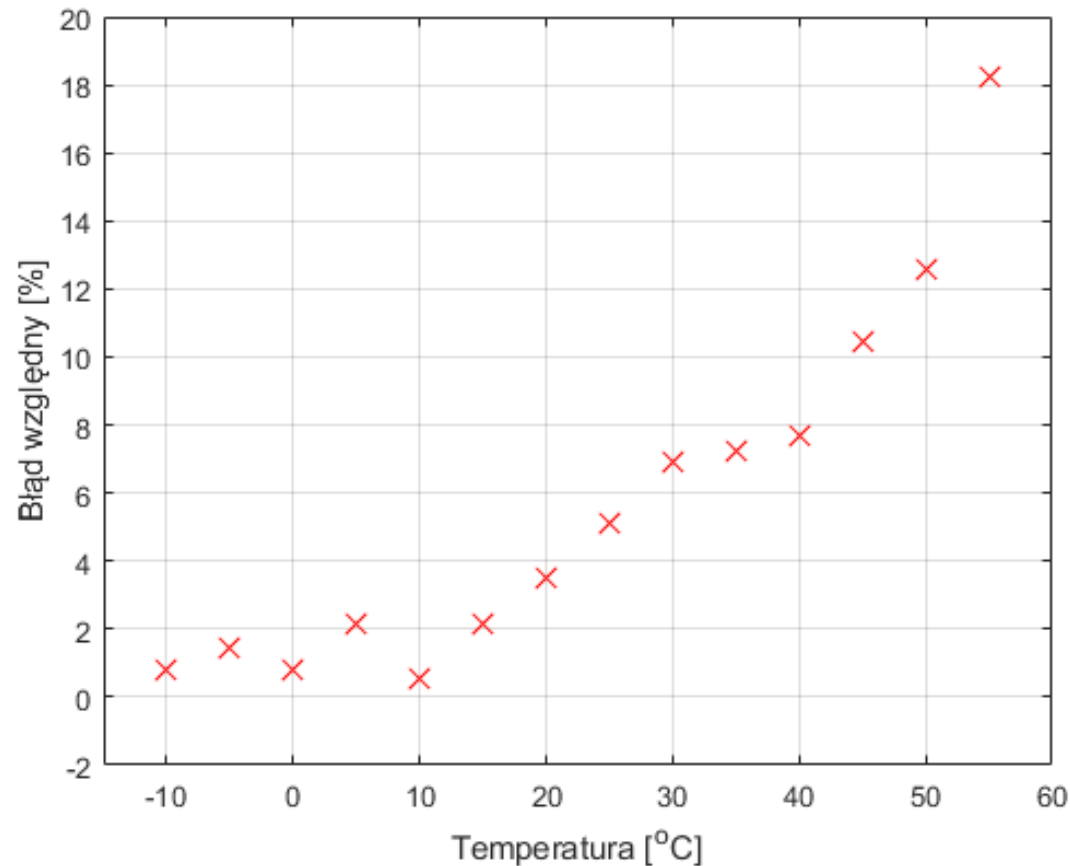
- oryginalny 24-bitowy obraz reprezentuje próba x_1, \dots, x_n ;
- mapa kolorów składa się 256 wektorów kodowych, tzw. **książka kodowa**, z_1, \dots, z_{256} , przy czym każdy wektor reprezentuje 24-bitowy kolor;
- każdy piksel oryginalnego obrazu zastępowany jest przez najbliższy wektor kodowy;
- dla przesłania pojedynczego piksela kanałem transmisyjnym wystarczy przesłać indeks jego wektora kodowego w książce kodowej.

Problemy wyboru książki kodowej i poszukiwania w niej najbliższej wartości tu nie rozpatrują się.

PRZYKŁAD



PRZYKŁADOWE WYNIKI EKSPERYMENTU



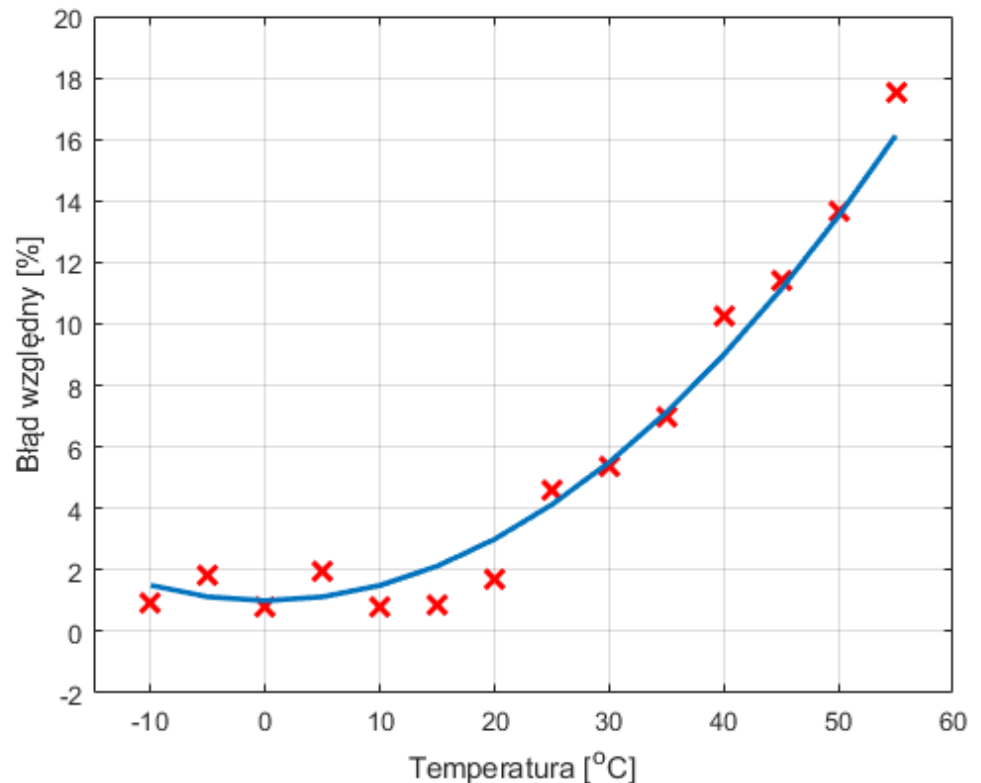
Charakterystyka temperaturowa dynamicznej wagi pojazdów, w której dokładność ważenia, charakteryzowana przez błąd względny (oś Y) zmienia się wraz ze zmianą temperatury (oś X).

ETAPY BUDOWANIA MODELU

- ***przyjęcie klasy modelu*** – a więc ogólnego wyrażenia matematycznego opisującego dane zjawisko lub obiekt, np.: wielomian 2-go stopnia. Klasę modelu wybiera inżynier na podstawie obserwacji kształtu charakterystyki, przesłanek teoretycznych, ogólnie wiedzy a’priori;
- ***obliczenia współczynników modelu*** – czyli konkretnych wartości liczbowych. Wartości te są obliczane przez algorytm identyfikacji, zazwyczaj w procesie iteracyjnego poszukiwania optimum funkcji celu.

APROKSYMACJA

Aproksymacja (ang. curve fitting) polega na dopasowaniu krzywej do punktów pomiarowych, przy czym krzywa nie przechodzi dokładnie przez te punkty, lecz odzwierciedla ogólny trend w danych. Celem jest więc znalezienie modelu matematycznego, który będzie reprezentatywny dla badanego zbioru danych.

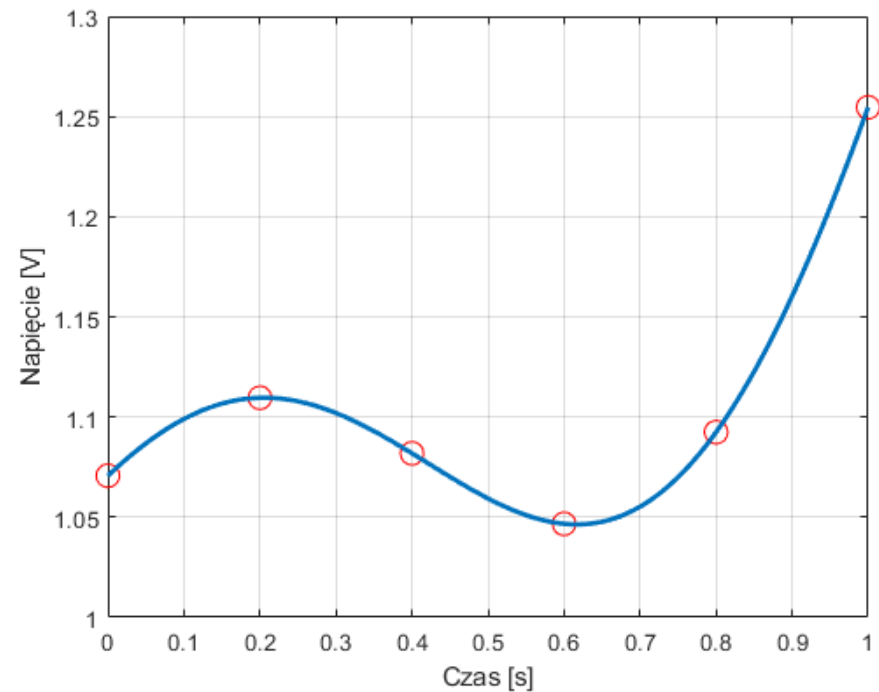


Aproksymacje stosujemy, gdy zależy nam na matematycznym opisie trendu, a dane pomiarowe charakteryzuje rozrzut wynikający np. z losowego charakteru wykonanego pomiaru.

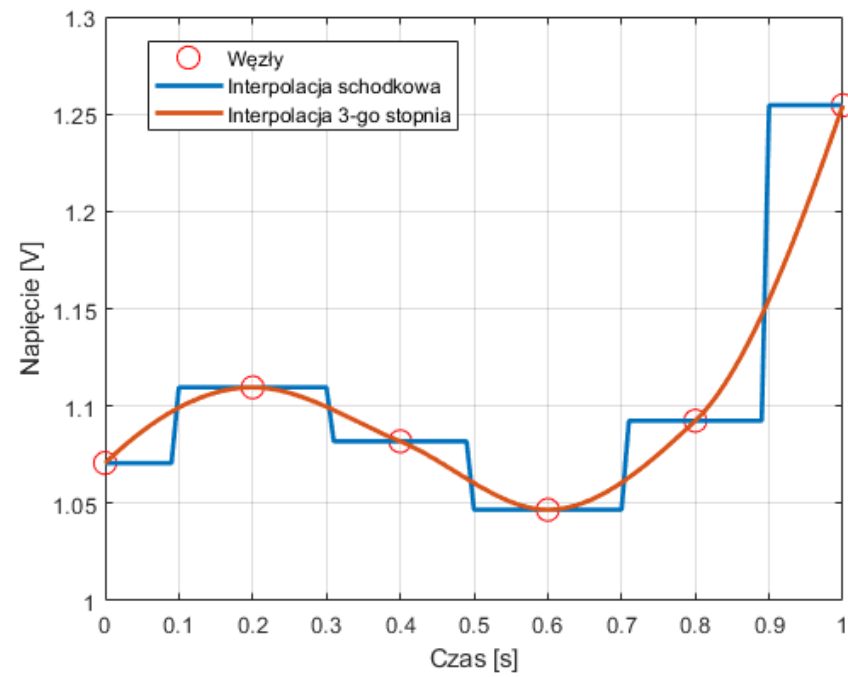
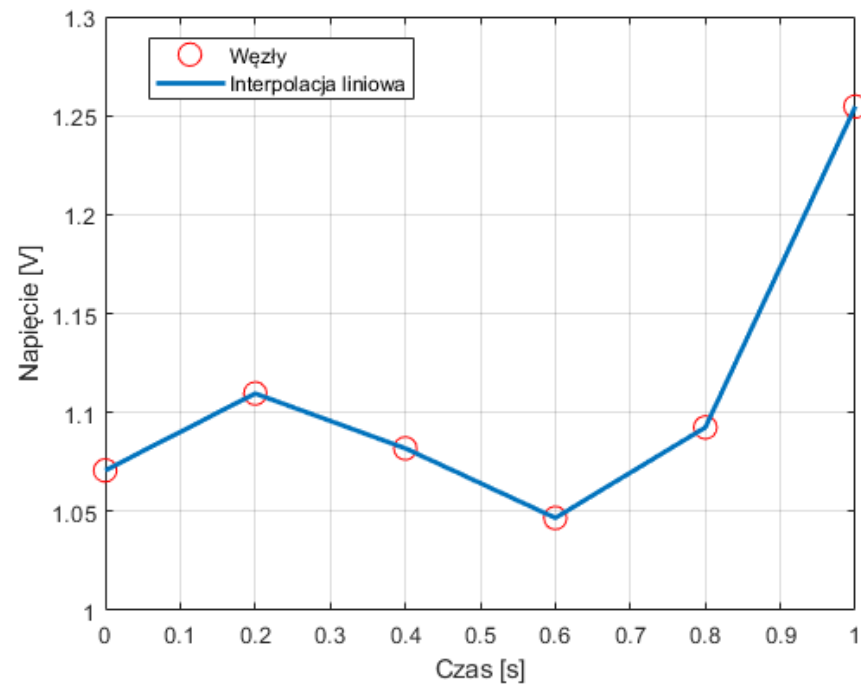
INTERPOLACJA (1)

Interpolacja (ang. interpolation) jest metodą numeryczną zbliżoną do aproksymacji z tą różnicą, że dopasowana do danych krzywa przechodzi dokładnie przez punkty pomiarowe. Innymi słowy interpolacja polega na dopasowaniu funkcji do danych w taki sposób, że funkcja ta przyjmuje konkretne wartości w punktach nazywanych **węzłami**.

Dla n punktów pomiarowych można dopasować wielomian rzędu $n-1$. Interpolacja jest często używana, do obliczania wartości funkcji pomiędzy węzłami.



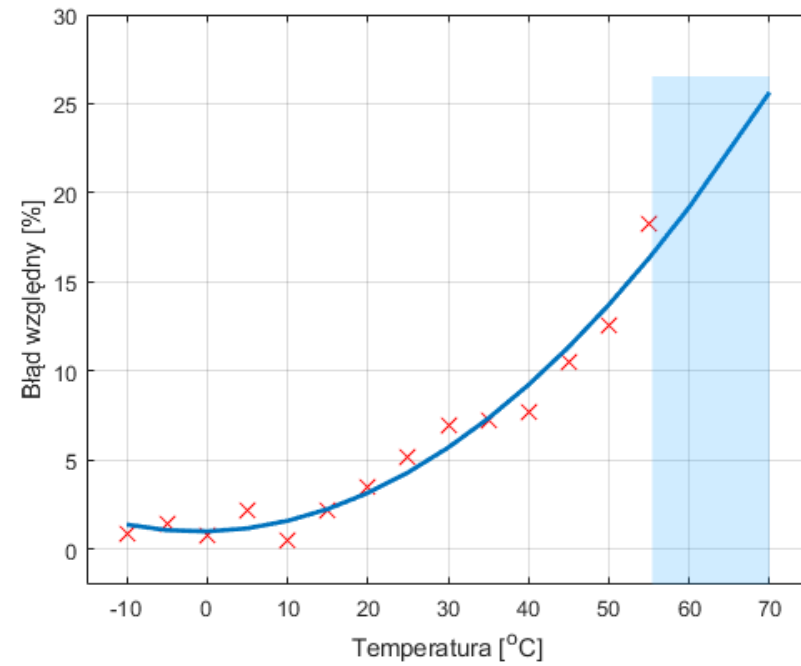
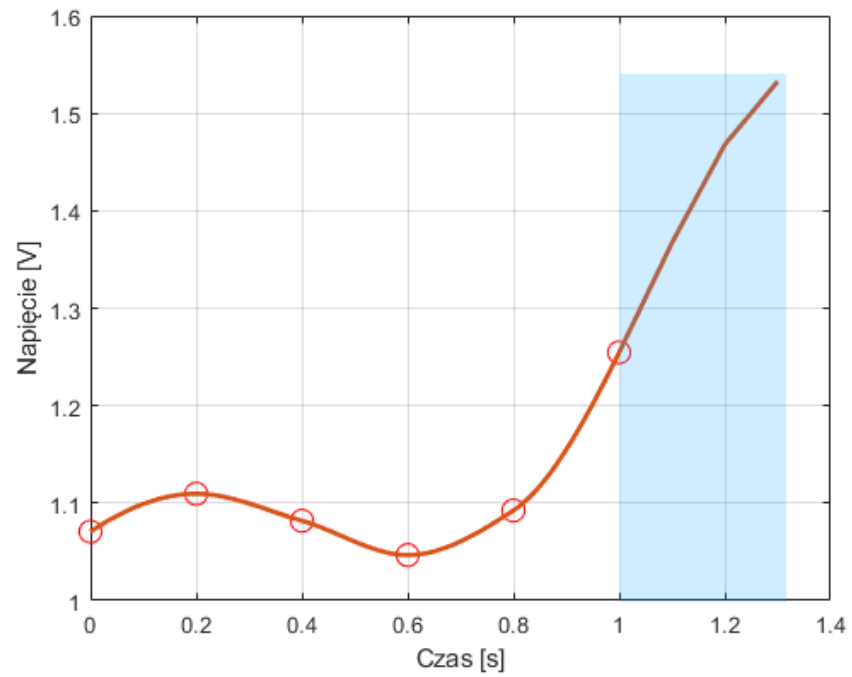
INTERPOLACJA (2)



EKSTRAPOLACJA (1)

Ekstrapolacja (*ang.* extrapolation) polega na obliczeniu wartości modelu, który został dopasowany do danych, poza zakresem dla którego te dane zgromadzono. Jest to więc wyznaczanie wartości funkcji na zewnątrz przedziału, w którym wartości tej funkcji są znane. Ekstrapolację można przeprowadzić zarówno dla modelu aproksymującego dane pomiarowe, jak i po interpolacji.

EKSTRAPOLACJA (2)



SPOSOBY UCZENIA

- **uczenie nadzorowane** (*ang.* supervised learning) – uczenie na podstawie przykładów;
- **uczenie częściowo nadzorowane** (*ang.* semi-supervised learning);
- **uczenie nienadzorowane** (*ang.* unsupervised learning) – nie są posiadane "klucze odpowiedzi" i muszą przez system zostać analizowane dane, szukane wzorce i odnajdywane relacje;
- **uczenie przez wzmacnianie** (*ang.* reinforcement learning) – otrzymane są gotowe zestawy dozwolonych działań, reguł i stwierdzeń; działając w ich ramach, dokonuje się analizy i obserwuje ich skutki; wykorzystuje reguły w taki sposób, aby osiągnąć pożądany efekt.

METODY OCENY DOKŁADNOŚCI MODELI

Podczas budowy modelu (regresji, klasyfikacji i innych), którego celem jest przewidywanie pewnych wartości na podstawie zbioru danych uczących poważnym problemem jest ocena jakości uczenia i zdolności poprawnego przewidywania.

Częstym błędem osób początkujących w zakresie analizy danych jest przeprowadzanie testów na tym samym zbiorze na którym system był uczony. Takie rozwiązanie nie jest poprawnym miernikiem jakości nauczonego modelu i prowadzi do wyników które są przeszacowane, czyli nadmiernie optymistyczne.

Ponieważ zwykle głównym celem budowy modeli predykcyjnych jest późniejsze wykorzystanie przewidywań modelu na danych niedostępnych podczas procesu uczenia więc opracowano szereg metod pozwalających na znacznie bardziej uczciwy pomiar dokładności.

KROSWALIDACJA

Teoria sprawdzianu krzyżowego została zapoczątkowana przez Seymoura Geissera. Pozwala ona właściwie ocenić trafność prognostyczną modelu predykcyjnego. Bez jej zastosowania nie można być pewnym czy model będzie dobrze działał dla danych, które nie były wykorzystywane do jego konstruowania.

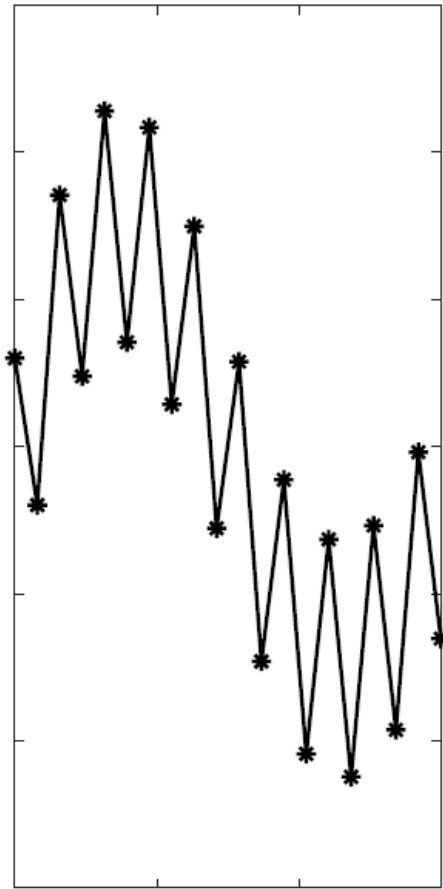
Walidacja krzyżowa, zwana także **kroswalidacją** lub **sprawdzian krzyżowy** jest techniką pozwalającą na oszacowanie skuteczności modelu, co daje możliwość zapobiec problemom przeuczenia modelu (*ang.* overfitting) lub niedouczenia (*ang.* underfitting).

PRZEUCZENIE I NIEDOUCZENIE (1)

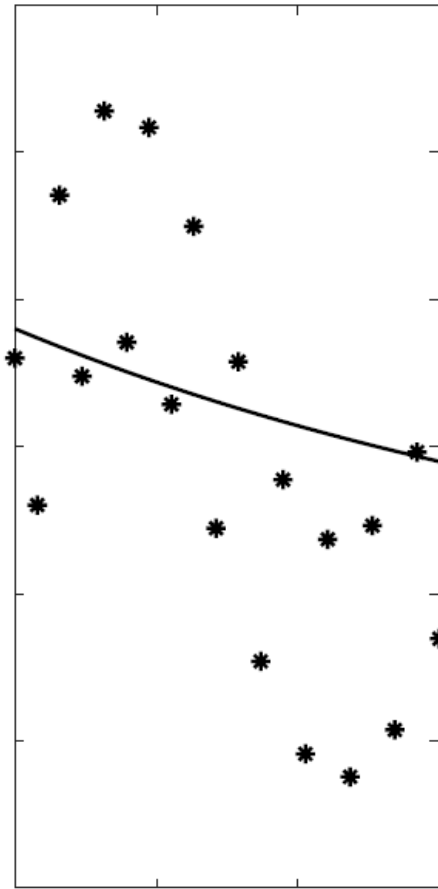
Przeuczenie modelu (nadmierne dopasowanie, przetrenowanie, czyli zbyt dużą liczbą parametrów w stosunku do rozmiaru próby na podstawie której model ten był konstruowany. Modele takie mogą świetnie pasować do danych uczących gdy w modelu jest wystarczającą złożoność, jednak będą dawały gorsze wyniki przy zastosowaniu do danych, z którymi się nie zetknęły podczas uczenia.

W celu uniknięcia nadmiernego dopasowania konieczne jest zastosowanie dodatkowych środków zapobiegawczych, które pozwalają stwierdzić, w którym momencie dalsze uczenie zaczyna prowadzić do powstania gorszego modelu.

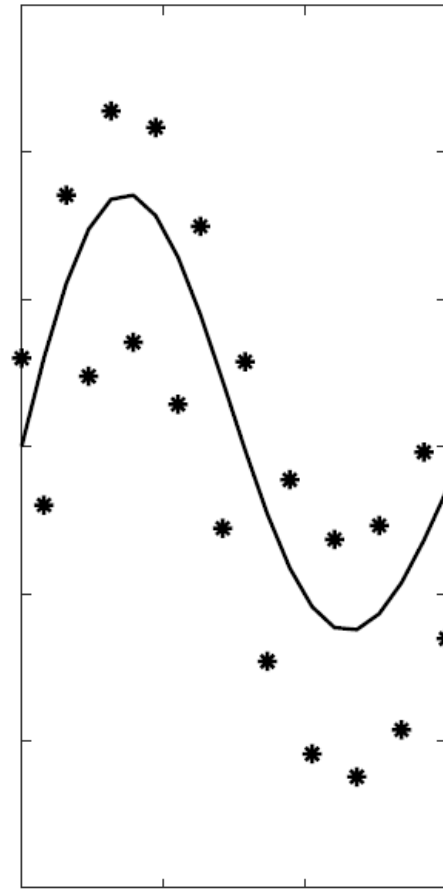
PRZEUCZENIE I NIEDOUCZENIE (2)



a)



b)



c)

PRZEUCZENIE I NIEDOUCZENIE (3)

Niedouczenie wynika z niewystarczającej liczby próbek uczących lub ze zbyt uproszczonego modelu zastosowanego w uczeniu.

Określenie, czy błędy wynikają z przeuczenia, niedouczenia lub jeszcze czegoś innego, jest problemem trudnym i często wymaga przeprowadzenia wielu eksperymentów.

Krosvalidacja polega na wielokrotnym powtarzaniu trzech kroków:

- podział zbioru danych na treningowy (uczący) i testowy;
- uczenie modelu;
- ocena jakości modelu;

a następnie uśrednieniu otrzymanych wyników.

WALIDACJA PROSTA

Metoda wydzielania (ang. holdout cross-validation) albo **walidacja prosta** jest jedną z klasycznych technik krosvalidacji pozwalających na oszacowanie skuteczności uogólniania modelu.

Polega ona na podziale całej próby na dwa niezależne podzbiory tzw. **uczący** i **testowy**. Taki podział zwykle realizowany jest w stosunku 70% przypadków to część ucząca zbioru, oraz 30% stanowiąca część testową. Idea oceny modelu lub doboru odpowiednich parametrów modelu sprowadza się wówczas do nauczania modelu na części uczącej oraz przetestowania go na części testowej, która nie była wykorzystywana w procesie uczenia modelu.

PODZIAŁ NA TRZY PODZBIORY

Udoskonalona metoda wydzielania bazuje na podziale danych na trzy podzbiory: ***treningowy, walidacyjny i testowy***. Zbiór treningowy służy do uczenia modelu, z kolei zbiór walidacyjny do weryfikacji skuteczności. Dopiero w momencie osiągnięcia satysfakcjonujących wyników wykorzystuje się zestaw testowy do oszacowania skuteczności modelu.

Jedną z częściej wymienianych *wad* metody wydzielania jest duża wrażliwość oszacowania na sposób podziału danych. Innymi słowy, wyniki mogą istotnie różnić się w zależności od wielkości poszczególnych podzbiorów.

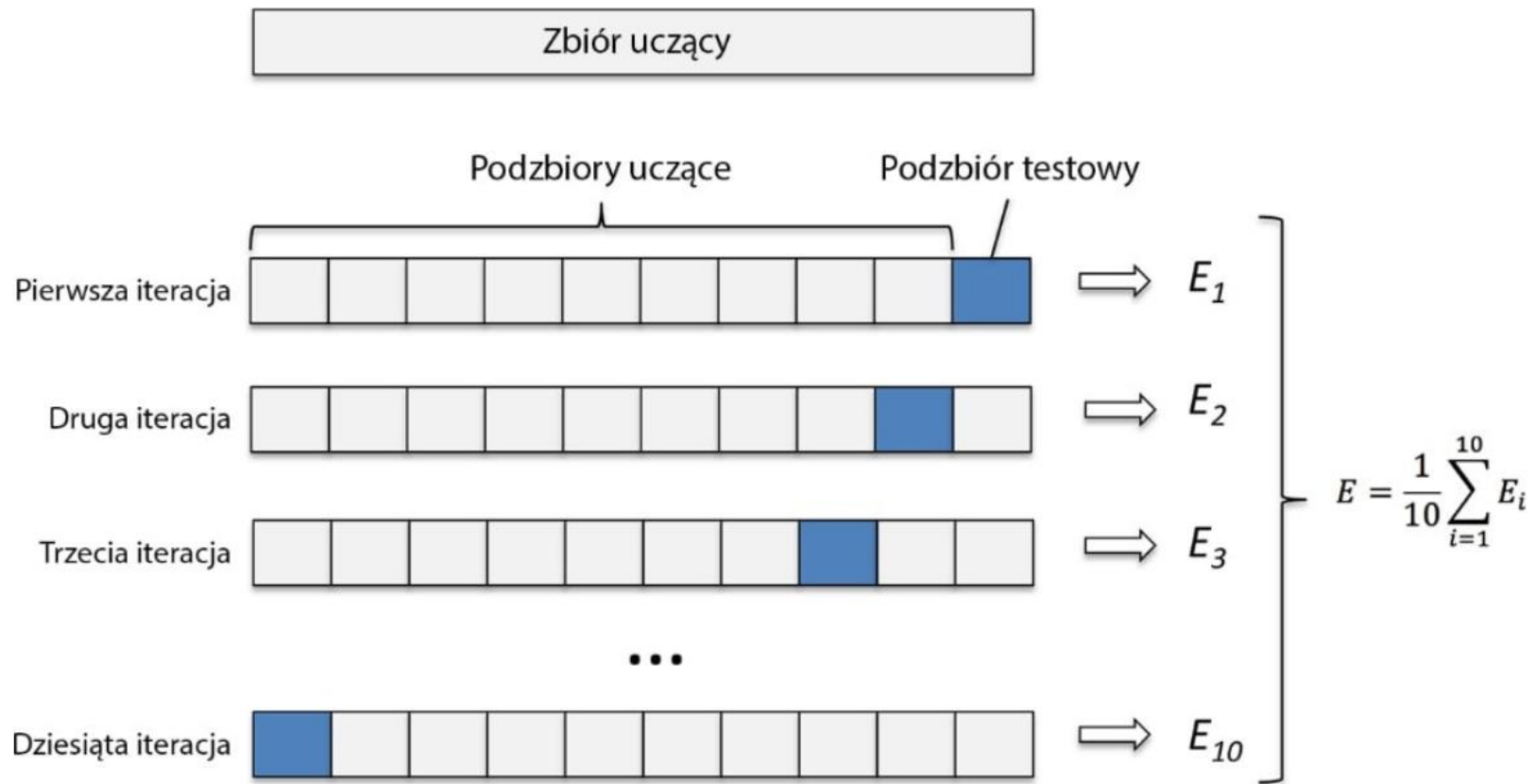
K-KROTNA WALIDACJA (1)

k-krotna walidacja krzyżowa (*ang. k-fold cross-validation*) dokonuje losowego podziału zbioru danych na k podzbiorów, gdzie $k-1$ podzbiorów wykorzystywanych jest do trenowania modelu.

Proces powtarzany jest k -krotnie, a w efekcie otrzymujemy k modeli i oszacowań skuteczności. Ostateczny wynik oszacowania uzyskiwany jest poprzez uśrednienie wyników uzyskanych ze wszystkich iteracji.

k -krotna walidacja krzyżowa dokonuje próbkowania bez zwracania, co redukuje wariancję oszacowania modelu. Innymi słowy, każda próbka będzie występowała tylko raz w zestawie treningowym lub testowym.

K-KROTNA WALIDACJA (2)



METODA LEAVE-ONE-OUT (LOO)

Metoda leave-one-out - metoda minus jednego elementu (*ang.* leave-one-out-method)

Jest to odmiana k -krotnej walidacji krzyżowej. W tym wypadku liczba podzbiorów jest równa liczbie próbek uczących $k = n$. W związku z tym, metoda LOO wymaga znacznie większej mocy obliczeniowej w stosunku do k -krotnego sprawdzianu krzyżowego, zatem zaleca się jej stosowanie w przypadku małych zbiorów danych.

BOOTSTRAP

Dla małych prób oprócz k -krotnej walidacji wykorzystywana jest tzw. metoda **bootstrap**. Polega ona na wielokrotnym repróbkowaniu elementów z próby uczącej. W metodzie tej wykonuje się losowanie ze zwracaniem z próby uczącej o liczności n (np. próba ucząca = {1,2,3,4,5} → próba bootstrap {1,1,2,5,5}), tworzy się m prób bootstrap.

Z reguły każda próba bootstrap nie zawiera wszystkich elementów próby oryginalnej. Dla $m = n$:

$$P(\text{nie wylosowania ustalonego elementu do próby bootstrap}) = \\ = \left(\frac{n-1}{n} \right)^n = \left(1 - \frac{1}{n} \right)^n \approx e^{-1} \approx 0,368$$

Oznacza to, że w próbie uczącej na podstawie bootstrapu średnio występuje $0,632 \cdot n$ różnych obserwacji.

TEORIA INFORMACJI

Teoria informacji – dyscyplina zajmująca się problematyką informacji oraz metodami przetwarzania informacji, np. w celu transmisji lub kompresji. Naukowo teoria informacji jest blisko powiązana z matematyką dyskretną, a z jej osiągnięć czerpią takie dyscypliny jak informatyka i telekomunikacja.

Za ojca teorii informacji uważa się Claude'a E. Shannona, który prawdopodobnie po raz pierwszy użył tego terminu w 1945 roku w swojej pracy zatytułowanej *A Mathematical Theory of Cryptography*. Natomiast w 1948 roku w kolejnej pracy pt. *A Mathematical Theory of Communication* przedstawił najważniejsze zagadnienia związane z tą dziedziną nauki.

NIEPEWNOŚĆ ZDARZEŃ LOSOWYCH

Niech zdarzenie losowe polega na występowaniu jednego z niezależnych doświadczeń x_1, x_2, \dots, x_n z odpowiadającym mu prawdopodobieństwem p_1, p_2, \dots, p_n , przy czym $p_1 + p_2 + \dots + p_n = 1$.

Prawdopodobieństwa te są znane, ale to jest wszystko, co znamy odnośnie zdarzenia, które wystąpi.

W praktyce ważnym jest umiejętność w postaci liczbowej określić taki stan niepewności dla różnych zdarzeń losowych, aby móc porównywać od tej strony.

Taka szukana charakterystyka musi być funkcją liczby n możliwych występowania doświadczeń.

- dla $n = 1$: $f(1) = 0$.
- przy zwiększeniu liczby możliwych występowania funkcja ta musi wzrastać.

ILOCZYN ZDARZEŃ

A i B - zdarzenia niezależne (A ma k równie prawdopodobnych wyników, a B ma l).

AB - iloczyn (wspólne występowanie) zdarzeń A i B

Niepewność pojawienia się zdarzenia AB jest większa, niż niepewność zdarzenia A , ponieważ do tej niepewności dochodzi jeszcze niepewność zdarzenia B .

Wymaganie: niepewność zdarzenia AB jest równa sumie niepewności zdarzeń A i B :

$$f(kl) = f(k) + f(l),$$

liczba kl wskazuje możliwą liczbę wyników zdarzenia AB .

Udowodnione zostało, że funkcja **logarytmiczna** $\log n$ jest jedyną ciągłą funkcją dla argumentu $n \in \mathbb{R}$, która spełnia powyższe równanie.

ENTROPIA

Miara niepewności występowania losowego zdarzenia X :

$$H(X) = -\sum_{i=1}^n p_i \log p_i.$$

nazywa się **entropią** zdarzenia X .

Oznaczenie: $H(X)$ lub $H(p_1, p_2, \dots, p_n)$.

podstawa logarytmowania	nazwa
2	bit
3	trit
10	dit
e	nat (<i>ang.</i> natural logarithm)

WŁASNOŚCI ENTROPII (1)

- $H(X)$ jest ciągłą po każdej zmiennej;
- jeśli wszystkie prawdopodobieństwa są równe

$$p_1 = p_2 = \dots = p_n = \frac{1}{n},$$

to $H(X)$ monotonicznie rosnąca funkcja po n :

$$H\left(\frac{1}{n}, \frac{1}{n}, \dots, \frac{1}{n}\right) \leq H\left(\frac{1}{n+1}, \frac{1}{n+1}, \dots, \frac{1}{n+1}\right),$$

czyli przy równych prawdopodobieństwach zdarzeń, niepewność będzie tym większa, im więcej jest wyników zdarzeń;

- gdy jakieś zdarzenie „rozpada się” na dwa kolejne, to entropię należy liczyć, jako sumę ważoną składowych.

WŁASNOŚCI ENTROPII (2)

➤ $H = 0$ wtedy i tylko wtedy, gdy jedno z prawdopodobieństw jest równe 1, a pozostałe są równe zero (miara niepewności na pewno występującego zdarzenia jest równa 0);

➤ dla liczby zdarzeń n funkcja H przyjmuje wartość maksymalną, gdy prawdopodobieństwa zejść zdarzeń są takie same:

$$\max H(p_1, p_2, \dots, p_n) = H\left(\frac{1}{n}, \frac{1}{n}, \dots, \frac{1}{n}\right) = \log n;$$

➤ entropia iloczynu niezależnych zdarzeń AB jest równa sumie entropii tych zdarzeń:

$$H(AB) = H(A) + H(B).$$

ENTROPIA WARUNKOWA

A i B - zależne.

$$H(AB) = - \sum_{i=1}^n \sum_{j=1}^m P(a_i b_j) \log P(a_i b_j).$$

Entropia warunkowa zdarzenia B pod warunkiem, że zdarzenie A przyjęło wartość a_i :

$$H(B | a_i) = - \sum_{j=1}^m P(b_j | a_i) \log P(b_j | a_i)$$

Entropia warunkowa zdarzenia B pod warunkiem zajścia zdarzenia A :

$$H(B | A) = \sum_{i=1}^n P(a_i) H(B | a_i).$$

Czyli dla entropii:

$$H(AB) = H(A) + H(B | A) = H(B) + H(A | B)$$

ENTROPIA KRZYŻOWA

Entropia krzyżowa (*ang.* cross entropy or log loss) to metoda porównywania dwóch rozkładów prawdopodobieństwa.

Prawdziwy (oczekiwany) rozkład prawdopodobieństwa zmiennej losowej X :

$$P(X) = \{p_1, p_2, \dots, p_n\}.$$

Model, który przewiduje rozkład zmiennej losowej X :

$$Q(X) = \{q_1, q_2, \dots, q_n\}.$$

Można użyć entropii krzyżowej jako metryki określającej jakość predykcji wykonywanych przez model:

$$H(P, Q) = -\sum_{i=1}^n p_i \log_b(q_i).$$

$$P = Q: H(P, P) = -\sum_{i=1}^n p_i \log_b(p_i).$$

DYWERGENCJA KULLBACKA-LEIBLERA

Dywersgencja Kullbacka-Leiblera (*ang.* Kullback–Leibler divergence, KL), zwana też **entropią względną** lub **relatywną entropią** jest różnicą pomiędzy entropią krzyżową i entropią. Określa ona rozbieżność między dwoma rozkładami prawdopodobieństwa:

$$D(P, Q) = H(P, Q) - H(P, P) = \sum_{i=1}^n p_i \log_b \frac{p_i}{q_i}.$$

Dla rozkładów ciągłych:

$$D(p, q) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} dx.$$

NIERÓWNOŚĆ GIBBSA

Klasyczna entropia względna jest wartością nieujemną:

$$D(P, Q) \geq 0.$$

Nierówność ta nazywa się nierównością Gibbsa

$$D(P, Q) = \sum_{i=1}^n p_i \log_b \frac{p_i}{q_i} = - \sum_{i=1}^n p_i \log_b \frac{q_i}{p_i} \stackrel{?}{\geq} 0$$

(Za pomocą rozłożenia w szereg Taylora można udowodnić, że $-\ln x \geq (1 - x)$).

$$- \sum_{i=1}^n p_i \log_b \frac{q_i}{p_i} \geq \sum_{i=1}^n p_i \left(1 - \frac{q_i}{p_i} \right) = \sum_{i=1}^n (p_i - q_i) = \sum_{i=1}^n p_i - \sum_{i=1}^n q_i = 1 - 1 = 0 \quad \text{co i}$$

należało udowodnić.

INFORMACJA WZAJEMNA (1)

Informacja wzajemna – miara zależności pomiędzy dwoma zmiennymi losowymi.

$$I(X, Y) = \sum_{y \in Y} \sum_{x \in X} p(x, y) \log \frac{p(x, y)}{p(x)p(y)},$$

gdzie:

$p(x, y)$ - wspólny rozkład prawdopodobieństwa (ang. *joint probability distribution*) X i Y ,

$p(x)$ i $p(y)$ - prawdopodobieństwa w rozkładach zmiennych X i Y .

Informacja wzajemna jest zerowa wtedy i tylko wtedy, gdy zmienne X i Y są niezależne, czyli $p(x, y) = p(x)p(y)$ i wtedy:

$$\log \frac{p(x, y)}{p(x)p(y)} = \log \frac{p(x)p(y)}{p(x)p(y)} = \log 1 = 0.$$

INFORMACJA WZAJEMNA (2)

Powiązanie z innymi funkcjami:

$$\begin{aligned} I(X, Y) &= H(X) - H(X | Y) = \\ &= H(Y) - H(Y | X) = \\ &= H(X) + H(Y) - H(X, Y) \end{aligned}$$

Indeks Giniego:

$$I_G(X) = \sum_i p_i (1 - p_i) = 1 - \sum_i p_i^2$$

SELEKCJA ZMIENNYCH

Celem *selekcji* zmiennych jest identyfikacja zmiennych, których eliminacja poprawia właściwości modelu.

Nadmierna liczba predyktorów jest niekorzystna z uwagi na:

- ryzyko nadmiernej współliniowości predyktorów i związanych z tym problemów;
- wprowadzenie do modelu niepotrzebnej informacji (szumu);
- koszt przygotowania i pozyskania obserwacji rozbudowanych o nadmiarowe predyktory;
- trudności w interpretacji najbardziej znaczącego wpływu predyktorów na zmienną objaśnianą.

TECHNIKI SELEKCJI ZMIENNYCH

- **wprowadzania** – wszystkie zmienne w określonym bloku są jednocześnie wprowadzane do modelu,
- **usuwania** – wszystkie zmienne w określonym bloku są jednocześnie usuwane z modelu,
- **eliminacji wstecznej** – po wprowadzeniu wszystkich zmiennych usuwana jest zmienna spełniająca kryteria usunięcia, aż do wyczerpania się zmiennych spełniających kryteria;
- **selekcji postępującej** – wprowadzanie do modelu kolejno zmiennych spełniających kryteria wprowadzenia, zaczynając od zmiennej, która w najwyższym stopniu spełnia przyjęte kryterium, aż do wyczerpania się zmiennych spełniających kryteria.

EKSTRAKCJA

Ekstrakcja to budowanie nowych zmiennych poprzez liniową lub nieliniową kombinację zmiennych oryginalnych. W odróżnieniu od selekcji, gdzie celem jest zawsze uzyskanie pewnego podzbioru wszystkich zmiennych, wymiarowość przestrzeni będącej wynikiem ekstrakcji może być mniejsza, taka sama, a nawet większa niż wymiar przestrzeni startowej.

ANALIZA SKŁADOWYCH GŁÓWNYCH

Początki techniki analizy składowych głównych pochodzą od Pearsona (1901r.). Jednak główny rozwój tej metody zawdzięcza się pracom amerykańskiego statystyka Hotellinga (1933), który wykorzystał ją do analizy testów osiągnięć szkolnych.

\mathbf{X} wymiarowości $n \times m$ oraz macierz transformacji \mathbf{Q} wymiarowości $m \times m$:

$\mathbf{Y} = \mathbf{XQ}$ wymiarowości $n \times m$.

Dla wektora \mathbf{x} mnożenie to ma postać $\mathbf{y} = \mathbf{Qx}$.

W metodzie PCA należy wybrać macierz \mathbf{Q} .

Oprócz tego jest to nauczanie bez nauczyciela, ponieważ mamy \mathbf{X} i nie mamy zmiennej zależnej.

GEOMETRYCZNY PUNKT WIDZENIA (1)

Z geometrycznego punktu widzenia ideą analizy składowych głównych jest opisanie zmienności układu n punktów w m wymiarowej przestrzeni zmiennych poprzez wprowadzenie nowego układu liniowych, ortogonalnych współrzędnych.

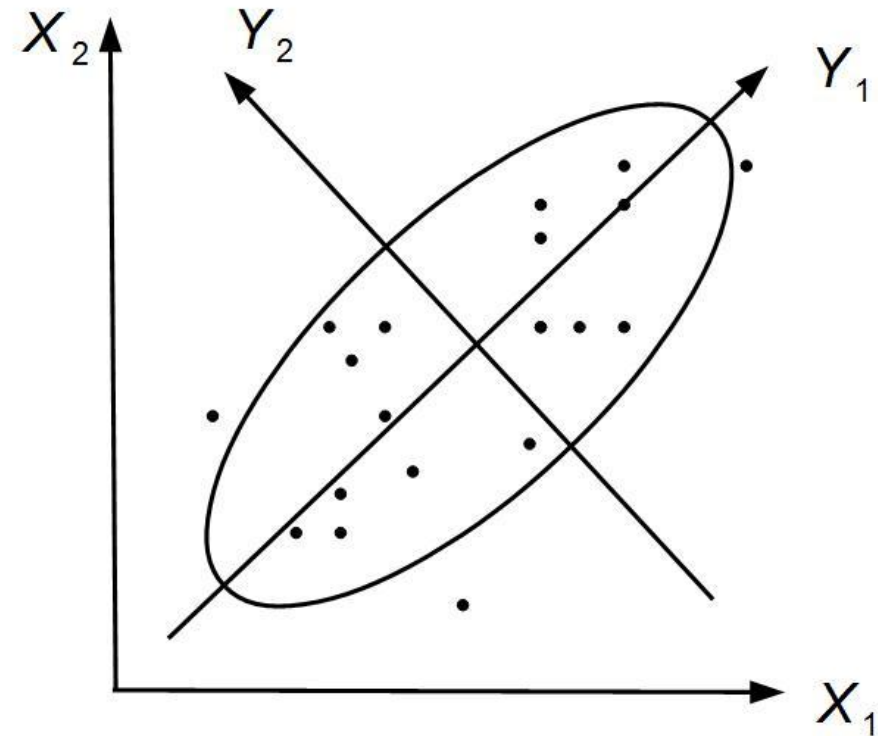
Wariancje danych punktów względem wprowadzonych współrzędnych są uporządkowane malejąco.

Rzuty punktów na pierwszą składową mają największą wariancję ze wszystkich możliwych liniowych współrzędnych.

GEOMETRYCZNY PUNKT WIDZENIA (2)

Kierunek, zgodnie z którym dane są bardziej rozproszone wyznacza nowa oś, która reprezentuje pierwszą główną składową Y_1 . Druga oś, biegnąca pod kątem 90 stopni od pierwszej, wyznacza kierunek drugiej składowej Y_2 . Obie osi współrzędnych X_1, X_2 są transformowane poprzez przesunięcie środka

do punktu średnich \bar{x}_1, \bar{x}_2 , a następnie obrócone w taki sposób, że otrzymaliśmy współrzędne składowych głównych Y_1, Y_2 .



UKRYTE ZMIENNE

Jeszcze jedno podejście polega na rozpatrywaniu przekształconej zmiennej Y jako **szeregu ukrytych zmiennych**, które nie korelują między sobą, czyli są niezależne i leżące w podstawie danych X . Ponieważ X to są dane, to ogólnie one mogą korelować między sobą.

Przy wykorzystaniu metody PCA zakładają, że dane X są liniową kombinacją zmiennych ukrytych Y .

Ponieważ liniowość rozpowszechnia się w obie strony, to nie tylko zmienne ukryte Y można rozpatrywać jako liniową kombinację zmiennych X , a i dane są liniową kombinacją zmiennych ukrytych

$$X = YQ^{-1}.$$

METODA PCA

- automatycznie tworzone są nowe „wymiary” (zmienne) będące kombinacjami istniejących wymiarów; zmienne te nazywa się **składowymi**;
- pozostawia się tylko zmienne, które mają największą zmienność, czyli niosą najwięcej informacji, czyli **składowe główne**.

Zastosowane analizy składowych głównych:

- wizualizacja danych wielowymiarowych;
- dekorrelacja;
- redukcja liczby zmiennych;
- redukcja wymiarowości;
- kompresja sygnału przy jego przetwarzaniu;
- analiza czynnikowa.

PODSTAWOWA IDEA PCA

Podstawową ideą metody jest transformacja wyjściowego zbioru zmiennych X_1, X_2, \dots, X_m na nowy zbiór zmiennych Y_1, Y_2, \dots, Y_m , zwanych składowymi głównymi.

Model matematyczny w analizie składowych głównych jest sformułowany w postaci następującego układu równań liniowych:

$$X_1 = a_{11}Y_1 + a_{12}Y_2 + \dots a_{1m}Y_m$$

$$X_2 = a_{21}Y_1 + a_{22}Y_2 + \dots a_{2m}Y_m$$

...

$$X_m = a_{m1}Y_1 + a_{m2}Y_2 + \dots a_{mm}Y_m$$

Współczynniki a_{ij} określają **wagę danej składowej** w opisie zmiennych empirycznych.

PCA: KROK 1

Założenia:

- **normalność rozkładu** – założenie to nie jest konieczne, gdy analizuje się duży zbiór danych;
- **liczebność i reprezentatywność próby** – do analizy przystępuje się, gdy próba liczy co najmniej 50 obserwacji; próbę należy pobrać w sposób losowy; zbiór obserwacji musi być jednorodny;
- **punkty odstające** – punkty odstające często zniekształcają prawdziwe zależności między zmiennymi; dobrze jest na początku analizy wykryć takie punkty i usunąć je z danych;
- **braki danych** – w przypadku brakujących danych w analizowanej próbie należy zastąpić braki przez średnie lub usunąć przypadki z brakującymi danymi.

PCA: KROK 2

Jeżeli analizowane zmienne są porównywalne (wyrażają się w tych samych jednostkach i są tego samego rzędu), to w dalszej analizie wykorzystuje się macierz kowariancji.

Jeżeli natomiast zmienne mają różne jednostki lub są różnego rzędu, analizę składowych głównych przeprowadza się wykorzystując macierz korelacji.

Użycie macierzy korelacji odpowiada wstępnej normalizacji (standaryzacji) zbioru wejściowego tak, aby każda zmienna miała na wejściu identyczną wariancję.

PCA:KROK 3 (1)

Wyznaczenie składowych głównych.

Szukamy kierunku (wersora \mathbf{a}), dla którego rozkład rzutu danych $(\mathbf{X} - \mu)\mathbf{a}$ ma największą wariancję s_a :

$$s_a = ((\mathbf{X} - \mu)\mathbf{a})^T ((\mathbf{X} - \mu)\mathbf{a}) = \mathbf{a}^T (\mathbf{X} - \mu)^T (\mathbf{X} - \mu)\mathbf{a} = \mathbf{a}^T \mathbf{C} \mathbf{a},$$

gdzie:

$\mu = (\mu_1 \ \mu_2 \ \dots \ \mu_m)$ - wektor wartości średnich;

$(\mathbf{X} - \mu)$ - macierz danych, każdy wiersz której jest równy:

$$(\mathbf{x}_i - \mu) = \mathbf{X}(i,:) - \mu;$$

$\mathbf{C} = \text{cov}(\mathbf{X})$ - macierz kowariancji, $\dim \mathbf{C} = m \times m$;

\mathbf{a} - wersor osi, tzn. $\|\mathbf{a}\| = \sqrt{\mathbf{a}^T \mathbf{a}} = 1$.

PCA: KROK 3 (2)

$$\mathbf{a}^T \mathbf{C} \mathbf{a} \rightarrow \max$$

przy ograniczeniach :

$$\mathbf{a}^T \mathbf{a} = 1.$$

Problem optymalizacji bez ograniczeń:

$$L(\mathbf{a}, \lambda) = \mathbf{a}^T \mathbf{C} \mathbf{a} - \lambda (\mathbf{a}^T \mathbf{a} - 1).$$

$$\frac{\partial L(\mathbf{a}, \lambda)}{\partial \mathbf{a}} = 2\mathbf{C}\mathbf{a} - 2\lambda\mathbf{a} = 0.$$

$$(\mathbf{C} - \lambda \mathbf{I}) \mathbf{a} = 0.$$

Wartości własne $(\lambda_1, \dots, \lambda_m)$ (wraz z odpowiednimi wektorami) są uporządkowane malejąco; m odpowiadających im wektorów własnych $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_m$ wyznaczają nowy układ współrzędnych w przestrzeni \mathbb{R}^m .

PCA: KROK 4

Redukcja wymiaru – kryteria wyboru.

➤ **Kryterium wystarczającej proporcji** – stopień wyjaśnionej wariancji oryginalnych zmiennych musi wynosić co najmniej 75%. W praktyce najczęściej już przy 2-3 głównych składowych stopień wyjaśnienia wariancji jest wystarczający.

➤ **Kryterium Kaisera** – eliminacja składowych głównych, których wartości własne są mniejsze od 1.

➤ **Wykres osypiska** – wyznaczenie na wykresie liniowym kolejnych wartości własnych. Interpretacja polega na znalezieniu miejsca, od którego na prawo występuje łagodny spadek wartości własnych. Nie powinno się uwzględniać więcej czynników, niż te znajdujące się po lewej stronie tego punktu.

Wybór odpowiedniego kryterium leży w gestii statystyka, dlatego też decyzja ta jest dosyć subiektywna i wpływa na rezultaty analizy.

PCA: KROK 5 (1)

Ładunki czynnikowe są współczynnikami korelacji pomiędzy daną zmienną a składowymi.

W przypadku, gdy analiza jest przeprowadzana na podstawie macierzy kowariancji, to współczynnik korelacji między i -tą zmienną X_i i j -tą składową Y_j oblicza się ze wzoru:

$$r_{ij} = \frac{\text{cov}(X_i, Y_j)}{s_i \sqrt{\lambda_j}} = \frac{\lambda_j a_{ij}}{s_i \sqrt{\lambda_j}} = \frac{\sqrt{\lambda_j} a_{ij}}{s_i}.$$

PCA: KROK 5 (2)

Suma wszystkich wartości własnych macierzy kowariancji $\lambda_1 + \lambda_2 + \dots + \lambda_m$ jest całkowita wariancja układu, dzięki czemu można zdefiniować część całkowitej wariancji wyznaczaną przez j -tą składową:

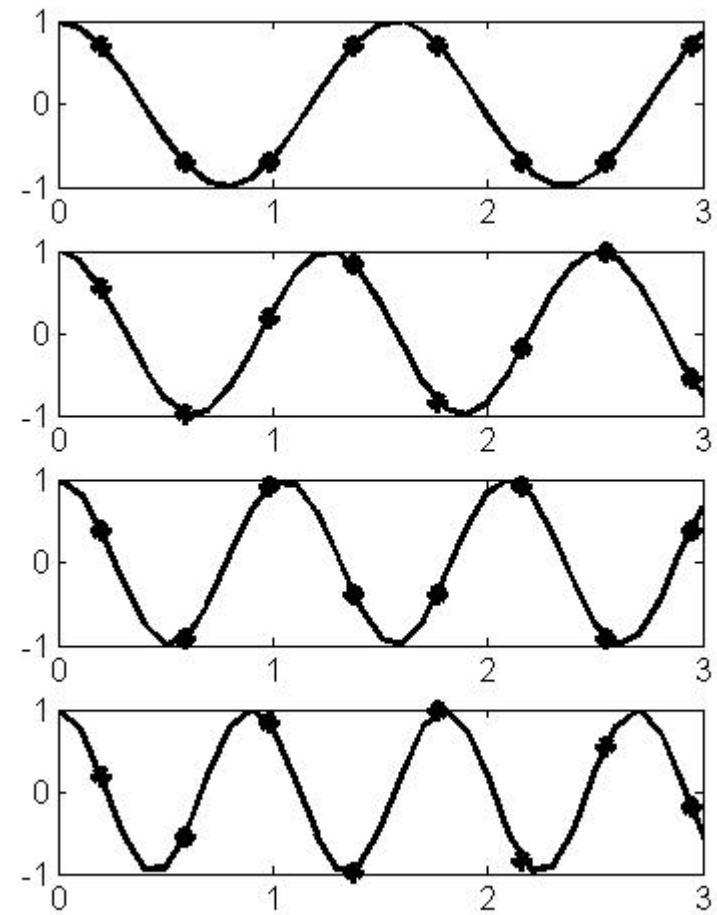
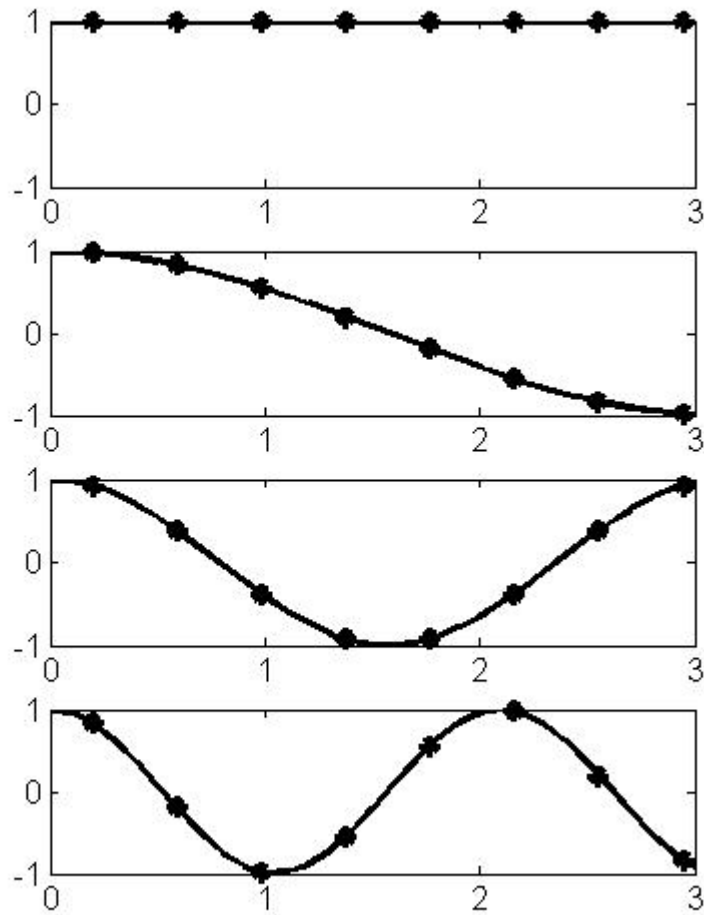
$$h_j = \frac{\lambda_j}{\lambda_1 + \lambda_2 + \dots + \lambda_m} \cdot 100\%.$$

Natomiast procentowy udział zmienności całkowitej wyjaśnionej przez k pierwszych składowych oblicza się następująco:

$$H_k = \sum_{j=1}^k h_j.$$

DCT

8 fal sinusowych $w(f) = \cos(f\varphi)$, $0 \leq \varphi \leq \pi$, $f = 0, 1, \dots, 7$



BAZA PRZEKSZTAŁCENIA COS

Na każdym wykresie zaznaczono osiem wartości funkcji $w(f)$ dla $\varphi = \left\{ \frac{\pi}{16}, \frac{3\pi}{16}, \frac{5\pi}{16}, \frac{7\pi}{16}, \frac{9\pi}{16}, \frac{11\pi}{16}, \frac{13\pi}{16}, \frac{15\pi}{16} \right\}$, które tworzą wektory bazowe \mathbf{w}_f (razem 64 liczby). Umieścimy ich w macierzy \mathbf{W} :

$$\mathbf{W} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0,981 & 0,831 & 0,556 & 0,195 & -0,195 & -0,556 & -0,831 & -0,981 \\ 0,924 & 0,383 & -0,383 & -0,924 & 0,924 & -0,383 & 0,383 & 0,924 \\ 0,831 & -0,195 & -0,981 & -0,556 & 0,556 & 0,981 & 0,195 & -0,831 \\ 0,707 & -0,707 & -0,707 & 0,707 & 0,707 & -0,707 & -0,707 & 0,707 \\ 0,556 & -0,981 & 0,195 & 0,831 & -0,831 & -0,195 & 0,981 & -0,556 \\ 0,383 & -0,924 & 0,924 & -0,386 & -0,383 & -0,924 & -0,924 & 0,383 \\ 0,195 & -0,556 & 0,831 & -0,981 & 0,981 & -0,831 & 0,556 & -0,195 \end{pmatrix}$$

JEDNOWYMIAROWE DCT

Np., niech wektor 8 wartości skorelowanych:

$$\mathbf{x} = (0,6 \ 0,5 \ 0,4 \ 0,5 \ 0,6 \ 0,5 \ 0,4 \ 0,55).$$

Zapiszmy ten wektor za pomocą sumy 8 wektorów bazowych \mathbf{w}_f

$$\mathbf{x} = \sum_f a_f \mathbf{w}_f.$$

Rozwiązując układ 8 równań liniowych szukamy wartości a_f :

$$a_0 = 0,506 \quad a_1 = 0,0143 \quad a_2 = 0,0115 \quad a_3 = 0,0439$$

$$a_4 = 0,0795 \quad a_5 = -0,0432 \quad a_6 = 0,0478 \quad a_7 = -0,0077$$

Można zauważyć, że wektor $a_0 = 0,506$ mało różni się od elementów wektora \mathbf{x} , a pozostałe wartości wag są znacznie mniejsze. Przykład ten pokazuje w jaki sposób dowolne przekształcenie ortogonalne wykonuje kompresję danych.

KWANTYZACJA

Dodatkowa kwantyzacja wag a_f zwiększa kompresję przy małej stracie danych i znacznie zmniejsza ilość danych, które należy przechowywać.

W praktyce współczynniki DCT łatwiej obliczać za pomocą wzoru:

$$G_f = \frac{1}{2} C_f \sum_{t=0}^7 x_t \cos\left(\frac{(2t+1)f\pi}{16}\right),$$

$$C_f = \begin{cases} \frac{1}{\sqrt{2}}, & f = 0 \\ 1, & f = 1, 2, \dots, 7 \end{cases}$$

Wzór ten jest łatwy, ale obliczenia wg tego wzoru są wolne, więc częściej stosuje się szybka transformata Fouriera.

IDCT

W celu odzyskania wektora \mathbf{x} , dla ósemek współczynników DCT wykonuje się odwrotna DCT (ang. inverse, IDCT):

$$x_t = \frac{1}{2} \sum_{j=0}^7 C_j G_j \cos\left(\frac{(2t+1)j\pi}{16}\right).$$

PRZYKŁAD (1)

$$\mathbf{x} = (12 \ 10 \ 8 \ 10 \ 12 \ 10 \ 8 \ 11)$$

Wykonujemy DCT i otrzymamy 8 współczynników:

$$\begin{array}{cccc} 28,6375 & 0,571202 & 0,46194 & 1,757 \\ 3,18198 & -1,72956 & 0,191342 & -0,308709 \end{array}$$

Zaokrąglimy te wartości (kwantyzacja):

$$28,6 \quad 0,6 \quad 0,5 \quad 1,8 \quad 3,2 \quad -1,7 \quad 0,2 \quad -0,3$$

Wykonamy IDCT:

$$\begin{array}{cccc} 12,0254 & 10,0233 & 7,96054 & 9,93097 \\ 12,0164 & 9,99321 & 7,94354 & 10,9989 \end{array}$$

PRZYKŁAD (2)

Kwantyzację współczynników

28 1 1 2 3 -2 0 0

IDCT:

12,1883 10,2315 7,74931 9,20863

11,7876 9,54549 7,82865 10,6557

W końcu jeszcze jedna kwantyzacja:

28 0 0 2 3 -2 0 0

IDCT:

11,236 9,62443 7,66286 9,54302

12,3471 10,0146 8,0534 10,6842

PRZYKŁAD (3)

W wektorze tym największa różnica między wartością początkową (12) a odzyskaną (11,236) równa jest 0,764 co stanowi 6,4% od 12.

Zbiór kwantowanych współczynników
28 0 0 2 3 -2 0 0 ma 4 własności, które pozwalają
wykorzystywać go do kompresji danych (przy czym dekompresja
ma małą stratę danych):

- zbiór ten zawiera tylko liczby całkowite;
- tylko 4 z nich nie są równe 0;
- współczynniki zerowe tworzą ciągi;
- wśród niezerowych współczynników tylko pierwszy ma dużą wartość, pozostałe są znacznie mniejsze od wartości danych.

DWUWYMIAROWE DCT (1)

$$G_{uv} = \frac{1}{\sqrt{2n}} C_u C_v \underbrace{\sum_{i=0}^{n-1} \sum_{j=0}^{n-1} x_{ij} \cos\left(\frac{(2i+1)u\pi}{16}\right) \cos\left(\frac{(2j+1)v\pi}{16}\right)}_{p_{ij}}, \quad (1)$$

$$C_u, C_v = \begin{cases} \frac{1}{\sqrt{2}}, & f = 0 \\ 1, & f = 1, 2, \dots, 7 \end{cases}$$

Jeśli to jest obraz, to on dzieli się na bloki pikseli x_{ij} $n \times n$ i dla każdego bloku wykonuje się obliczenie współczynników DCT.

Dekoder odzyskuje obraz:

$$x_{ij} = \frac{1}{4} \sum_{u=0}^{n-1} \sum_{v=0}^{n-1} C_u C_v G_{uv} \cos\left(\frac{(2i+1)u\pi}{16}\right) \cos\left(\frac{(2j+1)v\pi}{16}\right).$$

DWUWYMIAROWE DCT (2)

Dwuwymiarowe DCT można rozpatrywać jako podwójne obracanie w przestrzeni n -wymiarowej.

$$\begin{pmatrix} L & L & L & L & L & L & L & L \\ L & L & L & L & L & L & L & L \\ L & L & L & L & L & L & L & L \\ L & L & L & L & L & L & L & L \\ L & L & L & L & L & L & L & L \\ L & L & L & L & L & L & L & L \\ L & L & L & L & L & L & L & L \\ L & L & L & L & L & L & L & L \end{pmatrix} \Rightarrow \begin{pmatrix} L & S & S & S & S & S & S & S \\ L & S & S & S & S & S & S & S \\ L & S & S & S & S & S & S & S \\ L & S & S & S & S & S & S & S \\ L & S & S & S & S & S & S & S \\ L & S & S & S & S & S & S & S \\ L & S & S & S & S & S & S & S \\ L & S & S & S & S & S & S & S \end{pmatrix} \Rightarrow \begin{pmatrix} L & S & S & S & S & S & S & S \\ S & s & s & s & s & s & s & s \\ S & s & s & s & s & s & s & s \\ S & s & s & s & s & s & s & s \\ S & s & s & s & s & s & s & s \\ S & s & s & s & s & s & s & s \\ S & s & s & s & s & s & s & s \\ S & s & s & s & s & s & s & s \end{pmatrix}$$

Metody przyspieszania DCT:

1. Wykonanie działań arytmetycznych na liczbach stałoprzecinkowych zamiast zmiennoprzecinkowych, co wykonuje się znacznie szybciej.

DWUWYMIAROWE DCT (3)

1. w niezależności od rozmiaru danych (obrazów) wykonuje się obliczenie tylko 32 wartości cosinusów;
2. Podwójną sumę w (1) można zapisać w postaci:

$$c_{uv} = \begin{cases} \frac{1}{\sqrt{8}}, & u = 0 \\ \frac{1}{2} \cos\left(\frac{(2v+1)u\pi}{16}\right), & u > 0. \end{cases}$$

Można pokazać, że postać taka pozwala na zmniejszenie liczby działań $q = N/n$ razy, gdzie $N \times N$ - wymiar całego obrazu. Niestety q nie może przyjmować duże wartości.

MNW

Metoda największej wiarygodności (*ang.* maximum likelihood estimation) należy do jednych z najczęściej stosowanych metod punktowej estymacji parametrów. Polega ona na założeniu, że cała informacja o próbie statystycznej zawarta jest w funkcji wiarygodności.

MNW została przeanalizowana, zaproponowana i rozpowszechniona przez R.A.Fishera w latach 1912-1922 (choć wcześniej była ona wykorzystana przez Gaussa, Laplace'a i in.).

FUNKCJA WIARYGODNOŚCI

$\mathbf{w} = \{w_1, w_2, \dots, w_K\}$, $k = 1, 2, \dots, K$ interesujące nieznanne parametry.
 $\mathbf{x} = (x_1, x_2, \dots, x_n)^T$ - próba reprezentująca zmienną losową X ,
rozkład której określony przez funkcję $p(\mathbf{x}, \mathbf{w})$.

Funkcja $p(\mathbf{x}, \mathbf{w})$ jest prawdopodobieństwem w przypadku zmiennej dyskretnej lub gęstością prawdopodobieństwa w przypadku zmiennej ciągłej.

Niech pojedyncze i -te doświadczenie, czyli pobranie próby o liczebności 1, daje w wyniku x_i , wtedy dla próby n -wymiarowej (liczby doświadczeń n , serii n wyników) **funkcja wiarygodności** jest określona jako iloczyn prawdopodobieństw:

$$L(\mathbf{x}, \mathbf{w}) = \prod_{i=1}^n p(x_i, \mathbf{w})$$

LOGARYTMICZNA FUNKCJA WIARYGODNOŚCI

Funkcja $L(\mathbf{x}, \mathbf{w})$ osiąga maksimum w tych samych punktach, co jej logarytm i w praktyce często wykorzystuje się ***logarytmiczna funkcja wiarygodności***:

$$l(\mathbf{x}, \mathbf{w}) = \ln(L(\mathbf{x}, \mathbf{w})) = \ln \left(\prod_{i=1}^n p(x_i, \mathbf{w}) \right) = \sum_{i=1}^n \ln(p(x_i, \mathbf{w}))$$

Estymatorem maksymalnej wiarygodności nazywa się zbiór parametrów \mathbf{w}^* , dla którego logarytmiczna funkcja wiarygodności osiąga maksymalną wartość:

$$\mathbf{w}^* = \arg \max_{\mathbf{w}} l(\mathbf{x}, \mathbf{w}),$$

$$\frac{\partial l}{\partial w_k} = 0.$$

Należy rozwiązać układ K równań według parametrów \mathbf{w} .

PRZYKŁAD.

Wykorzystując metodę największej wiarygodności, ocenić parametry μ i σ rozkładu normalnego, jeśli w rezultacie n niezależnych doświadczeń zmienna losowa X przyjęła wartości X_1, X_2, \dots, X_n .

HISTOGRAM (1)

Histogram to zestawienie danych statystycznych w postaci wykresu powierzchniowego złożonego z przylegających do siebie słupków (prostokątów), których wysokość ilustruje liczebność występowania badanej cechy w populacji lub jej próbie, a podstawy (które znajdują się na osi odciętych) są rozpiętościami przedziałów klasowych.

Przedziały (niech x_0 - początek współrzędnych i h - szerokość przedziałów):

$$[x_0 + rh, x_0 + (r + 1)h),$$

gdzie r - pewne dodatnie i ujemne liczby całkowite.

HISTOGRAM (2)

Histogram:

$$\begin{aligned}\hat{f}(x) &= \frac{1}{n} \frac{\text{liczba } X_i \text{ w jednym przedziale z } x}{\text{szerokość przedziału, zawierającego } x} = \\ &= \frac{1}{nh} \sum_{i=1}^n I[X_i \text{ w jednym przedziale z } x]\end{aligned}$$

gdzie $I[A]$ - funkcja, która przyjmuje wartość 1, gdy A – prawdziwe oraz 0 w przeciwnym przypadku.

Szacowany parametr $\hat{f}(x)$ zależy od tych dwóch wartości: od **początku współrzędnych** oraz **szerokość przedziałów**, które badać musi wybrać sam. Najczęściej parametry te określane są za pomocą pewnych metod heurystycznych.

ESTYMACJA JĄDROWA

Metoda histogramów ma szereg wad:

- zły dobór wielkości przedziałów powoduje złe odwzorowanie funkcji gęstości;
- histogram ma ograniczone możliwości zastosowania dla danych dużej wymiarowości;
- przybliżona za pomocą histogramu funkcja gęstości jest nieciągła.

Z wadami tymi dobrze radzi ***estymacja jądrowa***, która została zapoczątkowana przez Rosenblatta (1956) i Parzena (1962) pół wieku temu. Pozwala ona na bezpośrednią estymację funkcji rozkładu prawdopodobieństwa na podstawie zaobserwowanych wartości badanej zmiennej losowej (próby losowej).

FUNKCJA JĄDROWA

Konstrukcja **estymatora jądrowego** $\hat{f}(x)$ (*ang.* kernel estimator) funkcji gęstości $f(x)$ zmiennej losowej X polega na przypisaniu każdemu elementowi próby x_i pewnej funkcji $K(x_i, x)$ zwanej **funkcją jądrową** w sposób następujący:

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^n K(x_i, x).$$

Funkcją jądrową $K(x)$ może być każda nieujemna funkcja:

$$K(x) \geq 0,$$

przyjmująca wyłącznie skończone wartości oraz taka, że:

$$\int_{-\infty}^{\infty} K(x) dx = 1.$$

WSPÓŁCZYNNIK GŁADKOŚCI

$$K(x) = -K(x),$$

$$K(x_i, x) = \frac{1}{h} K\left(\frac{x - x_i}{h}\right).$$

Parametr h nazywa się **współczynnikiem gładkości** lub **szerokością okna**.

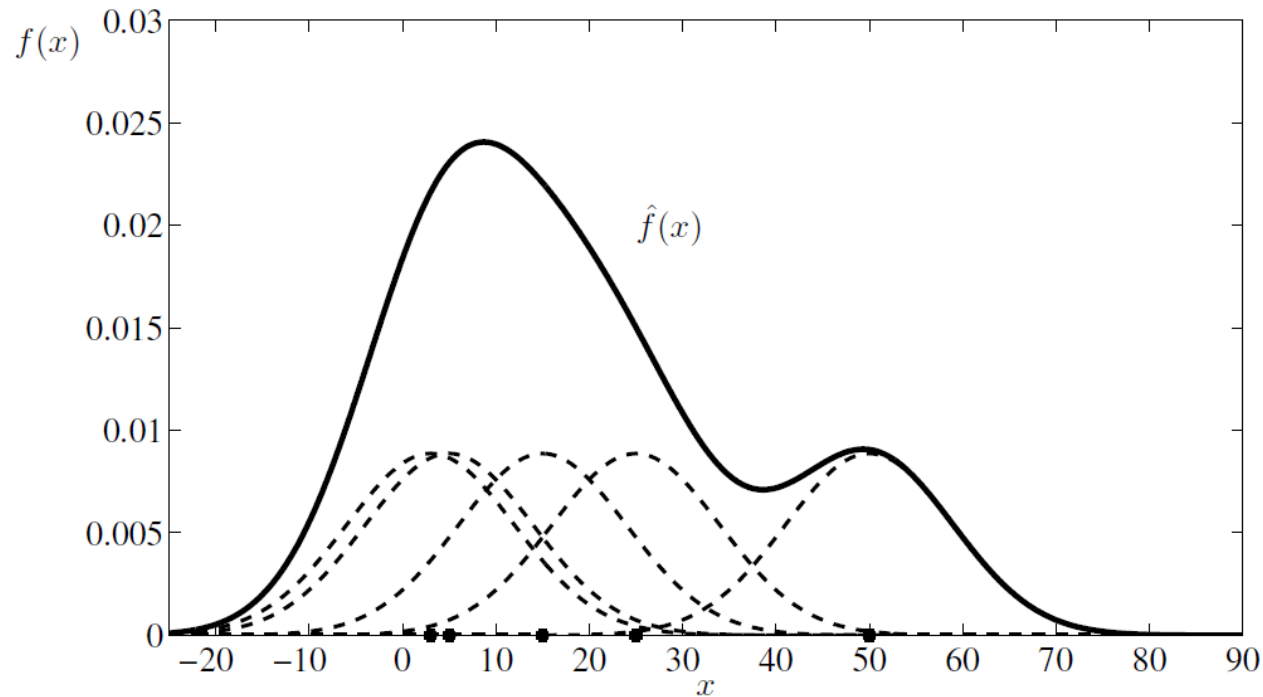
Estymator $\hat{f}(x)$:

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - x_i}{h}\right).$$

Końcowy estymator funkcji gęstości $\hat{f}(x)$ dziedziczy własności funkcji-jądra.

PRZYKŁADOWA POSTAĆ ESTYMATORA

Funkcja jądrowa Gaussa, $x = (3 \ 5 \ 15 \ 25 \ 50)$, $h = 9$.



Estymator jądrowy umożliwia oszacowanie funkcji gęstości praktycznie dowolnego rozkładu, bez żadnych założeń o jego przynależności do ustalonej klasy.

MIARY ROZBIEŻNOŚCI

Błąd kwadratowy MSE:

$$MSE_x(\hat{f}) = E\left[\left(\hat{f}(x) - f(x)\right)^2\right].$$

Średni scałkowany błąd kwadratowy MISE, (*ang.* Mean Integrated Square Error):

$$MISE_x(\hat{f}) = E\left[\int_{-\infty}^{\infty} \left(\hat{f}(x) - f(x)\right)^2 dx\right].$$

Miary te są nieujemne i tym mniejsze im bliżej rzeczywistej funkcji gęstości $f(x)$ znajduje się estymator $\hat{f}(x)$.

WYBÓR POSTACI JĄDRA

W przypadku jednowymiarowym jako funkcję $K(x)$ przyjmuje się klasyczne postacie gęstości rozkładów probabilistycznych.

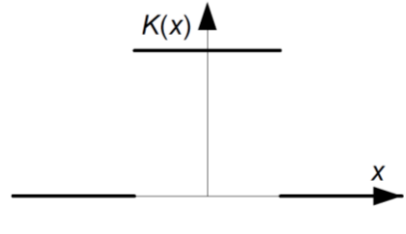
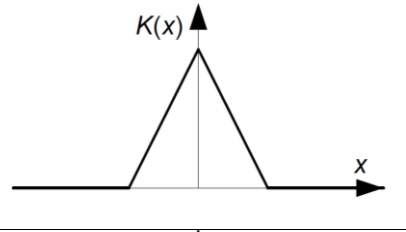
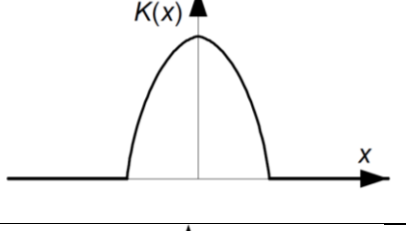
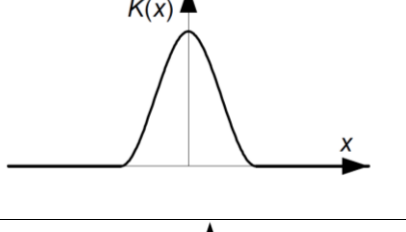
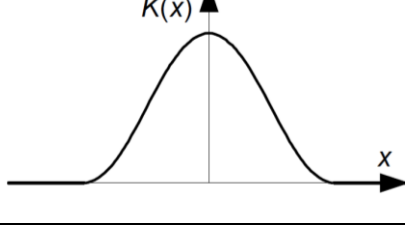
Dla funkcji Gaussa jądro:

$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

daje estymator jądrowy:

$$\hat{f}(x) = \frac{1}{nh\sqrt{2\pi}} \sum_{i=1}^n e^{-\frac{(x-x_i)^2}{2h}}$$

JĄDRA

jednostajne	$K(x) = \begin{cases} 1/2, & x \leq 1 \\ 0, & x > 1 \end{cases}$	
trójkątne	$K(x) = \begin{cases} 1 - x , & x \leq 1 \\ 0, & x > 1 \end{cases}$	
Epanecznikova	$K(x) = \begin{cases} 3/4(1 - x^2), & x \leq 1 \\ 0, & x > 1 \end{cases}$	
kwadratowe	$K(x) = \begin{cases} 15/6(1 - x^2), & x \leq 1 \\ 0, & x > 1 \end{cases}$	
normalne	$K(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$	

PRZYPADEK m-WYMIAROWY

$$\hat{f}(\mathbf{x}) = \frac{1}{nh^m} \sum_{i=1}^n K\left(\frac{\mathbf{x} - \mathbf{x}_i}{h}\right).$$

Jądro radialne:

$$K(\mathbf{x}) = CK\left(\sqrt{\mathbf{x}^T \mathbf{x}}\right) = CK\left(\left[\begin{matrix} x_1 & x_2 & \dots & x_m \end{matrix}\right]^T \left[\begin{matrix} x_1 & x_2 & \dots & x_m \end{matrix}\right]\right)^{\frac{1}{2}} = CK(\|\mathbf{x}\|)$$

Jądro produktowe:

$$K(\mathbf{x}) = K(\mathbf{x}) = K(x_1)K(x_2)\dots K(x_m),$$

gdzie K oznacza omówione uprzednie jądro jednowymiarowe, natomiast C jest dodatnią stałą, wyznaczoną tak, aby spełniony został warunek:

$$\int_{R^m} K(\mathbf{x}) d\mathbf{x} = 1.$$

JĄDRO RADIALNE A JĄDRO PRODUKTOWE

Dla dowolnie ustalonego jądra jednowymiarowego K bardziej efektywne jest jądro radialne niż produktowe, jednak z punktu widzenia zastosowania w aplikacjach różnica jest nieznaczna. Fakt ten powoduje, że w praktycznych zastosowaniach w przypadkach wielowymiarowych zalecają zastosowanie jądra produktowego, które wiąże się z łatwością wyznaczania współczynnika wygładzania. Także jest ono znacznie dogodniejsze w dalszej analizie, ponieważ procedury całkowania i różniczkowania jądra produktowego niewiele różnią się od przypadku jednowymiarowego.

WSPÓŁCZYNNIK GŁADKOŚCI

Wybór optymalnego współczynnika gładkości opiera się na minimalizacji miar rozbieżności między funkcją $f(x)$ a estymatorem $\hat{f}(x)$, czyli minimalizacji MSE, MISE i in.

$$h = \left[\frac{\int K^2(z) dz}{\left(\int z^2 K(z) dz \right)^2 \int [f''(z)]^2 dz} \right]^{\frac{1}{5}} n^{-\frac{1}{5}},$$

gdzie $z = \frac{x - x_i}{h}$.

METODA PRZYBLIŻONA

Najczęściej stosuje się **metoda przybliżona** estymacji jądrowej z funkcją Gaussa, dla której współczynnik gładkości w przypadku jednowymiarowym obliczany jest za pomocą tzw. Silverman-reguły:

$$h = 1,06 \min \left\{ s, \frac{IQR}{1,34} \right\} n^{-\frac{1}{5}},$$

gdzie s -odchylenie standardowe w próbie, a IQR -rozstęp międzykwartylowy w próbie.

W przypadku wielowymiarowym:

$$h = \sqrt[m+4]{\frac{4}{m^2(m+2)n} \prod_{j=1}^m s_j},$$

gdzie s_j odchylenia standardowe poszczególnych zmiennych.

METODA PODSTAWIEŃ

Oprócz metody przybliżonej także jest wykorzystywana **metoda podstawień** (*ang.* plug-in), w której do wzoru na wartość h podstawiane są oceny nieznaney wartości $\int [f''(z)]^2 dz$ obliczone na podstawie początkowej oceny tego parametru, która z kolei bazuje się na wstępnie obliczonym współczynniku gładkości, przykładowo na $h = 1,06sn^{-1/5}$. Wszystkie pozostałe parametry w tym wzorze po wybraniu jądra są znane.

KROSWALIDACJA (1)

Jednak jednym z najbardziej rozpowszechnionych podejść jest wykorzystywanie walidacji krzyżowej, polegającej na minimalizacji MISE. Metody polega na wyznaczaniu wartości h minimalizującej funkcję g :

$$g(x) = \frac{1}{n^2 h^m} \sum_{i=1}^n \sum_{j=1}^n \tilde{K}\left(\frac{x_j - x_i}{h}\right) + \frac{2}{nh^m} K(0),$$

gdzie:

$$\tilde{K}(x) = K^{*2}(x) - 2K(x),$$

$$K^{*2}(x) = \int_{-\infty}^{\infty} K(t)K(x-t)dt.$$

Minimalizacji funkcji $\tilde{K}(x)$ dokonuje się przy użyciu numerycznych metod optymalizacji.

KROSWALIDACJA (2)

Oprócz minimalizacji MISE wykorzystuje się metoda walidacji krzyżowej **na bazie funkcji wiarygodności**. Współczynnik gładkości h wybierany jest za pomocą maksymalizacji logarytmu funkcji wiarygodności, zbudowanej z całej próby oprócz i -go pomiaru:

$$l = \ln L = \sum_{i=1}^n \ln \hat{f}_{-1}(x),$$

gdzie $\hat{f}_{-1}(x)$ - estymator funkcji gęstości dla całej próby, oprócz i -go pomiaru, czyli:

$$\hat{f}_{-1}(x) = \frac{1}{(n-1)h} \sum_{j=1, j \neq i}^n K\left(\frac{x - x_j}{h}\right).$$

WYMAGANA LICZNOŚĆ PRÓBY

Niezbędna liczność próby n zależy przede wszystkim od liczby m badanych zmiennych losowych X_1, X_2, \dots, X_m . W celu zapewnienia 10-procentowej dokładności dla rozkładu normalnego można w przybliżeniu przyjąć jako $n \cong 4^m$. Sugeruje to minimalność liczebność próby, na przykład, $n = 4$ dla jednowymiarowej zmiennej. W praktyce, jednak, dla jednowymiarowej zmiennej losowej X , wymagana liczność próby n wynosi 20-50, odpowiednio zwiększając się wraz ze wzrostem wymiaru zmiennej. Jednak dzięki współczesnej technice komputerowej, nawet w złożonych zagadnieniach wielowymiarowych i przy niesprzyjających cechach rozkładów, nie musi to obecnie stanowić istotnej przeszkody, zwłaszcza dzięki możliwościom stosowania procedur redukcji wymiaru m i liczności próby n .

PRZYKŁAD

Korzystając ze wzoru:

$$h = \left[\frac{\int K^2(z) dz}{\left(\int z^2 K(z) dz \right)^2 \int [f''(z)]^2 dz} \right]^{\frac{1}{5}} n^{-\frac{1}{5}}$$

udowodnić, że dla rozkładu Gaussa $h = 1,06sn^{-1/5}$, jeśli wiadomo, że:

$$\int [f''(z)]^2 dz = \frac{3}{8\sqrt{\pi}s^5}.$$

MIESZANINA ROZKŁADÓW

Mieszaniną rozkładów nazywa się rozkład, który można opisać za pomocą funkcji:

$$f(x) = \sum_{k=1}^K \alpha_k f(x, w_k),$$

gdzie α_k - współczynniki wagowe (które też nazywa się czasami prawdopodobieństwami a’priori), przy czym:

$$\alpha_k \geq 0,$$

$$\sum_{k=1}^K \alpha_k = 1$$

ZADANIE OPTYMALIZACJI (1)

Dla próby n -elementowej:

$$\sum_{i=1}^n \ln \left(\sum_{k=1}^K \alpha_k f(x_i, w_k) \right) \rightarrow \max.$$

Najbardziej rozpowszechnionym algorytmem, który pozwala rozwiązać dane zadanie jest **metoda EM** (*ang.* Expectation-Maximization), która polega na rozpatrywaniu prawdopodobieństw a'posteriori g_{ik} przynależności x_i -go obiektu (pomiaru) do klasy k :

$$g_{ik} = \frac{\alpha_k f(x_i, w_k)}{\sum_{k=1}^K \alpha_k f(x_i, w_k)} \quad (1)$$

gdzie $g_{ik} \geq 0$, $\sum_{k=1}^K g_{ik} = 1$

ZADANIE OPTYMALIZACJI (1)

$$\sum_{k=1}^K \alpha_k f(x_i, w_k) = \frac{\alpha_k f(x_i, w_k)}{g_{ik}} \quad (2)$$

$$I = \sum_{i=1}^n \ln \left(\sum_{k=1}^K \alpha_k f(x_i, w_k) \right)$$

$$I = \sum_{i=1}^n \ln \left(\frac{\alpha_k f(x_i, w_k)}{g_{ik}} \right) = \sum_{i=1}^n \ln \alpha_k + \sum_{i=1}^n \ln f(x_i, w_k) - \sum_{i=1}^n \ln g_{ik} \quad (3)$$

Ponieważ $\sum_{k=1}^K g_{ik} = 1$, to wzór (3) można zapisać postaci:

$$I = \sum_{k=1}^K \sum_{i=1}^n g_{ik} \ln \alpha_k + \sum_{k=1}^K \sum_{i=1}^n g_{ik} \ln f(x_i, w_k) - \sum_{k=1}^K \sum_{i=1}^n g_{ik} \ln g_{ik}. \quad (4)$$

ALGORYTM ITERACYJNY (1)

Idea iteracyjnego algorytmu estymacji parametrów \mathbf{w} polega na tym, że należy wybrać początkowe przybliżenie tych parametrów \mathbf{w}^0 , obliczyć początkowe wartości prawdopodobieństw a'posteriori g_{ik}^0 zgodnie ze wzorem (1) i wracając do (4) znaleźć kolejne wartości parametrów \mathbf{w}^1 z warunku maksymalizacji w oddzielności dla każdego z pierwszych dwóch składników wzoru (4).

Rozwiązaniem zadania optymalizacji pierwszego składnika:

$$\sum_{k=1}^K \sum_{i=1}^n g_{ik} \ln \alpha_k \rightarrow \max_{\alpha_1, \dots, \alpha_K}$$

przy ograniczeniu $\sum_{k=1}^K \alpha_k = 1$. jest:

$$\alpha_k^{t+1} = \frac{1}{n} \sum_{i=1}^n g_{ik}^t, \text{ gdzie } t = 0, 1, 2, \dots - \text{numer iteracji.}$$

ALGORYTM ITERACYJNY (2)

Rozwiązanie zadania optymalizacji drugiego składnika:

$$\sum_{k=1}^K \sum_{i=1}^n g_{ik} \ln f(x_i, w_k) \rightarrow \max_{\alpha_1, \dots, \alpha_K}$$

znajduje się łatwiej, niż rozwiązanie zadania optymalizacja pierwszego składnika. Wzór na parametry \mathbf{w}^{t+1} zapisuje się z uwzględnieniem znanej postaci funkcji $f(x, \mathbf{w})$

W przypadku jednowymiarowego rozkładu normalnego (tzw. mieszanina rozkładów Gaussa) wzór na $f(x)$ przyjmuje postać:

$$f(x) = \sum_{k=1}^K \alpha_k f(x, \mu_k, \sigma_k)$$

$$\text{gdzie } f(x, \mu_k, \sigma_k) = \frac{1}{\sqrt{2\pi\sigma_k}} e^{-\frac{1}{2}\left(\frac{x-\mu_k}{\sigma_k}\right)^2}.$$

ALGORYTM ITERACYJNY (3)

$$\mu_k^{t+1} = \frac{1}{\sum_{i=1}^n g_{ik}^t} \sum_{i=1}^n g_{ik}^t x_i,$$

$$\sigma_k^{t+1} = \frac{1}{\sum_{i=1}^n g_{ik}^t} \sum_{i=1}^n g_{ik}^t (x_i - \mu_k)^2.$$

Wartości $\alpha_k^{t+1}, \mu_k^{t+1}, \sigma_k^{t+1}$ określają algorytm EM dla mieszaniny rozkładów normalnych.

EM: WNIOSEK (3)

Należy podkreślić, że chociaż najczęściej są badane modele z rozkładem normalnym, właśnie ten model nie spełnia warunków, które zapewniają prawidłową pracę algorytmu EM. Nie jest spełnione ograniczenie logarytmicznej funkcji wiarygodności (czyli przy pewnych warunkach funkcja ta cały czas wzrasta) oraz nie jest spełniony warunek wypukłości funkcji celu, co jest niezbędne do zbieżności algorytmu EM. To wszystko razem z dużą ilością maksimów lokalnych przy liczbie zmiennych $m \geq 2$ czyni, że algorytm staje się bardzo czuły i może doprowadzić nie do "prawidłowych" parametrów rozkładów, lecz do najbardziej prawdopodobnych. Oprócz tego, działania algorytmu mocno zależy od początkowego przybliżenia. Jednak pomimo wskazanych wad, udowodnione zostało, że algorytm EM jest bardziej efektywnym narzędziem do rozwiązywania zadania podziału mieszaniny normalnej, niż inne procedury metod numerycznych.

EM: PRZYPADEK WIELOWYMIAROWY (1)

Niech $\mathbf{X}_{n \times m}$ - próba m zmiennych, w której każdy obiekt $\mathbf{x}_i = \mathbf{X}(i,:)$ musi być przypisany do jednej z klas $1, 2, \dots, K$.

1. Początkowe parametry inicjalizacji algorytmu:

➤ współczynniki wagowe najczęściej są jednakowe i spełniają warunek $\sum_{k=1}^K \alpha_k = 1$, czyli $\alpha_k = \frac{1}{K}$;

➤ początkowe wartości średnich w każdej klasie $\mu_k = (\mu_{1k} \dots \mu_{mk})$ są wybierane w sposób losowy;

➤ początkowe macierze kowariancji \mathbf{C}_k są jednakowe w każdej klasie i są równe macierzy kowariancji danych \mathbf{X} .

EM: PRZYPADEK WIELOWYMIAROWY (2)

2. Dla każdego obiektu \mathbf{x}_i rozpatruje się mieszaninę rozkładów normalnych, w których funkcja gęstości dla każdego obiektu \mathbf{x}_i w klasie k wygląda następująco:

$$f(\mathbf{x}_i, \mu_k, \mathbf{C}_k) = \frac{1}{(2\pi)^{\frac{m}{2}} \sqrt{\det(\mathbf{C}_k)}} e^{-\frac{1}{2}(\mathbf{x}_i - \mu_k)^T \mathbf{C}_k^{-1} (\mathbf{x}_i - \mu_k)}.$$

3. Obliczenie oczekiwanych wartości prawdopodobieństw g_{ik} przynależności obiektu \mathbf{x}_i do klasy k :

$$g_{ik} = \frac{\alpha_k f(\mathbf{x}_i, \mu_k, \mathbf{C}_k)}{\sum_{k=1}^K \alpha_k f(\mathbf{x}_i, \mu_k, \mathbf{C}_k)}.$$

EM: PRZYPADEK WIELOWYMIAROWY (3)

4. Obliczenie logarytmicznej funkcji wiarygodności l :

$$l = \sum_{i=1}^n \ln \left(\sum_{k=1}^K \alpha_k f(\mathbf{x}_i, \mu_k, \mathbf{C}_k) \right).$$

5. Wyznaczenie nowych wartości parametrów modelu:

$$\alpha_k = \frac{1}{n} \sum_{i=1}^n g_{ik},$$

$$\mu_k = \frac{\sum_{i=1}^n g_{ik} \mathbf{x}_i}{n\alpha_k},$$

$$\mathbf{C}_k = \frac{\sum_{i=1}^n (\mathbf{x}_i - \mu_k)^T g_{ik} (\mathbf{x}_i - \mu_k)}{n\alpha_k}.$$

EM: PRZYPADEK WIELOWYMIAROWY (4)

6. Porównanie wartości funkcji wiarygodności z wartością funkcji z poprzedniej iteracji (dla pierwszej iteracji przyjmuje się, że na poprzedniej iteracji $l = 0$).

Jeśli wartość bezwzględna różnicy funkcji wiarygodności jest mniejsza, niż dopuszczalny błąd obliczeń δ , który zadaje wykonawca, czyli $\Delta l = |l^{t+1} - l^t| < \delta$, to algorytm się kończy, inaczej należy wrócić do kroku 2.

Jeszcze jednym dodatkowym ograniczeniem może być podanie maksymalnej liczby iteracji.

PRZYKŁAD (1)

Podzielić dane z tab. na dwie klasy.

X_1	X_2
1	1
6	7
3	2
4	6
5	7
2	1

$$\mathbf{X} = \begin{pmatrix} 1 & 1 \\ 6 & 7 \\ 3 & 2 \\ 4 & 6 \\ 5 & 7 \\ 2 & 1 \end{pmatrix}$$

Błąd dopuszczalny δ wybrać 0,001.

W zadaniu tym należy podzielić $n = 6$ obiektów macierzy danych \mathbf{X} na $K = 2$ klasy; liczba zmiennych $m = 2$.

PRZYKŁAD (2)

1. Początkowe parametry inicjalizacji algorytmu:

➤ współczynniki wagowe: $\alpha_1 = \frac{1}{2}$, $\alpha_2 = \frac{1}{2}$;

➤ początkowe wartości średnich w każdej klasie $\mu_k = (\mu_{1k} \dots \mu_{mk})$ są wybierane w sposób losowy, macierz μ musi być wymiarowości $K \times m$, czyli 2×2 :

$$\mu = \begin{pmatrix} 3,5 & 3,8 \\ 3,5 & 4 \end{pmatrix}$$

PRZYKŁAD (3)

➤ początkowe macierze kowariancji \mathbf{C}_k są jednakowe w każdej klasie i są równe macierzy kowariancji danych \mathbf{X} :

$$\mathbf{C}_1 = \mathbf{C}_2 = \frac{1}{6} \begin{pmatrix} 1-3,5 & 1-4 \\ 6-3,5 & 7-4 \\ 3-3,5 & 2-4 \\ 4-3,5 & 6-4 \\ 5-3,5 & 7-4 \\ 2-3,5 & 1-4 \end{pmatrix}^T \begin{pmatrix} 1-3,5 & 1-4 \\ 6-3,5 & 7-4 \\ 3-3,5 & 2-4 \\ 4-3,5 & 6-4 \\ 5-3,5 & 7-4 \\ 2-3,5 & 1-4 \end{pmatrix} = \begin{pmatrix} 2,9 & 4,3 \\ 4,3 & 7,3 \end{pmatrix},$$

gdzie $\bar{\mathbf{x}} = (3,5 \quad 4)$ wektor średnich wartości macierzy \mathbf{X} .

Wymiar każdej macierzy $\mathbf{C}_1, \mathbf{C}_2$ wynosi $m \times m$, czyli 2×2 .

PRZYKŁAD (4)

2. Wyniki obliczenia funkcji gęstości mieszaniny rozkładów normalnych dla każdego obiektu dla każdego obiektu \mathbf{x}_i w klasie k są zawarte w tab. Wprowadźmy oznaczenie $f_{ik} = f(\mathbf{x}_i, \mu_k, \mathbf{C}_k)$.

Funkcje gęstości w klasie 1:

	\mathbf{x}_i	μ_1	$(\mathbf{x}_i - \mu_1)$	f_{i1}
\mathbf{x}_1	(1 1)	(3,5 3,8)	(-2,5 -2,8)	0,02
\mathbf{x}_2	(6 7)	(3,5 3,8)	(2,5 3,2)	0,03
\mathbf{x}_3	(3 2)	(3,5 3,8)	(-0,5 -1,8)	0,05
\mathbf{x}_4	(4 6)	(3,5 3,8)	(0,5 2,2)	0,03
\mathbf{x}_5	(5 7)	(3,5 3,8)	(1,5 3,2)	0,04
\mathbf{x}_6	(2 1)	(3,5 3,8)	(-1,5 -2,8)	0,06

PRZYKŁAD (5)

Funkcje gęstości w klasie 2:

	\mathbf{x}_i	μ_2	$(\mathbf{x}_i - \mu_2)$	f_{i2}
\mathbf{x}_1	(1 1)	(3,5 4)	(-2,5 -3)	0,03
\mathbf{x}_2	(6 7)	(3,5 4)	(2,5 3)	0,03
\mathbf{x}_3	(3 2)	(3,5 4)	(-0,5 -2)	0,04
\mathbf{x}_4	(4 6)	(3,5 4)	(0,5 2)	0,04
\mathbf{x}_5	(5 7)	(3,5 4)	(1,5 3)	0,05
\mathbf{x}_6	(2 1)	(3,5 4)	(-1,5 -3)	0,05

PRZYKŁAD (5)

3. Wyniki obliczenia oczekiwanych wartości prawdopodobieństw g_{ik} przynależności obiektu \mathbf{x}_i do klasy k są zawarte w tab.:

f_{i1}	f_{i2}	$\alpha_1 f_{i1}$	$\alpha_2 f_{i2}$	$\sum_{k=1}^2 \alpha_k f_{ik}$	g_{i1}	g_{i2}
0,02	0,03	0,010	0,015	0,025	0,40	0,60
0,03	0,03	0,015	0,015	0,030	0,50	0,50
0,05	0,04	0,025	0,020	0,045	0,56	0,44
0,03	0,04	0,015	0,020	0,035	0,43	0,57
0,04	0,05	0,020	0,025	0,045	0,44	0,56
0,06	0,05	0,03	0,025	0,055	0,55	0,45

PRZYKŁAD (6)

4. Obliczenie logarytmicznej funkcji wiarygodności l :

$$l = \sum_{i=1}^6 \ln \left(\sum_{k=1}^K \alpha_k f(\mathbf{x}_i, \mu_k, \mathbf{C}_k) \right) =$$
$$= \ln 0,023 + \ln 0,030 + \dots + \ln 0,055 = -19,65$$

PRZYKŁAD (7)

5. Wyznaczenie nowych wartości parametrów modelu:

➤ Ponieważ:

$$\sum_{i=1}^n g_{i1} = 0,40 + 0,50 + 0,56 + \dots + 0,55 = 2,88$$

$$\sum_{i=1}^n g_{i2} = 0,60 + 0,50 + 0,44 + \dots + 0,45 = 3,12,$$

to nowe współczynniki wagowe są następujące:

$$\alpha_1 = \frac{1}{n} \sum_{i=1}^n g_{i1} = \frac{2,88}{6} = 0,48,$$

$$\alpha_2 = \frac{1}{n} \sum_{i=1}^n g_{i2} = \frac{3,12}{6} = 0,52,$$

czyli nowe wartości wag są równe $\alpha_k = \{0,48; 0,52\}$.

PRZYKŁAD (8)

➤ Nowe wartości średnich.

i	\mathbf{x}_i	g_{i1}	g_{i2}	$g_{i1}\mathbf{x}_i$	$g_{i2}\mathbf{x}_i$
1	(1 1)	0,40	0,60	(0,40 0,40)	(0,60 0,60)
2	(6 7)	0,5	0,5	(3,00 3,50)	(3,00 3,50)
3	(3 2)	0,55	0,44	(1,65 1,10)	(1,32 0,88)
4	(4 6)	0,43	0,55	(1,72 2,58)	(2,28 3,42)
5	(5 7)	0,44	0,55	(2,20 3,08)	(2,80 3,92)
6	(2 1)	0,55	0,45	(1,10 0,55)	(0,90 0,45)
$\sum_{i=1}^n g_{ik}\mathbf{x}_i$				(10,07 11,21)	(10,90 12,77)

PRZYKŁAD (9)

$$\mu_1 = \frac{\sum_{i=1}^n g_{i1} \mathbf{x}_i}{n\alpha_1} = \frac{(10,07 \quad 11,21)}{6 \cdot 0,48} = (3,50 \quad 3,90)$$

$$\mu_2 = \frac{\sum_{i=1}^n g_{i2} \mathbf{x}_i}{n\alpha_2} = \frac{(10,90 \quad 12,77)}{6 \cdot 0,52} = (3,49 \quad 4,09)$$

czyli nowa macierz średnich jest równa:

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} = \begin{pmatrix} 3,50 & 3,90 \\ 3,49 & 4,09 \end{pmatrix}.$$

PRZYKŁAD (10)

➤ nowe wartości macierzy kowariancji (są różne w każdej klasie):

$$\mathbf{C}_k = \frac{\sum_{i=1}^n (\mathbf{x}_i - \mu_k)^T g_{ik} (\mathbf{x}_i - \mu_k)}{n\alpha_k}.$$

W celu ułatwienia obliczeń wzór do obliczenia nowych macierzy kowariancji można zapisać w postaci macierzowej:

$$\mathbf{C}_k = \frac{\mathbf{A}_k^T \mathbf{B}_k}{n\alpha_k},$$

gdzie każdy i -ty wierz macierzy \mathbf{A}_k i \mathbf{B}_k w klasie k jest obliczony następująco:

$$\mathbf{a}_{ik} = (\mathbf{x}_i - \mu_k)$$

$$\mathbf{b}_{ik} = g_{ik} \mathbf{a}_{ik}$$

PRZYKŁAD (11)

Wyniki pośrednie do obliczenia macierzy kowariancji \mathbf{C}_1 :

g_{i1}	\mathbf{x}_i	μ_1	$\mathbf{a}_{i1} = (\mathbf{x}_i - \mu_1)$	$\mathbf{b}_{i1} = g_{i1} \mathbf{a}_{i1}$
0,40	(1 1)	(3,5 3,9)	(-2,5 -2,9)	(-1 -1,16)
0,50	(6 7)	(3,5 3,9)	(2,5 3,1)	(1,25 1,55)
0,55	(3 2)	(3,5 3,9)	(-0,5 -1,9)	(-0,28 -1,04)
0,43	(4 6)	(3,5 3,9)	(0,5 2,1)	(0,21 0,9)
0,44	(5 7)	(3,5 3,9)	(1,5 3,1)	(0,66 1,36)
0,55	(2 1)	(3,5 3,9)	(-1,5 -2,9)	(-0,83 1,59)

PRZYKŁAD (12)

$$\mathbf{C}_1 = \frac{1}{6 \cdot 0,48} \begin{pmatrix} -2,5 & -2,9 \\ 2,5 & 3,1 \\ -0,5 & -1,9 \\ 0,5 & 2,1 \\ 1,5 & 3,1 \\ -1,5 & 2,9 \end{pmatrix}^T \begin{pmatrix} -1 & -1,16 \\ 1,25 & 1,55 \\ -0,28 & -1,04 \\ 0,21 & 0,9 \\ 0,66 & 1,36 \\ -0,83 & 1,59 \end{pmatrix} = \begin{pmatrix} 2,81 & 4,23 \\ 4,23 & 7,26 \end{pmatrix}.$$

Analogicznie obliczana jest macierz \mathbf{C}_2 :

$$\mathbf{C}_2 = \begin{pmatrix} 3,00 & 4,40 \\ 4,40 & 7,32 \end{pmatrix}.$$

PRZYKŁAD (13)

6. Porównanie wartości logarytmicznej funkcji wiarygodności z wartością funkcji z poprzedniej iteracji (dla pierwszej iteracji przyjmuje się, że $l^0 = 0$):

$$\Delta l = |-19,65 - 0| = 19,65$$

Ponieważ $\Delta l > \delta$, czyli $19,65 > 0,001$, wracamy do kroku 2.

Po wykonaniu 4 iteracji powstał następujący podział: obiekty 1, 4, 5 należą do klasy 1, a obiekty 2, 3, 6 należą do klasy 2.

METODY HIERARCHICZNE

Klasteryzacja hierarchiczna (*ang.* hierarchical cluster analysis, HCA) jest metodą analizy danych, która ma na celu zbudowanie hierarchii klastrów. Służy do dzielenia obserwacji na klastry bazując na podobieństwach między nimi. Nie wymaga określenia liczby tworzonych klastrów.

➤ **Metody aglomeracyjne** (*ang.* agglomerative) - metody, w których początkowo każdy obiekt jest odrębnym klastrem. Następnie obiekty są stopniowo łączone według pewnej reguły w nowe klastry, aż do uzyskania jednego klastra.

➤ **Metody deglomeracyjne** (*ang.* divisive), podziałowe - metody, w których początkowo wszystkie obiekty tworzą jeden klaster, który kolejno jest dzielony (rozszczepiony) na mniejsze i bardziej jednorodne, aż do momentu uzyskania jednoelementowych klastry.

METODY HIERARCHICZNE

Wyniki klasteryzacji hierarchicznej stanowią zestaw zagnieżdżonych klastrów, które są zwykle prezentowane w postaci dendrogramu lub w postaci tabeli przebiegu aglomeracji (deglomeracji).

Stosując algorytmy klasteryzacji hierarchicznej konieczne jest dokonanie pomiaru odległości między obiektami, czyli głównym celem jest to aby odległości między obiektami tego samego klastra były możliwie jak najmniejsze, natomiast odległości między klastrami były jak największe.

Metody deglomeracyjne są znacznie bardziej złożone, niż metody aglomeracyjne, ponieważ trudno podać korzyści, które wynikają z odwrotnej do aglomeracji budowy dendrogramu. Dalej będziemy rozpatrywać metody aglomeracyjne.

ODMIENNOŚĆ

W klasteryzacji hierarchicznej istnieją dwa zasadnicze parametry: ***miara odmienności*** oraz ***metoda połączenia***.

Odmienność między klastrami i -tym a j -tym oznaczmy d_{ij} oraz przyjmiemy, że klaster i -ty zawiera n_i obserwacji, natomiast klaster j -ty n_j obserwacji.

W postaci miary odmienności najczęściej występuje ***miara odległości*** między klastrami.

ODLEGŁOŚCI

Odległość	Wzór
Euklidesowa	$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{j=1}^m (x_j - y_j)^2}$
Euklidesowa do kwadratu	$d(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^m (x_j - y_j)^2$
Czebyszewa	$d(\mathbf{x}, \mathbf{y}) = \max_j \{ x_j - y_j \}$
Manhattan	$d(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^m x_j - y_j $
Machalanobisa	$d(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x} - \mathbf{y}) \mathbf{C}^{-1} (\mathbf{x} - \mathbf{y})^T}$

INNE

Istnieją jednak metody, w których w postaci odmienności występuje nie odległość, a suma kwadratów odchyleń wewnątrz klastrów:

$$d_k = \sum_{i=1}^{n_k} x_i^2 - \frac{1}{n_k} \left(\sum_{i=1}^{n_k} x_i \right)^2 \quad (1)$$

w przypadku jednej zmiennej, albo:

$$d_k = \sum_{i=1}^{n_k} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$$

w przypadku wielowymiarowym.

METODY POŁĄCZENIA

Metody połączenia określają, w jaki sposób zdefiniowana jest odmiennność między dwoma klastrami.

Ważne jest, aby w danym eksperymencie wypróbować kilka metod łączenia oraz porównać ich wyniki.

W zależności od zbioru danych, niektóre metody mogą działać lepiej.

METODY POŁĄCZENIA

Odmienność między i -tym a j -tym klastrami jest równa w:

1. **metodzie najbliższego sąsiada** (*ang.* single linkage):
najmniejszej ...
2. **metodzie najdalszego sąsiada** (*ang.* complete linkage):
największej ...
3. **metodzie średniego połączenia** (*ang.* average linkage)
uśrednionej wartości...
...spośród $n_i n_j$ odmienności między parami obserwacji, z których jedna pochodzi z i -tego klastra, a druga z j -tego klastra.
4. **metodzie Warda**: minimalnej sumie kwadratów odchyleń od punktów do centroid klastrów.

METODY POŁĄCZENIA

W metodzie najbliższego sąsiada powstają klastry typu "łańcuchów", co oznacza, że są one połączone ze sobą tylko przez pojedyncze obiekty, które leżą najbliżej siebie. Pozwala ta metoda na wykrycie punktów odstających, nie należących do żadnego klastra i przydatna jest dla wstępnej obróbki danych w celu eliminacji takich punktów.

Metoda najdalszego sąsiada jest dobra w przypadkach, gdy obiekty tworzą oddzielone grupy, ale nie jest odpowiednia, gdy klastry są w jakiś sposób wydłużone.

Metoda średniego połączenia jest lepsza w sytuacji gdy istnieją (lub podejrzewamy, że istnieją) znaczne różnice w rozmiarach klastrów.

Metoda Warda empirycznie daje bardzo dobre wyniki, jednak ponieważ na każdym etapie metody spośród wszystkich możliwych do łączenia par klastrów wybiera się ta para, która w rezultacie łączenia daje klaster o minimalnym zróżnicowaniu, to zmierza ona do tworzenia klastrów o małej wielkości.

ALGORYTM AGLOMERACYJNY

1. Budowa macierzy odmienności o wymiarach $n \times n$, która zawiera odmienność każdej pary obiektów. Macierz ta jest symetryczna względem głównej przekątnej, którą stanowią same zera.

2. Na podstawie wybranej metody połączenia, wybierane są z macierzy odmienności (poza główną przekątną) dwa obiekty i -ty i j -ty, których w największy stopniu dotyczy wybrana miara odmienności i tworzą z tych obiektów nowy k -ty klaster.

3. Klaster ten zajmuje w macierzy odmienności pozycję pod numerem i (pod warunkiem, że $i < j$). Jednocześnie usuwa się z macierzy obiekt o numerze j . W ten sposób wymiary macierzy odmienności zmniejszają się o 1.

4. Ponownie wyznaczana jest macierz odmienności dla nowego, zredukowanego układu obiektów. W przypadku, gdy nie został stworzony jedyny wspólny klaster, wracamy do kroku 2, inaczej kończymy proces klasteryzacji.

ALGORYTM AGLOMERACYJNY

Po każdym połączeniu klastrów w trakcie działania metody aglomeracyjnej konieczne jest obliczenie odmienności między powstałymi klastrami i klastrami innymi, powstałymi wcześniej.

Jeśli przykładowo klaster i -ty został połączony z klastrem j -tym, to odmienność między wcześniej powstałym skupieniem k -tym i nowym połączeniem klastrów i -tego i j -tego w jeden klaster można obliczać rekurencyjnie za pomocą wzoru Lance'a-Willims'a

WZÓR LANCE'A-WILLIMS'A

$$d_{k,ij} = a_i d_{ki} + a_j d_{kj} + b d_{ij} + c |d_{ki} - d_{kj}| \quad (2)$$

Wartości parametrów dla różnych metod łączenia a_i, a_j, b, c są przedstawione w tab.:

metoda łączenia	a_i	a_j	b	c
najbliższy sąsiad	0,5	0,5	0	-0,5
najdalszy sąsiad	0,5	0,5	0	0,5
średnie połączenie	$\frac{n_i}{n_i + n_j}$	$\frac{n_j}{n_j + n_i}$	0	0
metoda Warda	$\frac{n_k + n_i}{n_k + n_i + n_j}$	$\frac{n_k + n_j}{n_k + n_j + n_i}$	$-\frac{n_k}{n_k + n_i + n_j}$	

PRZYKŁAD 1

Wykorzystując metodę Warda, wykonać podział na klastry dla następujących danych jednowymiarowych:

2,5,9,10,15.

Tworzymy pierwsze klastry z par obiektów i budujemy macierz odmienności:

$$\begin{pmatrix} 0 & 4,5 & 24,5 & 32 & 84,5 \\ 4,5 & 0 & 8 & 12,5 & 50 \\ 24,5 & 8 & 0 & \mathbf{0,5} & 18 \\ 32 & 12,5 & \mathbf{0,5} & \mathbf{0} & 12,5 \\ 84,5 & 50 & 18 & 12,5 & \mathbf{0} \end{pmatrix}.$$

PRZYKŁAD 1

Przykładowo odmiennosc dla klastra, zawierającego pierwszy a drugi element zgodnie ze wzorem

$$d_k = \sum_{i=1}^{n_k} x_i^2 - \frac{1}{n_k} \left(\sum_{i=1}^{n_k} x_i \right)^2$$

jest równa:

$$d_{12} = (2^2 + 5^2) - \frac{1}{2}(2 + 5)^2 = 29 - 24,5 = 4,5.$$

PRZYKŁAD 1

0	4,5	24,5	32	84,5
4,5	0	8	12,5	50
24,5	8	0	0,5	18
32	12,5	0,5	0	12,5
84,5	50	18	12,5	0

W macierzy odmienności minimalna wartość sumy kwadratów odchyień jest dla obiektów x_3 oraz x_4 , więc tworzą one pierwszy klaster **K₁**.

PRZYKŁAD

Budujemy nową macierz odmienności dla elementów $x_1, x_2, \mathbf{K}_1, x_5$ odpowiednio:

$$\begin{pmatrix} 0 & 4,5 & 38 & 84,5 \\ 4,5 & 0 & 14 & 50 \\ 38 & 14 & 0 & 20,66 \\ 84,5 & 50 & 20,66 & 0 \end{pmatrix}.$$

Przykładowo dla obiektów x_5 i \mathbf{K}_1 (1):

$$d_{5,34} = (9^2 + 10^2 + 15^2) - \frac{1}{3}(9 + 10 + 15)^2 = 20,67$$

Obliczenie za pomocą wzoru Lance'a-Willims'a:

$$d_{5,34} = \frac{1+1}{1+1+1}18 + \frac{1+1}{1+1+1}12,5 + \frac{1+1}{1+1+1}0,5 = 20,17.$$

PRZYKŁAD

Można zauważyć, że wyniki nieznacznie się różnią, co jednak nie wpływa na końcowy podział na klastry. Dalej w przykładzie będziemy korzystać ze wzoru (1).

$$\begin{matrix} & x_1 & x_2 & \mathbf{K}_1 & x_5 \\ \left(\begin{array}{cccc} 0 & 4,5 & 38 & 84,5 \\ 4,5 & 0 & 14 & 50 \\ 38 & 14 & 0 & 20,66 \\ 84,5 & 50 & 20,66 & 0 \end{array} \right) \end{matrix} \cdot$$

Obiekty x_1, x_2 tworzą nowy klaster $\mathbf{K}_2 = \{x_1, x_2\}$

■

PRZYKŁAD 1

Nowa macierz odmienności dla elementów $\mathbf{K}_2, \mathbf{K}_1, x_5$:

$$\begin{matrix} & \mathbf{K}_2 & \mathbf{K}_1 & x_5 \\ \begin{pmatrix} 0 & 41 & 92,67 \\ 41 & 0 & 20,66 \\ 92,67 & 20,66 & 0 \end{pmatrix}, \end{matrix}$$

czyli tym razem dołączamy obiekt x_5 do klastra \mathbf{K}_1 i tworzymy klaster \mathbf{K}_3 , który zawiera obiekty x_3, x_4, x_5 .

Na ostatnim kroku pozostały tylko dwa obiekty \mathbf{K}_3 i \mathbf{K}_2 , które możemy połączyć w jeden klaster, obejmujący wszystkie 5 obiektów.

PRZYKŁAD 2

Dla danych z tab. wykonać klasteryzację obiektów $\mathbf{x}_i = \mathbf{X}(i,:)$.
W postaci miary odmienności wybrać odległość Euklidesową, w postaci metody łączenia - metodę najbliższego sąsiada.

X_1	X_2	X_3	X_4
39,8	38	22,2	23,2
53,7	37,2	18,7	18,5
47,3	39,8	23,3	22,1
41,7	37,6	22,8	22,3
44,7	38,5	24,8	24,4
47,9	39,8	22,0	23,3

PRZYKŁAD 2

0,0	4,08	2,35	0,75	1,78	2,31
4,08	0,0	3,93	3,68	4,70	3,89
2,35	3,93	0,0	2,30	1,87	0,88
0,75	3,68	2,30	0,0	1,75	2,43
1,78	4,70	1,87	1,75	0,0	2,00
2,31	3,89	0,88	2,43	2,00	0,0

Wybieramy najmniejsze wartości w macierzy odległości (poza główną przekątną) i tworzymy nowy klaster z obiektów, których ta najmniejsza odległość dotyczy.

Jest to odległość między obiektem pierwszym \mathbf{x}_1 a czwartym \mathbf{x}_4 . Łączymy te obiekty w jeden nowy klaster $\mathbf{K}_1 = \{\mathbf{x}_1, \mathbf{x}_4\}$ i stawiamy go na miejsce obiektu \mathbf{x}_1 a \mathbf{x}_4 usuwamy z macierzy.

PRZYKŁAD 2

Budujemy nową macierz odległości dla nowego, zredukowanego układu obiektów: \mathbf{K}_1 oraz wszystkie obiekty oprócz \mathbf{x}_1 i \mathbf{x}_4 .

Przykładowo odległość między obiektem \mathbf{x}_2 a klastrem \mathbf{K}_1 jest równa:

$$d_{2,14} = \min\{d_{21}, d_{24}\} = \min\{4,08; 3,68\} = 3,68,$$

albo inaczej za pomocą wzoru Lance'a-Willims'a:

$$\begin{aligned} d_{2,14} &= \frac{1}{2}(d_{21} + d_{24} - |d_{21} - d_{24}|) = \\ &= \frac{1}{2}(4,08 + 3,68 - |4,08 - 3,68|) = 3,68 \end{aligned}$$

PRZYKŁAD 2

0,0	3,68	2,30	1,75	2,31
	0,0	3,93	4,70	3,89
		0,0	1,87	0,88
			0,0	2,00
				0,0

W macierzy tej pierwszy wiersz (kolumna) należą do klastra K_1 , a dalej są odpowiednio odległości dla obiektów x_2, x_3, x_5, x_6 po kolei. Ze względu na symetrię macierzy pokazana tylko połowa.

Wykorzystując nową macierz odległości, znajdujemy kolejną najmniejszą odległość. Jest to odległość między obiektem x_3 i x_6 . Łączymy te obiekty w nowy klaster $K_2 = \{x_3, x_6\}$ stawiamy go na miejsce obiektu x_3 , usuwamy obiekt x_6 i powtarzamy krok 2.

PRZYKŁAD 2

Zgodnie z metodą najbliższego sąsiada odległość między klastrami \mathbf{K}_1 i \mathbf{K}_2 jest równa:

$$d_{14,36} = \min\{d_{13}, d_{16}, d_{43}, d_{46}\} = \min\{2,35; 3,68; 2,30; 2,43\} = 2,30$$

Kolejna macierz odległości, w której pierwszy wiersz \mathbf{K}_1 drugi wiersz \mathbf{x}_2 , dalej klaster \mathbf{K}_2 i obiekt \mathbf{x}_5 :

$$\begin{pmatrix} 0 & 3,68 & 2,30 & \mathbf{1,75} \\ & \mathbf{0} & 3,89 & 4,70 \\ & & 0 & 1,87 \\ & & & 0 \end{pmatrix}$$

Najmniejsza odległość jest pomiędzy obiektem \mathbf{x}_5 i klastrem \mathbf{K}_1 , więc 5 obiekt dołączony jest do tego klastra, czyli $\mathbf{K}_3 = \{\mathbf{x}_5, \mathbf{K}_1\}$.

PRZYKŁAD 2

Nowa macierz odległości (dla $\mathbf{K}_3, \mathbf{K}_2, \mathbf{x}_2$):

$$\begin{pmatrix} 0 & 1,87 & 3,68 \\ & 0 & 3,89 \\ & & 0 \end{pmatrix},$$

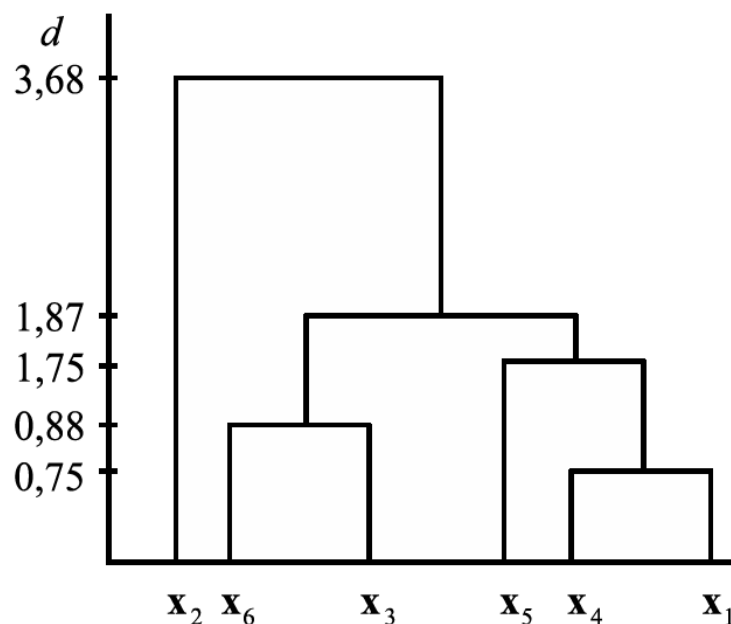
czyli łączone są klastry \mathbf{K}_2 i \mathbf{K}_3 w jeden klaster $\mathbf{K}_4 = \{\mathbf{K}_2, \mathbf{K}_3\} = \{\mathbf{x}_1, \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5, \mathbf{x}_6\}$.

Ostatnia macierz odległości (między \mathbf{K}_4 i \mathbf{x}_2):

$$\begin{pmatrix} 0 & 3,68 \\ & 0 \end{pmatrix}.$$

DENDROGRAM

Otrzymane rezultaty możemy przedstawić za pomocą **dendrogramu** (wykresu drzewkowego), ilustrującego hierarchiczną strukturę zbioru obiektów ze względu na zmniejszające się podobieństwo między nimi:



OKREŚLENIE LICZBY KLASTRÓW

W metodach aglomeracyjnych zalecana jest standaryzacja danych przed rozpoczęciem klasteryzacji, ponieważ różne wymiary zmiennych X mogą wpływać na końcowy wynik.

Przy wykorzystaniu metod hierarchicznych należy odpowiedzieć na pytanie, jaka powinna być liczba klastrow na które należałoby podzielić dany zbiór obiektów.

W tym celu należy analizować dendrogram pod względem różnic odległości między kolejnymi węzłami. Duża wartość różnic oznacza, że klastry są odległe (odległość między kolejnymi węzłami jest duża) i w tym miejscu można dokonać podziału.

Nie ma obiektywnej reguły ustalenia liczby klastrow, ale istnieją pewne reguły, które pomagają podnieść decyzję.

OKREŚLENIE LICZBY KLASTRÓW

- Reguła maksimum.
- Reguła Grabińskiego.
- Reguła Mojena: punktem odcięcia jest odległość d_i , dla której spełniona jest nierówność:

$$d_i > \bar{d} + \lambda s_d,$$

gdzie \bar{d}, s_d - średnia i odchylenie standardowe dla wartości odległości, a λ - pewna stała.

Mojen zasugerował, że wartość $\lambda \in [2,75; 3,50]$ daje zadowalające wyniki. Z kolei Miligan i Cooper (1985) sugerują, że wartość $\lambda = 1,25$ daje wyniki najlepsze.

PRZYKŁAD 3

Dla danych (standaryzowanych) z tab. określić optymalną liczbę klastrow.

X_1	X_2	X_3	X_4
-1,02	-0,90	-1,34	0,82
0,89	0,94	0,49	-1,36
-0,55	-1,30	-0,31	-0,96
-0,80	-0,39	1,87	1,22
0,96	0,94	0,26	1,02
-1,04	-0,90	-0,88	-0,96
0,99	1,04	1,01	-1,36
-1,09	-1,00	-0,77	1,02
0,92	0,84	0,66	0,43
0,94	0,94	0,26	1,02
0,94	1,04	0,49	-0,17
-1,11	-1,30	-1,00	-0,56
-1,06	-0,79	-1,57	0,82
1,01	-0,84	0,83	-0,96

PRZYKŁAD 3

Przebieg aglomeracji:

	d														
1	0,10	x_5	x_{10}												
2	0,42	x_7	x_{14}												
3	0,46	x_1	x_{13}												
4	0,51	x_9	x_{11}												
5	0,57	x_6	x_{12}												
6	0,68	x_9	x_{11}	x_5	x_{10}										
7	0,72	x_9	x_{11}	x_2	x_5	x_{10}									
8	0,77	x_9	x_{11}	x_2	x_5	x_{10}	x_7	x_{14}							
9	0,95	x_6	x_{12}	x_8											
10	1,05	x_6	x_{12}	x_8	x_1	x_{13}									
11	2,27	x_6	x_{12}	x_8	x_1	x_{13}	x_3								
12	4,17	x_6	x_{12}	x_8	x_1	x_{13}	x_3	x_4							
13	6,45	x_6	x_{12}	x_8	x_1	x_{13}	x_3	x_4	x_9	x_{11}	x_2	x_5	x_{10}	x_7	x_{14}

PRZYKŁAD 3

Wyniki obliczenia wartości różnic i ilorazów odległości d_i .

	d_i	d_{i-1}	$d_i - d_{i-1}$	d_i / d_{i-1}
1	0,10			
2	0,42	0,10	0,32	4,20
3	0,46	0,42	0,03	1,08
4	0,51	0,46	0,05	1,11
5	0,57	0,51	0,06	1,13
6	0,68	0,57	0,10	1,18
7	0,72	0,68	0,05	1,07
8	0,77	0,72	0,05	1,06
9	0,95	0,77	0,17	1,22
10	1,05	0,95	0,10	1,11
11	2,27	1,05	1,21	2,16
12	4,17	2,27	1,9	1,84
13	6,45	4,17	2,27	1,54

PRZYKŁAD 3

Maksymalna wartość różnic odległości jest równa 2,27, krok 13, ale ponieważ na kroku 13 wszystkie obiekty zostały połączone w jedną grupę, to wybiera się krok 12, czyli dane za pomocą a reguły maksimum są podzielone na dwa klastry. Do pierwszego klastra należą obiekty 9,11,2,5,10,7,14, a do drugiego obiekty 6,12,8,1,13,3,4.

Reguła Grabińskiego: Jeżeli nie brać pod uwagę kilka pierwszych wartości ilorazu $\frac{d_i}{d_{i-1}}$, to można zauważyć, że iloraz ten przyjmuje wartość największą na kroku 11, czyli stworzone są trzy klastry. Do pierwszego klastra należą obiekty 9,11,2,5,10,7,14, do drugiego klastra należą obiekty 6,12,8,1,13,3 i obiekt 4 tworzy nowy jednoelementowy klaster.

PRZYKŁAD 3

Reguła Mojena: ponieważ dla odległości d_i średnia i odchylenie standardowe są odpowiednio równe

$$\bar{d} = 1,47, s_d = 1,84,$$

to wybierając $\lambda = 1,25$ można określić:

$$d_i > \bar{d} + 1,25 \cdot s_d = 1,47 + 1,25 \cdot 1,84 = 3,77$$

czyli za pomocą reguły Mojena sugerowany jest podział na kroku **12**, dla którego $d_{12} = 4,17 > 3,77$, czyli podział jest taki sam jak dla reguły maksimum:

do pierwszego klastra należą obiekty 9,11,2,5,10,7,14, a do drugiego obiekty 6,12,8,1,13,3,4.

K-MEANS

$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1 \\ \dots \\ \mathbf{x}_n \end{pmatrix}_{n \times m}$ - zbiór danych n obserwacji, każda obserwacja jest

opisana przez m zmiennych (cech, atrybutów).

Klasteryzacja polega na znalezieniu K -elementowej partycji $C = \{C_1 \ C_2 \ \dots \ C_K\}$ zbioru \mathbf{X} **maksymalizującej** pewną miarę jakości grupowania danych $F(C)$, czyli należy znaleźć K parami rozłącznych zbiorów $C_1 \ C_2 \ \dots \ C_K$, takich, że $C_1 \cup C_2 \cup \dots \cup C_K = \mathbf{X}$.

W jaki sposób zdefiniować miarę jakości grupowania? Chcemy, aby każde dwa elementy należące do tej samej grupy były do siebie podobne, zaś każde dwa elementy należące do dwóch różnych grup były do siebie niepodobne.

ZAŁOŻENIA

Założmy, że potrafimy określić:

1. pewną miarę podobieństwa $\rho(x, y)$ między obserwacjami x i y ;
2. pewną miarę odległości $d(x, y)$ mierzącą odległości między obserwacjami x i y ;
3. zazwyczaj podobieństwo jest ujemnie skorelowane z odległością, na przykład $d(x, y) = \frac{1}{\rho(x, y)}$.

Ogólnie możliwe są różne podejścia do mierzenia jakości grupowania, które prowadzą do różnych algorytmów oraz różnych wyników grupowania tych samych danych. W konkretnej sytuacji wybór podejścia powinien zależeć od charakterystyki analizowanych danych oraz konkretnych potrzeb i konkretnych oczekiwań analityka danych.

PODEJŚCIE 1

Całkowita jakość grupowania:

$$F(C) = \frac{\sum_{k=1}^K WCV(C_k)}{\sum_{1 \leq k < l \leq K} BCV(C_k, C_l)},$$

$$WCV(C_k) = \frac{1}{|C_k| - 1} \sum_{\substack{x \in C_k \\ y \in C_k \\ x \neq y}} \rho(\mathbf{x}, \mathbf{y}) \quad - \text{ średnie podobieństwo}$$

elementów w grupie (zmienność wewnątrz grupy – *ang.* within-cluster variation) C_k ; $|C_k|$ - liczba elementów w grupie C_k ;

$$BCV(C_k, C_l) = \frac{1}{|C_k| |C_l|} \sum_{x \in C_k} \sum_{y \in C_l} \rho(\mathbf{x}, \mathbf{y}) \quad - \text{ średnie podobieństwo}$$

elementów w każdych dwóch grupach (zmienność pomiędzy grupami - *ang.* between-cluster variation).

PODEJŚCIE 1

Podobną funkcję możemy określić w oparciu o odległości.

Podejście takie jest niepraktyczne ze względu na złożoność obliczeniową.

PODEJŚCIE 2

Dla każdej grupy C_k można wyznaczyć centrum grupy \mathbf{c}_k określone jako centrum ciężkości punktów:

$$\mathbf{c}_k = \frac{1}{|C_k|} \sum_{x \in C_k} \mathbf{x}.$$

Chcielibyśmy, aby nasza metodyka grupowania maksymalizowała zmienność pomiędzy grupami względem zmienności wewnątrz grupy.

PODEJŚCIE 2

Obliczane są dwie miary - odchylenie wewnątrzgrupowe $WCV(C)$ oraz odchylenie międzygrupowe $BCV(C)$:

$$WCV(C) = \sum_{k=1}^K WCV(C_k) = \sum_{k=1}^K \sum_{x \in C_k} d(\mathbf{x}, \mathbf{c}_k),$$

$$BCV(C) = \sum_{1 \leq k < l \leq K} \sum_{x \in C_k} d(\mathbf{c}_k, \mathbf{c}_l),$$

za pomocą których określa się jakość grupowania.

Przykładowo może być to miara postaci: $F(C) = \frac{BCV(C)}{WCV(C)}$.

Także w postaci $WCV(C)$ można brać wartość średniokwadratową $SSE = \sum_{k=1}^K \sum_{x \in C_k} d^2(\mathbf{x}, \mathbf{c}_k)$.

PODEJŚCIE 2

Jeśli każda zmienna jest numeryczna, a odległość jest Euklidesowa, to dla każdej grupy C_k określa się macierz kowariancji \mathbf{W}_k (nieunormowana) elementów grupy:

$$\mathbf{W}_k = \sum_{\mathbf{x} \in C_k} (\mathbf{x} - \mathbf{c}_k)^T (\mathbf{x} - \mathbf{c}_k), \dim \mathbf{W} = m \times m;$$

Wtedy macierz rozrzutu wewnątrzgrupowego:

$$\mathbf{W} = \sum_{k=1}^K \mathbf{W}_k.$$

Macierz rozrzutu międzygrupowego:

$$\mathbf{M} = \sum_{k=1}^K |C_k| (\mathbf{c}_k - \mu)^T (\mathbf{c}_k - \mu), \dim \mathbf{M} = m \times m,$$

gdzie μ - estymowana wartość średnia wszystkich punktów danych \mathbf{X} .

PODEJŚCIE 2

Popularne funkcje oceny jakości grupowania danych opierają się na macierzach **W**, **M**, m.in.:

1. $tr(\mathbf{W})$;
2. $\det(\mathbf{W})$;
3. $tr(\mathbf{MW}^{-1}) \dots$

$tr(\mathbf{W}) = \sum_{i=1}^m w_{ii}$ - ślad macierzy – suma elementów na głównej przekątnej macierzy kwadratowej

PODEJŚCIE 2

Wadą miary $tr(\mathbf{W})$ jest zależność od skali poszczególnych zmiennych. Zmieniając bowiem jednostkę jednej ze zmiennych, możemy otrzymać inną strukturę grupowania. Miara ta zazwyczaj prowadzi do kulistych kształtów grup, często dość zwartych i równolicznych.

Miara $\det(\mathbf{W})$ nie ma zależności skali, więc wykrywa też grupy eliptyczne. Preferuje również grupy równoliczne.

Miara $tr(\mathbf{MW}^{-1})$ preferuje grupy równoliczne o podobnych kształtach. Często tworzy grupy współliniowe.

ODLEGŁOŚĆ

Różne miary jakości grupowania danych prowadzą do różnych algorytmów grupowania. Algorytmy wykrywające grupy definiowane w oparciu o centra grup to algorytm *k*-means, LAD oraz EM.

Kluczowe znaczenie dla grupowania ma miara odległości $d(\mathbf{x}, \mathbf{y})$ w przestrzeni danych mierząca odległość między wektorami danych \mathbf{x} i \mathbf{y} .

ODLEGŁOŚĆ

1. odległość **euklidesowa**:

$$d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\| = \sqrt{\sum_{j=1}^m (x_j - y_j)^2} = \sqrt{(\mathbf{x} - \mathbf{y})(\mathbf{x} - \mathbf{y})^T};$$

2. odległość **Minkowskiego** (niekiedy zwana I_m):

$$d(\mathbf{x}, \mathbf{y}) = \left(\sum_{j=1}^m (x_j - y_j)^r \right)^{\frac{1}{r}}, \text{ gdzie } r \text{ jest pewną stałą.}$$

W przestrzeni euklidesowej:

- dla $r = 2$, otrzymujemy odległość euklidesową (I_2);
- dla $r = 1$, otrzymujemy odległość Manhattan (I_1);
- dla $r \rightarrow \infty$, otrzymujemy odległość Czebyszewa (I_∞).

ODLEGŁOŚĆ

Częstym problemem jest nieodporność algorytmów grupowania na skalowanie poszczególnych wymiarów, na przykład zmiana jednostek jednej ze zmiennych (z mm na km) może prowadzić do zupełnie innych wyników grupowania. Uniknąć tego można wprowadzając ważenie wymiarów w definicji odległości.

3. **Ważona** odległość **euklidesowa**:

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{j=1}^m a_j (x_j - y_j)^2} = \sqrt{(\mathbf{x} - \mathbf{y}) \mathbf{A} (\mathbf{x} - \mathbf{y})^T},$$

gdzie a_j - wagi kolejnych wymiarów (pewne stałe), \mathbf{A} - macierz diagonalna z wartościami a_j na głównej przekątnej. Ważenie wymiarów można rozszerzyć, dopuszczając, aby macierz \mathbf{A} nie była diagonalna.

ODLEGŁOŚĆ

4. Odległość ***Mahalanobisa*** w przypadku, gdy $\mathbf{A} = \mathbf{V}^{-1}$, gdzie $\mathbf{V} = \frac{1}{n} \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})^T (\mathbf{x}_i - \bar{\mathbf{x}})$ - macierz kowariancji zbioru danych:

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{j=1}^m v_j (x_j - y_j)^2} = \sqrt{(\mathbf{x} - \mathbf{y}) \mathbf{V}^{-1} (\mathbf{x} - \mathbf{y})^T}.$$

ZADANIE

\mathbf{X} - zbiór danych n obserwacji, każda obserwacja jest opisana przez m zmiennych; K - liczba grup, które należy utworzyć; każda grupa C_k reprezentowana jest przez centrum grupy \mathbf{c}_k .

Każdy wektor danych (obiekt) $\mathbf{x}_i = \mathbf{X}(i,:)$ jest przypisywany do grupy C_k , której centrum \mathbf{c}_k jest mu najbliższe (w przypadku równych odległości od kilku centrów, decyduje ustalona kolejność rozpatrywania grup lub przypisanie jest losowe).

Zdanie polega na znalezieniu takich podzbiorów zbioru \mathbf{X} $C_1 \cup C_2 \cup \dots C_K = \mathbf{X}$ minimalizujących funkcje:

$$\sum_{k=1}^K \sum_{\mathbf{x} \in C_k} \|\mathbf{x} - \mathbf{c}_k\|^2.$$

KROKI MINIMALIZACJI

Minimalizacja taka może przebiegać w dwóch krokach powtarzanych iteracyjnie:

1. znając wektory \mathbf{c}_k , wyznaczyć optymalne przypisanie wektorów danych \mathbf{x}_i do grup: każdy wektor danych powinien być przypisany do grupy reprezentowanej przez najbliższy mu wektor \mathbf{c}_k ;
2. znając przypisanie wektorów danych do grup, wyznaczyć wektory \mathbf{c}_k : najczęstszym rozwiązaniem jest ustalenie wektorów \mathbf{c}_k w środkach geometrycznych zbioru punktów, tworzących grupę.

ALGORYTM

Krok 1 Wprowadzamy liczbę grup k (wybór użytkownika).

Krok 2 Losowo przypisujemy k obserwacji jako początkowe środki grup.

Krok 3 Dla każdej obserwacji wyszukujemy najbliższy środek grupy. Każda obserwacja jest następnie przypisywana do najbliższego jej środka. Mamy zatem k grup (każda z grup jest niepusta, bo zawiera przynajmniej swój środek).

Krok 4 Następnie ponownie obliczamy środki (średnie wartości współrzędnych reprezentujących wartości zmiennych dla danej obserwacji) oparte na obserwacjach, które są obecnie związane z każdym klastrem. To zapewni nam nowy zbiór k środków.

Krok 5 Powtarzaj kroki od 3 do 5, aż do zakończenia działania algorytmu.

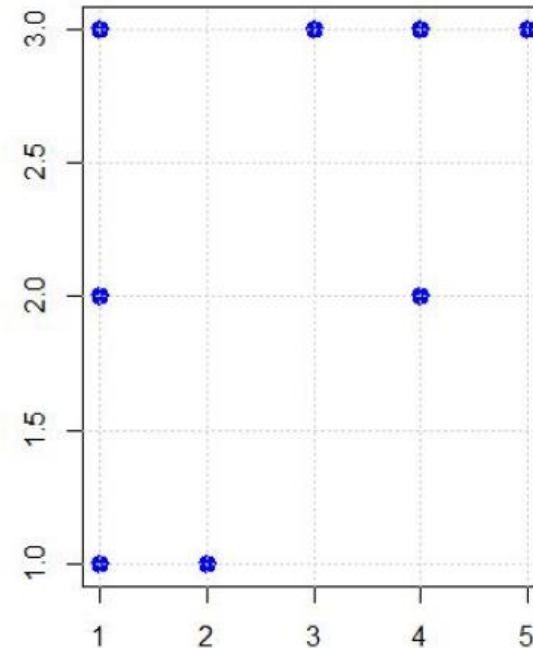
ZAKOŃCZENIE ALGORYTMU

- klastry nie ulegają zmianie, tzn. obserwacje nie „przechodzą” z jednego klastra do innego;
- spełnione jest pewne kryterium zbieżności.

PRZYKŁAD

Dla zbioru \mathbf{X} wykonać podział na 2 klasy. W postaci miary jakości wybrać $F(C) = \frac{BCV}{SSE}$, gdzie $SSE = \sum_{k=1}^K \sum_{x \in C_k} d^2(\mathbf{x}, \mathbf{c}_k)$.

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1 \\ \dots \\ \mathbf{x}_n \end{pmatrix}_{8 \times 2} = \begin{pmatrix} 1 & 3 \\ 3 & 3 \\ 4 & 3 \\ 5 & 3 \\ 1 & 2 \\ 4 & 2 \\ 1 & 1 \\ 2 & 1 \end{pmatrix}$$



I ITERACJA: MACIERZ ŚRODKÓW

Krok 2. Macierz środków:

$$\mathbf{C} = \begin{pmatrix} \mathbf{c}_1 \\ \mathbf{c}_2 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 2 & 1 \end{pmatrix}.$$

Krok 3 .Dla każdej obserwacji wyszukujemy najbliższy środek grupy. Skorzystamy ze wzoru na odległość euklidesową. Przykładowo dla pierwszego obiektu:

$$d(\mathbf{x}_1, \mathbf{c}_1) = \sqrt{(1-1)^2 + (3-1)^2} = \sqrt{4} = 2 \text{ albo w postaci wektorowej:}$$

$$(\mathbf{x}_1 - \mathbf{c}_1) = (1 \ 3) - (1 \ 1) = (0 \ 2),$$

$$d(\mathbf{x}_1, \mathbf{c}_1) = \sqrt{(\mathbf{x}_1 - \mathbf{c}_1)(\mathbf{x}_1 - \mathbf{c}_1)^T} = \sqrt{(0 \ 2) \begin{pmatrix} 0 \\ 2 \end{pmatrix}} = \sqrt{4} = 2.$$

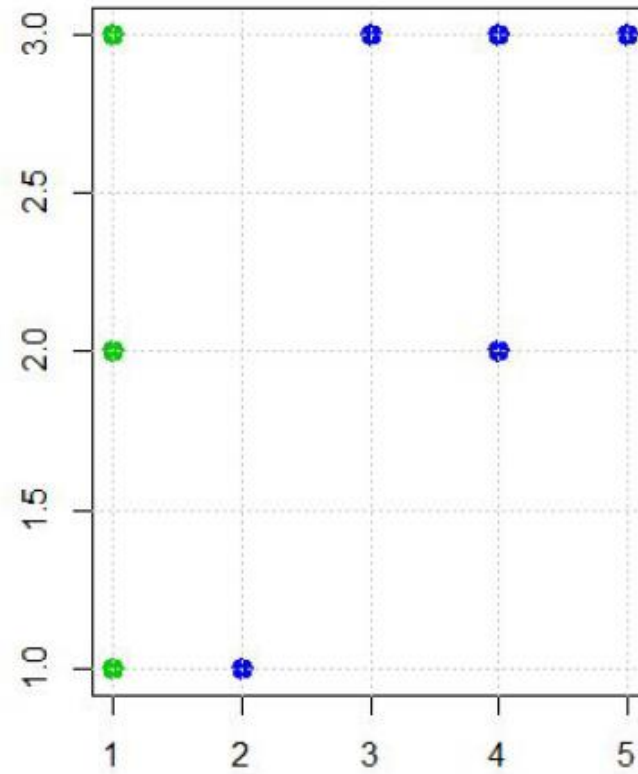
I ITERACJA: ODLEGŁOŚCI

Wyniki obliczenia odległości między każdym punktem a każdym środkiem grupy są zawarte w tabeli:

punkty	$d(\mathbf{x}_i, \mathbf{c}_1)$	$d(\mathbf{x}_i, \mathbf{c}_2)$	Przynależność do klasy
1	2,00	2,24	C_1
2	2,83	2,24	C_2
3	3,61	2,83	C_2
4	4,47	3,61	C_2
5	1,00	1,41	C_1
6	3,16	2,24	C_2
7	0,00	1,00	C_1
8	1,00	0,00	C_2

Czyli obiekty 1,5,7 należą do 1 grupy, a obiekty 2,3,4,6,8 do grupy 2.

I ITERACJA: ODLEGŁOŚCI



I ITERACJA: F(C)

Kiedy przynależność do grup została już przypisana, można obliczyć sumę błędów kwadratowych:

$$SSE = \sum_{k=1}^K \sum_{\mathbf{x} \in C_k} d^2(\mathbf{x}, \mathbf{c}_k) = 2^2 + 2,83^2 + 3,61^2 + \dots + 0^2 + 1^2 + 2,24^2 + \\ + 2,24^2 + \dots + 1,41^2 + 2,24^2 + 1^2 + 0^2 = 36$$

$$(\mathbf{c}_1 - \mathbf{c}_2) = \begin{pmatrix} 1 & 1 \end{pmatrix} - \begin{pmatrix} 2 & 1 \end{pmatrix} = \begin{pmatrix} -1 & 0 \end{pmatrix}$$

$$BCV = d(\mathbf{c}_1, \mathbf{c}_2) = \sqrt{(\mathbf{c}_1 - \mathbf{c}_2)(\mathbf{c}_1 - \mathbf{c}_2)^T} = \sqrt{\begin{pmatrix} -1 & 0 \end{pmatrix} \begin{pmatrix} -1 \\ 0 \end{pmatrix}} = \sqrt{1} = 1.$$

$$F(C) = \frac{BCV}{SSE} = \frac{1}{36} = 0,0278.$$

Oczekujemy, że ten współczynnik będzie zwiększany w kolejnych iteracjach, ponieważ chcemy zwiększyć zmienność między grupami względem zmienności wewnątrz grup.

I ITERACJA: NOWE ŚRODKI

Krok 4. Dla każdego klastra C_k szukamy nowy środek.

$$\mathbf{c}_1 = \left(\frac{1+1+1}{3} \quad \frac{3+2+1}{3} \right) = (1 \quad 2)$$

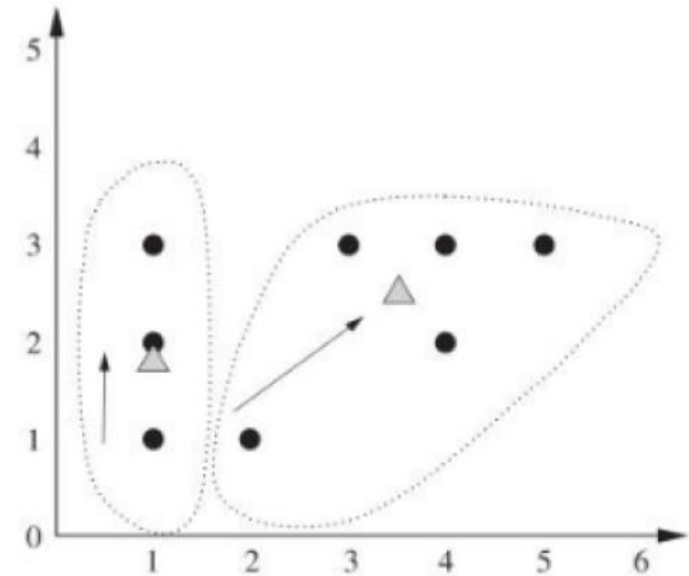
$$\mathbf{c}_2 = \left(\frac{3+4+5+4+2}{5} \quad \frac{3+3+3+2+1}{5} \right) = (3,6 \quad 2,4)$$

I ITERACJA: NOWE ŚRODKI

➤ c_1 z punktu (1 1) przesunęło się do punktu (1 2), do góry, do środka trzech punktów w grupie 1.

➤ Natomiast drugi środek c_2 przesunął się z punktu (2 1) do (3,6 2,4) do góry i w prawo.

➤ Ponieważ środki zostały przesunięte, to wracamy do kroku 3 i powtarzamy kroki 3 i 4 aż do zbieżności lub zakończenia.



Grupy i środki po pierwszej iteracji.

II ITERACJA: ODLEGŁOŚCI

Krok 3. Dla każdej obserwacji wyznaczamy najbliższy środek grupy. Tabela pokazuje odległości pomiędzy każdym punktem a każdym uaktualnionym środkiem grupy $\mathbf{c}_1 = (1 \ 2)$, $\mathbf{c}_2 = (3,6 \ 2,4)$ razem z przynależnością do grupy.

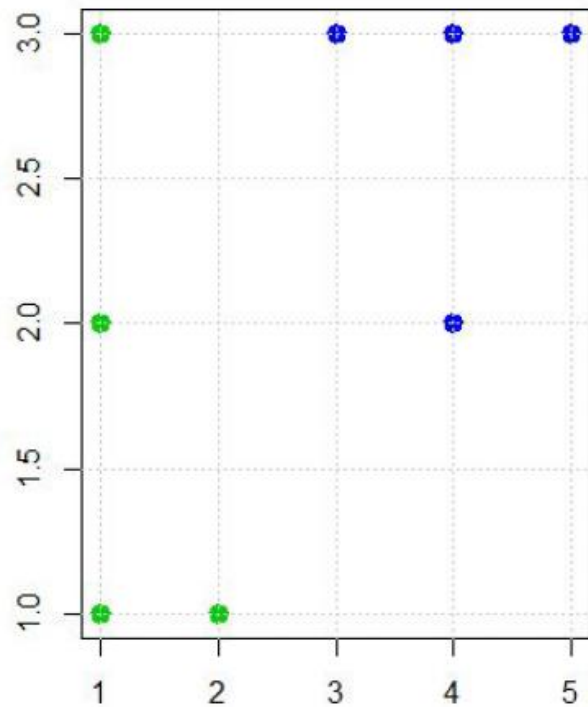
punkty	$d(\mathbf{x}_i, \mathbf{c}_1)$	$d(\mathbf{x}_i, \mathbf{c}_2)$	Przynależność do klasy
1	1,00	2,67	C_1
2	2,24	0,85	C_2
3	3,16	0,72	C_2
4	4,12	1,52	C_2
5	0,00	2,63	C_1
6	3,00	0,57	C_2
7	1,00	2,95	C_1
8	1,41	2,13	C_1

II ITERACJA: ODLEGŁOŚCI

punkty	$d(\mathbf{x}_i, \mathbf{c}_1)$	$d(\mathbf{x}_i, \mathbf{c}_2)$	Przynależność do klasy
1	1,00	2,67	C_1
2	2,24	0,85	C_2
3	3,16	0,72	C_2
4	4,12	1,52	C_2
5	0,00	2,63	C_1
6	3,00	0,57	C_2
7	1,00	2,95	C_1
8	1,41	2,13	C_1

Zauważmy, że mamy przesunięcie pojedynczego obiektu z grupy 2 do grupy 1. Stosunkowo duża zmiana wartości \mathbf{c}_2 sprawiła, że obserwacja 8 znajduje się teraz bliżej \mathbf{c}_1 niż \mathbf{c}_2 , dlatego obiekt 8 należy teraz do grupy 1. Wszystkie pozostałe obserwacje pozostają w tych samych grupach co poprzednio.

II ITERACJA: $F(C)$



Punkty w drugiej iteracji

$SSE = 7,86$, (mniej, niż poprzednio);
 $BCV = 2,63$.

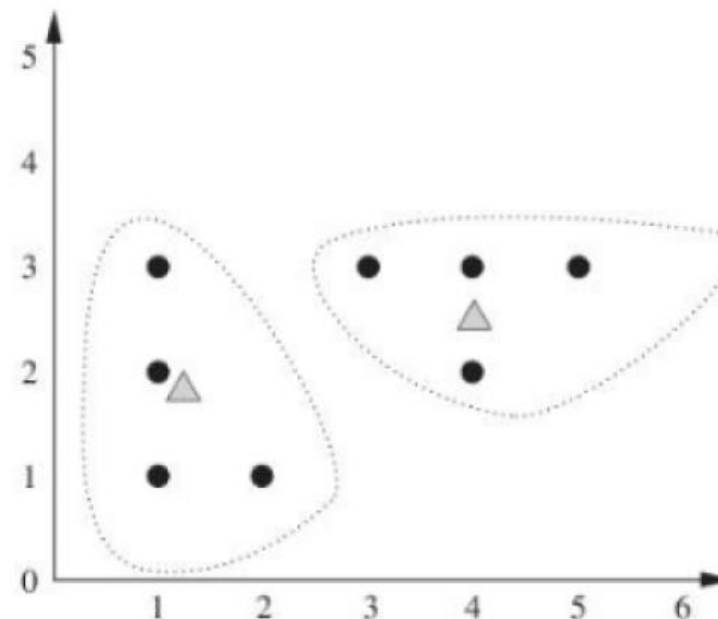
$F(C) = 0,3346$ ma wartość **większą**, niż poprzednio.

II ITERACJA: NOWE ŚRODKI

Krok 4.

$$\mathbf{c}_1 = \left(\frac{1+1+1+2}{4} \quad \frac{3+2+1+1}{4} \right) = (1,25 \quad 1,75)$$

$$\mathbf{c}_2 = \left(\frac{3+4+5+4}{4} \quad \frac{3+3+3+2}{5} \right) = (4 \quad 2,75)$$



Grupy i środki po drugiej iteracji.

Środki zostały przesunięte, więc wracamy do kroku 3 w trzeciej iteracji.

III ITERACJA: ODLEGŁOŚCI

Krok 3. Zauważmy, że obserwacje nie zmieniły przynależności do grupy z poprzedniej iteracji:

punkty	$d(\mathbf{x}_i, \mathbf{c}_1)$	$d(\mathbf{x}_i, \mathbf{c}_2)$	Przynależność do klasy
1	1,70	3,01	C_1
2	2,15	1,03	C_2
3	3,02	0,25	C_2
4	3,95	1,03	C_2
5	0,35	3,09	C_1
6	2,75	0,75	C_2
7	0,79	3,47	C_1
8	1,06	2,66	C_1

III ITERACJA: $F(C)$

$SSE = 6,23$, jest niewiele mniejsza, niż poprzednia wartość (7,86), co wskazuje na bliskość najlepszego rozwiązania grupowania.

$$BCV = 2,97.$$

$F(C) = 0,4703$ ma wartość **większą**, niż poprzednio (0,3347). Wskazuje to na zwiększenie zmienności pomiędzy grupami względem zmienności wewnątrz grup.

Krok 4. Dla każdej z k grup szukamy środka grupy. Ponieważ żadna obserwacja nie zmieniła przynależności do grupy, środki grup również pozostaną niezmiennione.

Krok 5: Ponieważ środki pozostały niezmiennione, algorytm kończy działanie.

Końcowa klasyfikacja: 1,5,7,8 – grupa 1, 2,3,4,6 – grupa 2.

INNE

Problemem algorytmu k -środków jest decyzja, ile grup należy szukać. W tym celu pomaga sprawdzian krzyżowy, w wyniku działania którego wybiera się wartości k z minimalną wartością SSE.

Istnieje wiele różnych modyfikacji algorytmów k-means: algorytm Fuzzy C-Means (FCM), algorytm Possibilistic C-Means (PCM), algorytm Gustafsona-Kessela, algorytm Fuzzy Maximum Likelihood Estimation (FMLE). Rozpatrywanie tych algorytmów jest poza zakresem naszych wykładów.

Algorytm k-means nazywa się czasem Hard C-Means (HCM) i realizowany on jest trochę inaczej, niż rozpatrywaliśmy dotychczas.

HCM: MACIERZE WEJŚCIOWE

$\mathbf{P}_{n \times K}$ - macierz przynależności wektorów danych $\mathbf{x}_i, i = 1, 2, \dots, n$ do grupy C_k :

➤ $p_{ik} = \{0, 1\};$

➤ dla każdego $i = 1, 2, \dots, n$: $\sum_{k=1}^K p_{ik} = 1;$

➤ dla każdego $k = 1, 2, \dots, K$ $0 < \sum_{i=1}^n p_{ik} < n.$

$\mathbf{C}_{K \times m}$ - macierz środków, której każdy k -ty wierz jest wektorem \mathbf{c}_k

ALGORYTM HCM

Krok 1. Wektory macierzy **C** inicjowane są losowo.

Krok 2. Dla każdego $i = 1, 2, \dots, n$, dla każdego $k = 1, 2, \dots, K$:

➤ $p_{ik} = 1$, jeśli dla każdego $l \neq k$ zachodzi $d(\mathbf{x}_i, \mathbf{c}_k) < d(\mathbf{x}_i, \mathbf{c}_l)$;

Jeśli dla pewnego wektora minimalna odległość jest realizowana przez więcej, niż jeden środek grupy, to należy wybrać losowo, bądź w inny ustalony sposób;

➤ $p_{ik} = 0$ w przeciwnym przypadku;

Krok 3. Dla każdego $k = 1, 2, \dots, K$:

$$\mathbf{c}_k = \frac{\sum_{i=1}^n p_{ik} \mathbf{x}_i}{\sum_{i=1}^n p_{ik}}.$$

Krok 4. Powtarzać kroki 2 i 3 dopóki grupowanie nie ustabilizuje się (macierze **P**, **C** przestaną się zmieniać).

WYBRANE METODY KLASYFIKACJI

Klasyfikacja - sformalizowane zadanie, w którym zbiór obiektów należy podzielić w pewny sposób na klasy.

Zadanie klasyfikacji polega na tym, że na podstawie danych początkowych (danych, dla których z góry wiadomo, do jakich klas one należą) zostaje opracowana metoda rozróżniania na grupy nowych (nie należących do danych początkowych) obiektów.

W większości metod klasyfikacji zbiór danych początkowych rozbija się na dwa zbiory: **uczący** i **testowy**. Zbiór uczący wykorzystany jest do nauczania (konstruowania) modelu. Zbiór testowy wykorzystany jest w celu sprawdzania wiarygodności zbudowanego modelu.

MIARY OCENY JAKOŚCI MODELI KLASYFIKACYJNYCH

Oceniając jakość modeli klasyfikacyjnych, możemy skorzystać z dwóch kategorii wskaźników:

- ***liczbowe wskaźniki jakości*** – statystyki wyrażające jakość klasyfikacji przy pomocy wymiernych wartości liczbowych.
- ***graficzne „wskaźniki”*** – graficzne przedstawienie jakości klasyfikacji, polegające na wizualizacji i odpowiednim zestawieniu różnych wskaźników liczbowych.

Metody graficzne ułatwiają ocenę i prezentację wyników klasyfikacji. Przykładami tego typu wskaźników są:

- macierz konfuzji;
- krzywa ROC;
- AUC itd.

KLASY BINARNE

W przypadku klasyfikacji dwuklasowej (binarnej) wyróżniają dwie klasy: klasa pozytywna (1) i klasa negatywna (0).

Do pozytywnej klasy dolicza się grupa obiektów, która nas interesuje w modelowaniu.

Klasa negatywna – to pozostałe obiekty. (W przypadku klasyfikacji wieloklasowej do klasy negatywnej należą wszystkie klasy łącznie, oprócz pozytywnej).

WYNIKI KLASYFIKACJI

- **True-Positive (TP - prawdziwie pozytywna)**: obserwacje poprawnie zaklasyfikowane do klasy pozytywnej;
- **True-Negative (TN - prawdziwie negatywna)**: obserwacje poprawnie zaklasyfikowane do klasy negatywnej;
- **False-Positive (FP - fałszywie pozytywna)**: obserwacje negatywne niepoprawnie zaklasyfikowane do klasy pozytywnej;
- **False-Negative (FN - fałszywie negatywna)**: obserwacje pozytywne niepoprawnie zaklasyfikowane do klasy negatywnej.

Wymienione wyniki klasyfikacji często zapisuje się w postaci tzw. **macierzy konfuzji** (macierzy błędów):

MACIERZ KONFUZJI

Stan faktyczny ↓ \predykcja →	0	1
0	<i>TN</i>	<i>FP</i>
1	<i>FN</i>	<i>TP</i>

Dla idealnego klasyfikatora (czyli wszystko poprawnie zakwalifikowaliśmy i nasz model się nie pomylił) mamy:

- ***FP*** = 0;
- ***FN*** = 0;
- ***TP*** = *liczba obserwacji należących do klasy pozytywnej*;
- ***TN*** = *liczba obserwacji należących do klasy negatywnej*.

Nieidealny klasyfikator:

- ***TP*** i ***FN*** tworzą klasę pozytywnych obserwacji: ***P*** = ***TP*** + ***FN***;
- ***TN*** i ***FP*** tworzą klasę negatywnych obserwacji: ***N*** = ***TN*** + ***FP***.

MIARY KLASYFIKACJI

➤ **dokładność:**

$$ACC = \frac{TP + TN}{TN + TP + FN + FP}$$

procent prawidłowo zaklasyfikowanych przypadków (obiektów, obserwacji);

➤ **błąd:**

$$ERR = 1 - ACC = \frac{FN + FP}{TN + TP + FN + FP}$$

procent nieprawidłowo zaklasyfikowanych przypadków;

➤ **czułość** (*True-Positive Rate*):

$$TPR = \frac{TP}{TP + FN}$$

miara wskazująca w jakim procencie klasa faktycznie pozytywna została pokryta przewidywaniem pozytywnym;

MIARY KLASYFIKACJI

➤ **specyficzność** (*True-Negative Rate*):

$$TNR = \frac{TN}{TN + FP}$$

miara wskazująca w jakim procencie klasa faktycznie negatywna została pokryta przewidywaniem negatywnym;

➤ **precyzja**:

$$\frac{TP}{TP + FP}$$

procent przewidywań faktycznie pozytywnych wśród tych, które klasyfikator zaklasyfikował jako pozytywne;

MIARY KLASYFIKACJI

➤ F_1 :

$$F_1 = \frac{2TP}{2TP + FP + FN}$$

wykorzystuje się przy porównywaniu jakości różnych klasyfikatorów – im wyższa wartość, tym lepsza jakość klasyfikatora;

➤ ***zbalansowana dokładność:***

$$\frac{1}{2}TPR + \frac{1}{2}TNR.$$

PRZYKŁAD

Do grupy 2000 osób skierowano komunikację marketingową zachęcającą do skorzystania z produktu. Spośród 2000 osób produkt zakupiło 600.

Grupę 2000 podzielono losowo na dwie równoliczne części, każda po 1000 osób (w tym w każdej po 300 klientów, którzy skorzystali z produktu). Pierwszej grupie przydzielono rolę danych uczących, zaś drugiej rolę danych testowych.

Wykorzystując dane uczące, dostępne charakterystyki klientów oraz informacje o fakcie zakupienia produktu, *przygotowano* (nauczono) klasyfikator umożliwiający przewidywanie czy dany klient skorzysta z produktu.

PRZYKŁAD

Oceny jakości klasyfikatora dokonano przy wykorzystaniu danych testowych. Wyniki oceny zaprezentowano w postaci macierzy konfuzji:

Stan faktyczny ↓ \predykcja →	0	1
0	<i>TN</i> 600	<i>FP</i> 100
1	<i>FN</i> 50	<i>TP</i> 250

Określić miary jakości klasyfikacji.

PRZYKŁAD

$TP + FN + TN + FP = 250 + 50 + 600 + 100 = 1000$ - liczba klientów (baza, na której dokonano oceny);

$P = TP + FN = 250 + 50 = 300$ - liczba klientów, którzy kupili produkt;

$N = TN + FP = 600 + 100 = 700$ - liczba klientów, którzy nie skorzystali z produktu;

$TP + TN = 250 + 600 = 850$ - liczba poprawnych klasyfikacji;

$FP + FN = 100 + 50 = 150$ - liczba błędnych klasyfikacji;

Dokładność: $ACC = \frac{TP + TN}{P + N} = \frac{850}{1000} = 85\%$ - procent poprawnych klasyfikacji;

PRZYKŁAD

Błąd klasyfikacji: $ERR = \frac{FP + FN}{P + N} = \frac{150}{1000} = 15\%$ - procent nieprawidłowych klasyfikacji;

Czułość: $TPR = \frac{TP}{TP + FN} = \frac{250}{250 + 50} = \frac{250}{300} = 0,83$ - miara wskazująca w jakim procencie liczba klientów, którzy skorzystali z produktu została zaklasyfikowana do tej klasy;

Specyficzność: $TNR = \frac{TN}{TN + FP} = \frac{600}{600 + 100} = 0,86$ - miara wskazująca w jakim procencie liczba klientów, którzy nie skorzystali z produktu została zaklasyfikowana do tej klasy;

PRZYKŁAD

Precyzja: $\frac{TP}{TP + FP} = \frac{250}{250 + 100} = 0,71$ - procent klientów, którzy skorzystali z produktu wśród tych wszystkich klientów, których klasyfikator zaklasyfikował do grupy osób skorzystawszych z produktu;

$$F_1 = \frac{2 \cdot TPR \cdot Precyzja}{TPR + Precyzja} = \frac{2 \cdot 0,83 \cdot 0,71}{0,83 + 0,71} = 0,77;$$

Zbalansowana dokładność:

$$\frac{1}{2} TPR + \frac{1}{2} TNR = \frac{1}{2} 0,83 + \frac{1}{2} 0,86 = 0,845.$$

MODEL PREDYKCYJNY A PUNKT ODCIĘCIA

Opisane miary określone są dla klasyfikatora binarnego, jednak w praktyce najczęściej stosuje się modele predykcyjne z **ciągłą zmienną odpowiedzi** (np. estymator prawdopodobieństwa skorzystania z produktu, gdzie wynikiem działania modelu jest wartość z przedziału $[0, 1]$ interpretowana właśnie jako wspomniane prawdopodobieństwo określane również **skłonnością**).

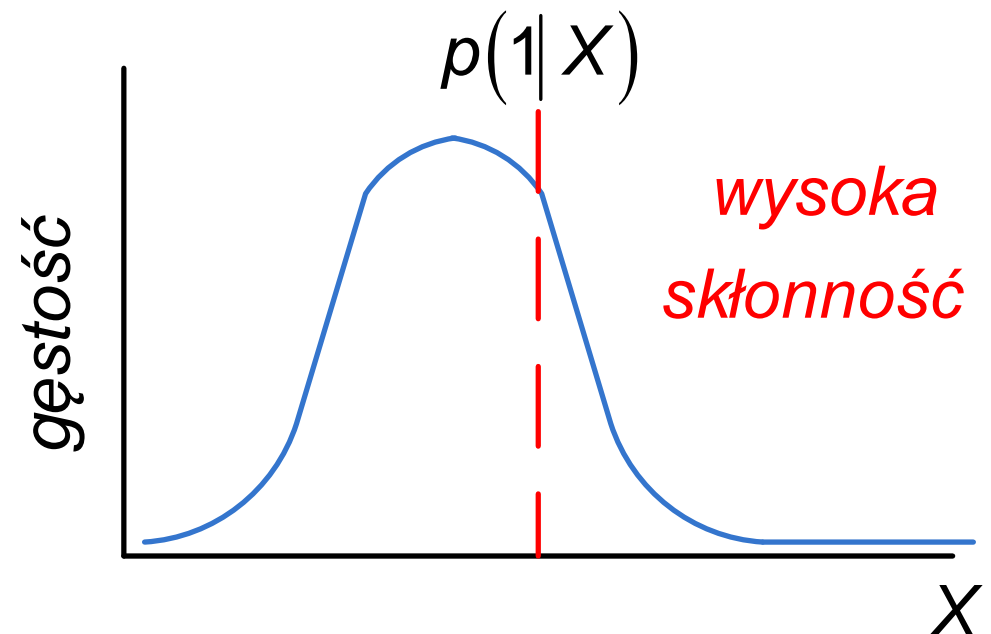
Rozpatrzmy przypadek dwóch grup. Często znalezienie reguły decyzyjnej sprowadza się do wybrania pewnej wartości zmiennej, która najlepiej dzieli badaną zbiorowość na dwie grupy: jedną, w której często występowało dane zdarzenie (wysoka skłonność) i druga, w której częstość występowania zdarzenia była mała (niewysoka skłonność).

MODEL PREDYKCYJNY A PUNKT ODCIĘCIA

Oznaczmy przez $p_0 \in [0,1]$ rozgraniczający segment wysokiej od segmentu niewysokiej skłonności, a $p(1|X)$ - prawdopodobieństwo warunkowe, że wystąpi pierwsza klasa dla obiektu X .

p_0 jest **punktem odcięcia** (**progiem**).

Jeśli szacowane prawdopodobieństwo $p(1|X) \geq p_0$, to obiekt X należy do pierwszej klasy, w przeciwnym przypadku do klasy drugiej.



KRZYWA ROC

Krzywa ROC (*Receiver Operating Characteristics*) jest graficzną reprezentacją efektywności modelu predykcyjnego poprzez wykreślenie charakterystyki jakościowej klasyfikatorów binarnych powstałych z modelu przy zastosowaniu wielu różnych punktów odcięcia.

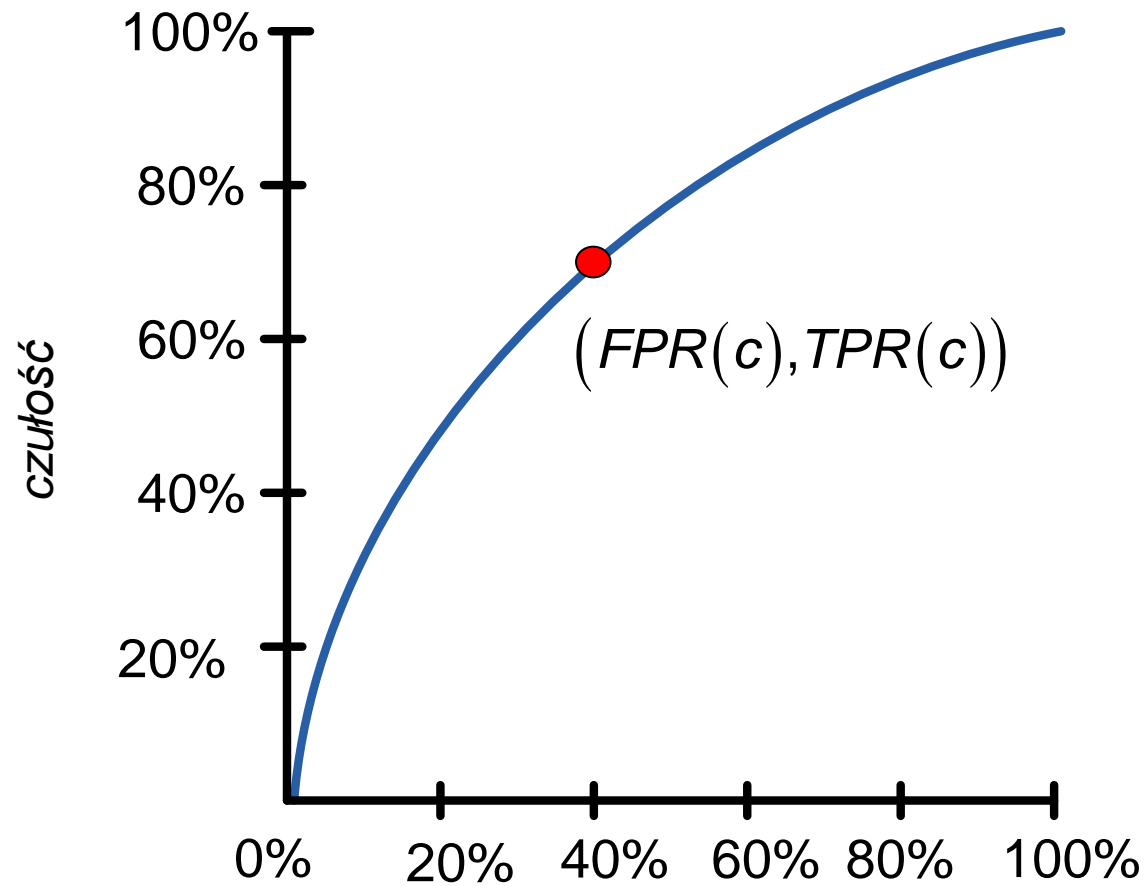
Każdy punkt krzywej ROC odpowiada innej macierzy konfuzji. Im więcej różnych punktów odcięcia zbadamy, tym więcej uzyskamy punktów na krzywej ROC.

Finalnie na wykres nanosimy czułość TPR (oś pionowa) oraz jeden minus specyficzność $FPR = 1 - TNR$ (*False-Positive Rate* - oś pozioma).

Punkt $(FPR(c), TPR(c))$ krzywej ROC reprezentuje jakość klasyfikatora binarnego otrzymanego z modelu predykcyjnego, przyjmując punkt odcięcia, równy c .

KRZYWA ROC

$$TPR = \text{True Positive Rate} = P(1|1) = TPR$$



$$FPR = \text{False Positive Rate} = P(1|0) = 1 - P(0|0) = 1 - TNR$$

1 - specyficzność

KRZYWA ROC

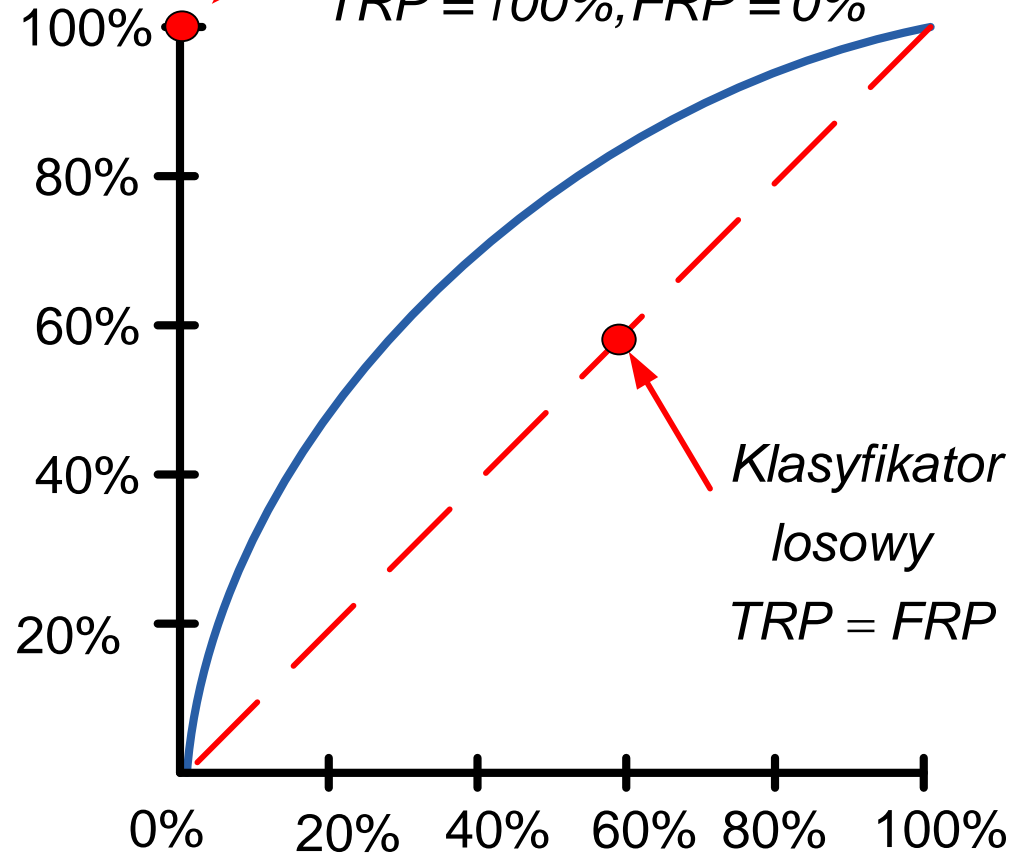
Klasyfikator

teoretycznie idealny

$TRP = 100\%, FRP = 0\%$

$TPR = \text{True Positive Rate} = P(1|1) = TPR$

czułość



Klasyfikator

losowy

$TRP = FRP$

$FPR = \text{False Positive Rate} = P(1|0) = 1 - P(0|0) = 1 - TNR$

$1 - \text{specyficzność}$

AUC

Bardzo popularnym podejściem jest także wyliczanie pola pod wykresem krzywej *ROC*, oznaczonego jako *AUC* (*Area Under Curve*) i traktowanego jako miarę dobroci i trafności danego modelu. Wartość wskaźnika *AUC* przyjmuje wartości z przedziału $[0,1]$; im większa, tym lepszy model.

LINIOWA ANALIZA DYSKRYMINACYJNA

Zadaniem analizy dyskryminacyjnej jest rozstrzyganie, które zmienne w najlepszy sposób dzielą dany zbiór przypadków na występujące w naturalny sposób klasy.

Zmienne dyskryminacyjne są to zmienne użyte do rozróżniania całej zbiorowości na klasy.

LINIOWA ANALIZA DYSKRYMINACYJNA

W praktyce analiza dyskryminacyjna to ogólny termin, odnoszący się do kilku powiązanych zadań statystycznych. W uproszczeniu te zadania możemy podzielić na:

1. Opis i interpretacja różnic międzygrupowych:

- czy możemy, dysponując zbiorem kilku zmiennych, wyodrębnić jedną grupę od drugiej;
- na ile dobrze zmienne dyskryminujące rozróżniają dane początkowe na grupy;
- które zmienne najlepiej dyskryminują zbiór danych.

2. Klasyfikacja obiektów, czyli określenie na podstawie cech uzyskanych z obserwacji lub z doświadczenia do której klasy należy nowy obiekt.

LINIOWA ANALIZA DYSKRYMINACYJNA

Inaczej mówiąc formalnie analiza dyskryminacyjna rozkłada się na dwa etapy: pierwszy - odzyskanie i opisanie różnic między istniejącymi klasami (grupami, podzbiorami) obiektów obserwowanych i wyodrębnienie zmiennych, mających istotny wpływ na proces rozbijania na klasy;

oraz drugi etap - opracowanie algorytmu klasyfikacji nowych obiektów, tzn. algorytmu odniesienia nowego obiektu do jednej z istniejących grup.

LINIOWA ANALIZA DYSKRYMINACYJNA

- jest to metoda poszukiwania liniowej kombinacji zmiennych X_1, X_2, \dots, X_m , która najlepiej rozdziela całą zbiorowość na dwie i więcej klasy.

Należy ta metoda do metod granicznych, czyli w podstawie tej metody leży założenie, że granicy między klasami można aproksymować za pomocą funkcji liniowych. Wtedy zadanie polega na odnalezieniu parametrów tych funkcji.

Ogólne równanie hiperpłaszczyzny dyskryminacyjnej ma postać:

$$D = a_1 X_1 + a_2 X_2 + \dots + a_m X_m$$

gdzie

a_1, a_2, \dots, a_m - współczynniki funkcji dyskryminacyjnej D . Mogą to być także funkcję, które zawierają wyraz wolny a_0 .

ZAŁOŻENIA

- do klasy muszą należeć co najmniej 2 obiekty, czyli $n_k \geq 2$, gdzie n_k - liczebność k -tej klasy;
- liczba zmiennych dyskryminacyjnych m może być dowolna pod warunkiem, że spełniony jest warunek $m < n - 2$, gdzie n jest liczbą obiektów (obserwacji);
- zmienne dyskryminacyjne X_1, X_2, \dots, X_m są liniowo niezależnymi wielowymiarowymi zmiennymi, które mają rozkład normalny;
- macierzy kowariancji klas są bliskie sobie.

W statystyce matematycznej udowodniono, że jeśli zbiór danych przedstawia próbę z wielowymiarowej populacji mającej rozkład normalny oraz wszystkie macierze kowariancji klas są równe, to analiza dyskryminacyjna jest optymalna, tzn. żaden inny sposób nie pozwala osiągnąć takiego małego prawdopodobieństwa błędu.

DANE DO ANALIZY DYSKRYMINACYJNEJ

Niech liczba zmiennych dyskryminacyjnych X_1, X_2, \dots, X_m wynosi m ; liczba obiektów w próbie jest równa n , liczba obiektów k -tej klasy jest równa n_k , $k = 1, 2, \dots, K$.

Wektor $\bar{\mathbf{x}}_k$ wartości średnich dla zmiennych X_1, X_2, \dots, X_m w każdej klasie jest równy:

$$\bar{\mathbf{x}}_k = \begin{pmatrix} \bar{x}_{1k} & \bar{x}_{2k} & \dots & \bar{x}_{mk} \end{pmatrix}.$$

DANE DO ANALIZY DYSKRYMINACYJNEJ

Macierz podobna do macierzy kowariancji w każdej klasie:

$$\mathbf{C}_k = \left(\mathbf{X}_k^* \right)^T \mathbf{X}_k^* \quad (1)$$

gdzie:

$$\mathbf{X}_k^* = \begin{pmatrix} x_{11k} - \bar{x}_{1k} & x_{12k} - \bar{x}_{2k} & \dots & x_{1mk} - \bar{x}_{mk} \\ x_{21k} - \bar{x}_{1k} & x_{22k} - \bar{x}_{2k} & \dots & x_{2mk} - \bar{x}_{mk} \\ \dots & \dots & \dots & \dots \\ x_{n1k} - \bar{x}_{1k} & x_{n2k} - \bar{x}_{2k} & \dots & x_{nmk} - \bar{x}_{mk} \end{pmatrix}$$

Etapy wyznaczenia funkcji dyskryminacyjnych rozróżniają w zależności od liczby klas.

ALGORYTM DLA $K=2$

Dla podziału całej zbiorowości na 2 klasy wystarczy znaleźć jedną zmienną dyskryminacyjną D .

1. Wyznaczyć wektory wartości średnich $\bar{\mathbf{x}}_k$ w każdej klasie.
2. Wyznaczyć macierz \mathbf{C}_k w każdej klasie.
3. Wyznaczyć wspólną macierz kowariancji:

$$\mathbf{C} = \frac{\mathbf{C}_1 + \mathbf{C}_2}{n_1 + n_2 - 2}$$

4. Wyznaczyć macierz \mathbf{C}^{-1} , odwrotna do wspólnej macierzy kowariancji.
5. Wyznaczyć wektor (kolumnę) współczynników funkcji dyskryminacyjnej:

$$\mathbf{a}^T = \mathbf{C}^{-1} \left(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 \right)^T$$

KLASYFIKACJA DLA $K=2$

W liniowej analizie dyskryminacyjnej przy podziale na dwie klasy funkcja dyskryminacyjna jest używana także i w celu klasyfikacji. W przypadku podziału na liczbę klas większą, niż 2, oprócz funkcji dyskryminacyjnych, wyznaczane są też funkcje klasyfikacyjne.

KLASYFIKACJA DLA K=2

1. Dla każdej klasy wyznaczyć wektor ocen wartości funkcji dyskryminacyjnej na podstawie zbioru początkowego danych:

$$\mathbf{d}_k = \mathbf{X}_k \mathbf{a}^T.$$

2. Wyznaczyć wartości średnie arytmetyczne \bar{d}_1, \bar{d}_2 dla wszystkich elementów wektora \mathbf{d}_k w każdej klasie.

3. Wyznaczyć stałą dyskryminacyjną c :

$$c = \frac{\bar{d}_1 + \bar{d}_2}{2}$$

4. W celu określenia do jakiej klasy należy nowy obiekt X_1, X_2, \dots, X_m , należy najpierw wyznaczyć dla niego wartość funkcji dyskryminacyjnej D i porównać ją ze stałą dyskryminacyjną c . Jeśli wyznaczona wartość D jest większa bądź równa stałej c , to nowy obiekt należy do klasy pierwszej, w przeciwnym przypadku - do klasy drugiej.

PRZYKŁAD

Na podstawie danych z tab. wyznaczyć funkcję dyskryminacyjną oraz klasę, do której należy nowy obiekt $X_1 = 0,2$, $X_2 = 0,75$.

	klasa1		klasa 2	
nr	X_1	X_2	X_1	X_2
1	0,15	1,91	0,48	0,88
2	0,34	1,68	0,41	0,62
3	0,09	1,89	0,62	1,09
4	0,21	2,30	0,50	1,32
5			1,20	0,68

1. Wektory wartości średnich $\bar{\mathbf{x}}_k$:

$$\bar{\mathbf{x}}_1 = \begin{pmatrix} \bar{x}_{11} & \bar{x}_{21} \end{pmatrix} = (0,1975 \quad 0,1945)$$

$$\bar{\mathbf{x}}_2 = \begin{pmatrix} \bar{x}_{12} & \bar{x}_{22} \end{pmatrix} = (0,6420 \quad 0,9180)$$

PRZYKŁAD

2. Klasa 1, $n_1 = 4$:

$$\mathbf{X}_1^* = \begin{pmatrix} 0,15 - 0,1975 & 1,91 - 1,954 \\ 0,34 - 0,1975 & 1,68 - 1,954 \\ 0,09 - 0,1975 & 1,89 - 1,954 \\ 0,21 - 0,1975 & 2,30 - 1,954 \end{pmatrix} = \begin{pmatrix} -0,0475 & -0,035 \\ 0,1425 & -0,265 \\ -0,1075 & -0,055 \\ 0,0125 & 0,355 \end{pmatrix}.$$

Macierz \mathbf{C}_1 jest równa:

$$\mathbf{C}_1 = \begin{pmatrix} -0,0475 & -0,035 \\ 0,1425 & -0,265 \\ -0,1075 & -0,055 \\ 0,0125 & 0,355 \end{pmatrix}^T \begin{pmatrix} -0,0475 & -0,035 \\ 0,1425 & -0,265 \\ -0,1075 & -0,055 \\ 0,0125 & 0,355 \end{pmatrix} = \begin{pmatrix} 0,0344 & -0,0256 \\ -0,0256 & 0,2004 \end{pmatrix}$$

PRZYKŁAD

Analogicznie macierz \mathbf{C}_2 :

$$\mathbf{C}_2 = \begin{pmatrix} 0,4120 & -0,1185 \\ -0,1185 & 0,3380 \end{pmatrix}$$

3. Wspólna macierz kowariancji:

$$\begin{aligned} \mathbf{C} &= \frac{1}{4 + 5 - 2} \left[\begin{pmatrix} 0,0344 & -0,0256 \\ -0,0256 & 0,2004 \end{pmatrix} + \begin{pmatrix} 0,4120 & -0,1185 \\ -0,1185 & 0,3380 \end{pmatrix} \right] = \\ &= \begin{pmatrix} 0,0638 & -0,0206 \\ -0,0206 & 0,0769 \end{pmatrix} \end{aligned}$$

4. Macierz, odwrotna do wspólnej macierzy kowariancji:

$$\mathbf{C}^{-1} = \begin{pmatrix} 17,1661 & 4,5938 \\ 4,5938 & 14,2266 \end{pmatrix}$$

PRZYKŁAD

5. Współczynniki funkcji dyskryminacyjnej:

$$\begin{aligned}\mathbf{a}^T &= \mathbf{C}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T = \begin{pmatrix} 17,1661 & 4,5938 \\ 4,5938 & 14,2266 \end{pmatrix} \cdot \begin{pmatrix} 0,1975 - 0,643 \\ 1,945 - 0,918 \end{pmatrix} = \\ &= \begin{pmatrix} 17,1661 & 4,5938 \\ 4,5938 & 14,2266 \end{pmatrix} \cdot \begin{pmatrix} -0,445 \\ 1,027 \end{pmatrix} = \begin{pmatrix} -2,912 \\ 12,569 \end{pmatrix}\end{aligned}$$

Funkcja dyskryminacyjna wygląda następująco:

$$D = -2,912 \cdot X_1 + 12,569 \cdot X_2$$

PRZYKŁAD

6. Dla każdej klasy wyznaczyć wektor ocen wartości funkcji dyskryminacyjnej na podstawie zbioru początkowego danych:

$$\mathbf{d}_1 = \begin{pmatrix} 0,15 & 1,91 \\ 0,34 & 1,68 \\ 0,09 & 1,89 \\ 0,21 & 2,30 \end{pmatrix} \cdot \begin{pmatrix} -2,912 \\ 12,569 \end{pmatrix} = \begin{pmatrix} 23,569 \\ 20,125 \\ 23,493 \\ 28,297 \end{pmatrix}$$

$$\mathbf{d}_2 = \begin{pmatrix} 0,48 & 0,88 \\ 0,41 & 0,62 \\ 0,62 & 1,09 \\ 0,50 & 1,32 \\ 1,20 & 0,68 \end{pmatrix} \cdot \begin{pmatrix} -2,912 \\ 12,569 \end{pmatrix} = \begin{pmatrix} 9,663 \\ 6,599 \\ 11,894 \\ 15,135 \\ 5,052 \end{pmatrix}$$

PRZYKŁAD

7. Wartości średnie arytmetyczne dla elementów wektorów \mathbf{d}_k :

$$\bar{d}_1 = \frac{1}{4} \cdot (23,569 + 20,125 + 23,493 + 28,297) = 23,871$$

$$\bar{d}_2 = \frac{1}{5} \cdot (9,633 + 6,599 + 11,894 + 15,135 + 5,052) = 9,668$$

8. Stała dyskryminacyjną:

$$c = \frac{23,871 + 9,668}{2} = 16,770$$

PRZYKŁAD

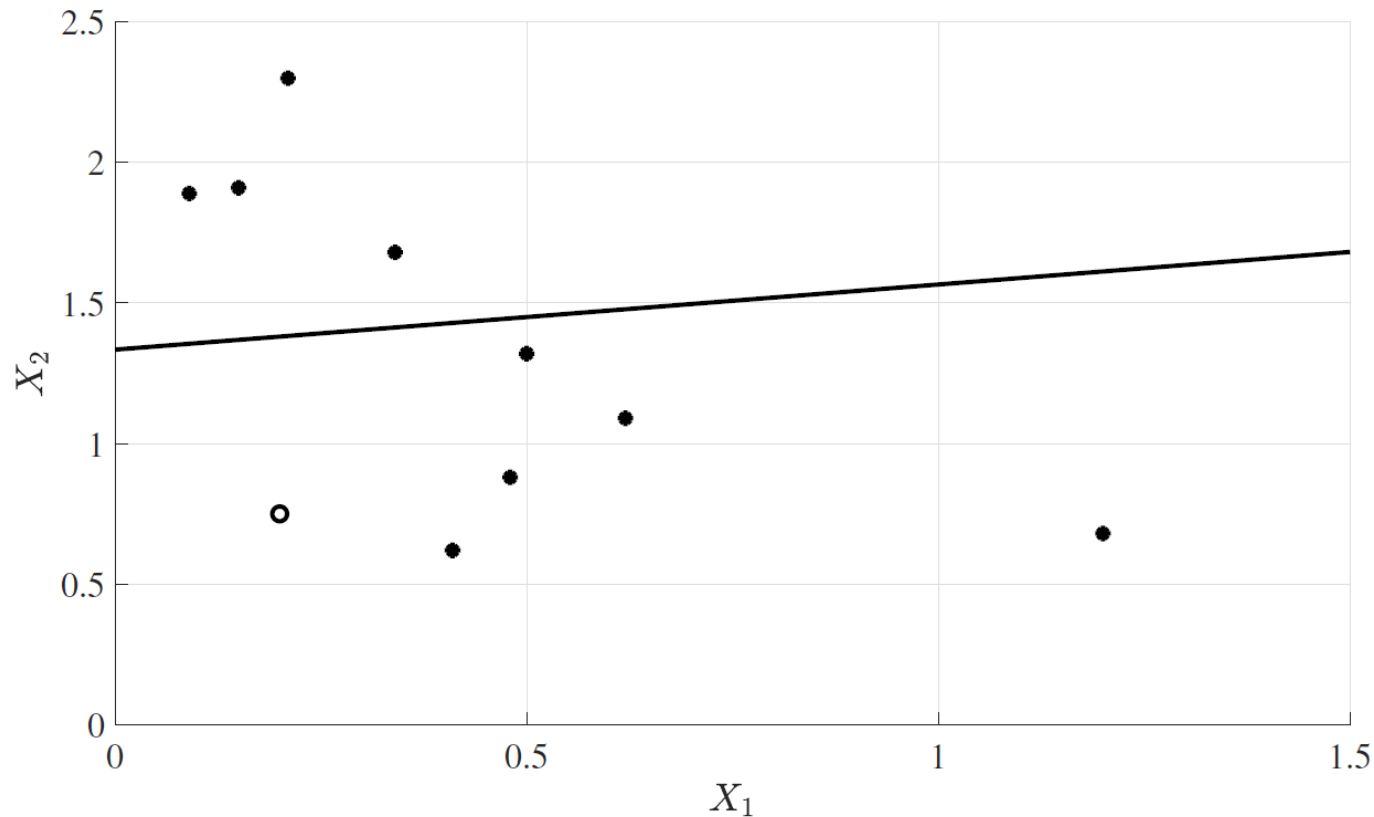
8. W celu określenia do jakiej klasy należy nowy obiekt $X_1 = 0,2$, $X_2 = 0,75$, należy wyznaczyć wartość funkcji dyskryminacyjnej:

$$D = -2,912 \cdot X_1 + 12,569 \cdot X_2 = -2,912 \cdot 0,2 + 12,569 \cdot 0,75 = 8,844$$

Ponieważ $D < c$, czyli $8,844 < 16,770$, to obiekt $X_1 = 0,2$, $X_2 = 0,75$ należy do klasy drugiej.

PRZYKŁAD

Na rys. przedstawiony jest rzut na płaszczyznę X_1, X_2 danych początkowych, prostej dyskryminacyjnej D , oraz obiektu do klasyfikacji.



$K > 2$

Dla podziału całej zbiorowości na liczbę klas $K > 2$ pod warunkiem, że $m \leq K$, wystarczy znaleźć co najwyżej $K - 1$ funkcji dyskryminacyjnych D_i .

ALGORYTM, $K > 2$

1. Wyznaczyć wektor $\bar{\mathbf{x}}_k$ wartości średnich w każdej klasie.
2. Wyznaczyć wektor $\bar{\mathbf{x}}$ wartości średnich dla każdej zmiennej we wszystkich klasach:

$$\bar{\mathbf{x}} = (\bar{x}_1, \bar{x}_2, \dots, \bar{x}_m)$$

3. Wyznaczyć macierz rozrzutu międzygrupowego \mathbf{M} :

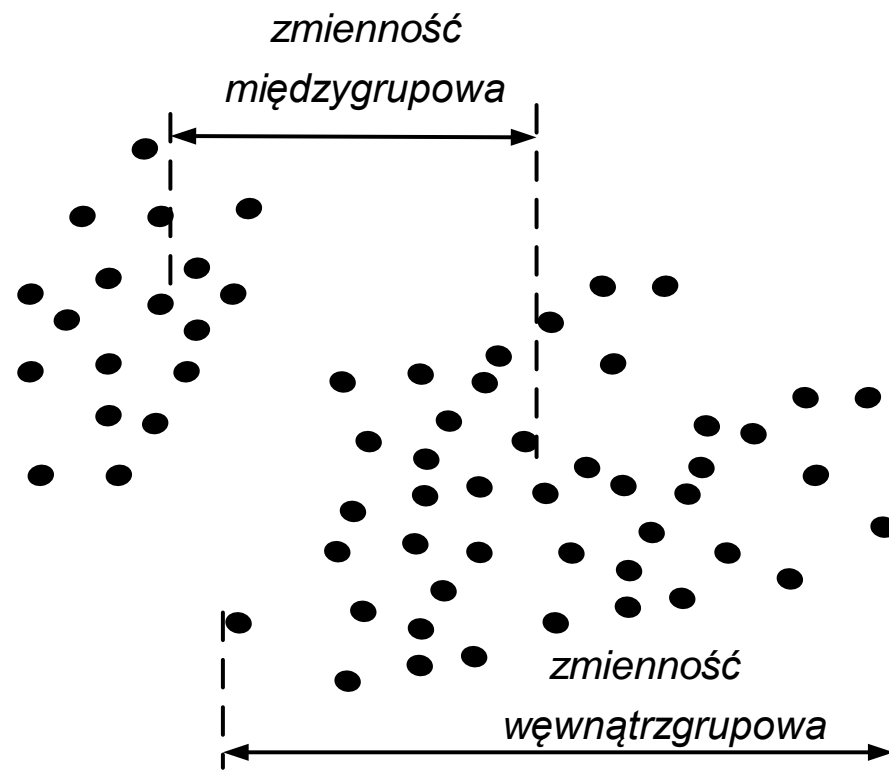
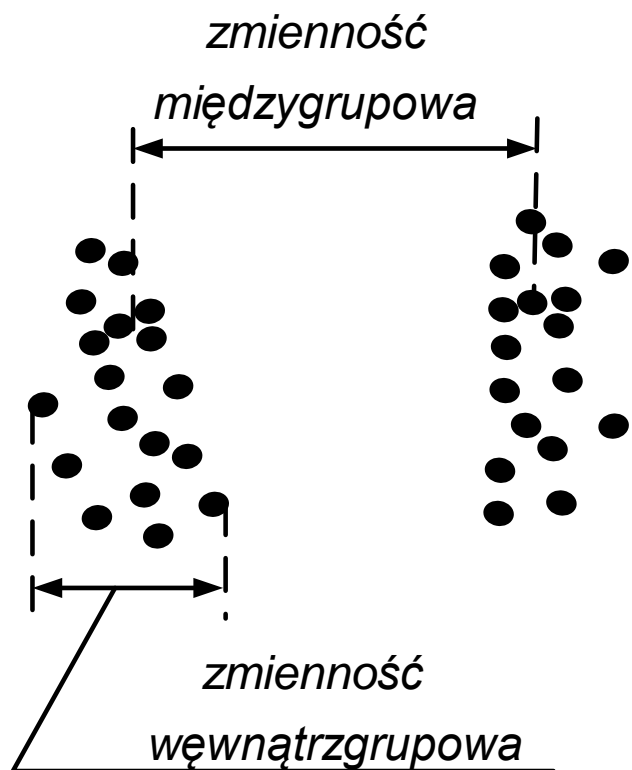
$$\mathbf{M} = \sum_{k=1}^K n_k \left[(\bar{\mathbf{x}}_k - \bar{\mathbf{x}})^T (\bar{\mathbf{x}}_k - \bar{\mathbf{x}}) \right]$$

4. Wyznaczyć macierz rozrzutu wewnątrzgrupowego \mathbf{W} :

$$\mathbf{W} = \sum_{k=1}^K \mathbf{C}_k$$

gdzie \mathbf{C}_k - macierze wyznaczone zgodnie ze wzorem (1).

ALGORYTM, $K > 2$



ALGORYTM, $K > 2$

5. Ponieważ Fisher udowodnił, że najlepsza dyskryminacja będzie osiągnięta w przypadku, gdy stosunek rozrzutu międzygrupowego do rozrzutu wewnątrzgrupowego osiąga maksimum i w praktyce warunek ten sprowadza się do znalezienia wektorów własnych macierzy $\mathbf{M}\mathbf{W}^{-1}$, to rozwiązujemy w tym celu równanie charakterystyczne:

$$|\mathbf{M} - \lambda \mathbf{W}| = 0,$$

gdzie $\{\lambda_1, \lambda_2, \dots\}$ - zbiór pierwiastków równania charakterystycznego (zbiór wartości własnych).

ALGORYTM, $K > 2$

6. Dla każdej wartości własnej λ_k wyznaczamy wektor (kolumnę) własny \mathbf{a}_k^T z równania:

$$(\mathbf{M} - \lambda_k \mathbf{W}) \mathbf{a}_k^T = 0,$$
$$\mathbf{a}_k = (a_{1k} \ a_{2k} \ \dots \ a_{mk}).$$

ALGORYTM, $K > 2$

7. Funkcję dyskryminacyjną są liniową kombinacją zmiennych dyskryminacyjnych:

$$\mathbf{D} = \mathbf{A}\mathbf{x}^T,$$

gdzie \mathbf{A} - macierz parametrów zmiennych dyskryminacyjnych:

$$\mathbf{A} = \begin{pmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \\ \dots \end{pmatrix}$$

a \mathbf{x} - wektor zmiennych X :

$$\mathbf{x} = (X_1 \ X_2 \ \dots \ X_m)$$

Inaczej, każda k -ta funkcja dyskryminacyjna wygląda następująco:

$$D_k = a_{1k}X_1 + a_{2k}X_2 + \dots + a_{mk}X_m.$$

LICZBA FUNKCJI DYSKRYMINACYJNYCH

Ogólnie, jeżeli liczba klas jest równa K , to liczba funkcji dyskryminacyjnych będzie co najmniej o jedną mniej, niż K ; jeżeli natomiast liczba zmiennych dyskryminacyjnych m jest mniejsza, niż liczba klas K , to liczba funkcji dyskryminacyjnych będzie co najmniej o jedną mniej, niż liczba zmiennych m .

LICZBA FUNKCJI DYSKRYMINACYJNYCH

Udział funkcji dyskryminacyjnych w rozróżnieniu na klasy zależy od wartości własnych, które można pogrupować w ciąg malejący $\lambda_1, \lambda_2, \dots$. Im większa wartość własna, tym lepiej rozdzielone są klasy, czyli jest lepszy podział. Oznacza to, że największą mocą dyskryminacji dysponuje pierwsza funkcja dyskryminacyjna (dla której λ_k jest największa). Z kolei druga funkcja dostarcza najlepszego rozdziału w tym "co pozostawiła" funkcja pierwsza itd.

Ponieważ mamy do czynienia z wylosowaną próbą, a nie z całą populacją, więc powstaje pytanie: "Jakie jest prawdopodobieństwo tego, że uzyskany w próbie pewien poziom zróżnicowania faktycznie w populacji występuje?", czyli czy wszystkie funkcje dyskryminacyjne można uznawać za statystycznie istotne.

LICZBA FUNKCJI DYSKRYMINACYJNYCH

Najpopularniejszy stosowany test nie testuje wprost funkcji dyskryminacyjnych, ale stopniowo testuje "pozostałości" po wprowadzeniu tej funkcji.

Przykładowo w pierwszym kroku oceniamy, czy warto wprowadzać pierwszą funkcję dyskryminacyjną. Z kolei w drugim kroku usuwamy informację dyskryminacyjną niesioną przez pierwszą funkcję, a następnie testujemy, czy dość dużo zróżnicowania pozostało, aby usprawiedliwić wprowadzenie drugiej funkcji dyskryminacyjnej itd. Taki sposób postępowania mówi nam, ile funkcji dyskryminacyjnych możemy traktować jako istotne.

Wykorzystuje się w tym celu **statystyka lambda Wilksa**. Jest to standardowa statystyka stosowana do wyznaczania istotności statystycznej mocy dyskryminacyjnej.

STATYSTYKA LAMBDA WILKSA

Aby przetestować, czy więcej, niż q funkcji mają nieistotną miarę dyskryminacyjną stosujemy statystykę postaci:

$$\Lambda = \prod_{k=q+1}^m \frac{1}{1 + \lambda_k}$$

Im bliżej ta wartość jest do zera, tym większe jest rozróżnienie na klasy, im dalej od zera, tym gorsze. Statystyka ta ma w przybliżeniu rozkład χ^2 , więc w praktyce często wykorzystuje się statystyka:

$$\chi^2 = - \left(n - \frac{m + K}{2} - 1 \right) \cdot \ln \Lambda$$

o $(m - K) \cdot (q - K - 1)$ stopniach swobody, gdzie
 $n = n_1 + n_2 + \dots + n_K$.

KLASYFIKACJA, $K > 2$

Są trzy główne metody **klasyfikacji** obiektów za pomocą analizy dyskryminacyjnej:

1. klasyfikacja liniowa - wyznacza się K funkcji liniowych i obiekt należy do takiej klasy k , wartość funkcji której jest największa dla danego obiektu;
2. metody, związane z odległością - obiekt należy do takiej klasy k , odległość do centrum której jest minimalna; najczęściej do wyznaczania odległości wykorzystuje się odległość Mahalanobisa;
3. metody probabilistyczne - obiekt należy do klasy k w przypadku, gdy odpowiednie prawdopodobieństwo a posteriori przynależności do tej klasy jest największe.

KLASYFIKACJA, $K > 2$

Rozpatrzmy metodę wyznaczenia liniowych funkcji klasyfikacyjnych, zaproponowaną przez angielskiego statystyka R. Fishera. Zasugerował on, że klasyfikacja powinna bazować na liniowej kombinacji zmiennych klasyfikacyjnych:

$$h_k = b_{k0} + b_{k1} \cdot X_1 + b_{k2} \cdot X_2 + \dots + b_{km} \cdot X_m$$

gdzie

h_k - funkcja klasyfikacyjna dla klasy k ;

b_{ki} - parametry funkcji klasyfikacyjnych.

KLASYFIKACJA, $K > 2$

Funkcji klasyfikacyjnych jest tyle, ile jest klas. Nie należy mylić funkcję klasyfikacyjnych h_k z funkcjami dyskryminacyjnymi D . Każdej klasie odpowiada своя funkcja klasyfikacyjna h_k . Obiekt należy do tej klasy k , dla której przyjmuje funkcja h_k największą wartość.

Współczynniki b_{ki} wyznaczone są za pomocą wzorów:

$$\mathbf{b}_k = (n - K) \cdot \bar{\mathbf{x}}_k \cdot \mathbf{W}^{-1},$$

gdzie $n = n_1 + n_2 + \dots + n_K$.

$$b_{k0} = -\frac{1}{2} \cdot \mathbf{b}_k \cdot \bar{\mathbf{x}}_k^T.$$

PRZYKŁAD

Na podstawie danych z tab. wyznaczyć funkcję klasyfikacyjną oraz klasę, do której należy nowy obiekt $X_1 = 25, X_2 = 30$.

	klasa 1		klasa 2		klasa 3	
1	X_1	X_2	X_1	X_2	X_1	X_2
2	10	10	20	10	32	50
3	9	8	15	12	40	62
4	8,7	9	30	20	43	57
5	6	7,5	22	15	45	70
6	7	9,8	31	27	35	60
7			35	30	31	54
8			40	18	41	58
9					42	61
10					38	67

PRZYKŁAD

Liczba pomiarów w każdej klasie wynosi $n_1 = 5$, $n_2 = 7$, $n_3 = 9$.

1. Wyznaczamy wektory $\bar{\mathbf{x}}_k$ wartości średnich dla zmiennych X_1, X_2 w każdej klasie:

$$\bar{\mathbf{x}}_1 = (8,14 \quad 8,86)$$

$$\bar{\mathbf{x}}_2 = (27,57 \quad 18,86)$$

$$\bar{\mathbf{x}}_3 = (38,56 \quad 59,89)$$

PRZYKŁAD

2. Wyznaczamy wektor $\bar{\mathbf{x}}$ wartości średnich dla każdej zmiennej X_1, X_2 we wszystkich klasach:

$$\bar{\mathbf{x}} = (27,65 \ 34,06).$$

Przykładowo dla zmiennej X_1 wartość średnia jest równa:

$$\bar{x}_1 = \frac{10 + \dots + 7 + 20 + \dots + 40 + 32 + \dots + 38}{5 + 7 + 9} = 27,65$$

PRZYKŁAD

3. Wyznaczamy macierz rozrzutu międzygrupowego **M**.
W tym celu obliczymy wartości $(\bar{\mathbf{x}}_k - \bar{\mathbf{x}})$ dla każdej klasy:

$$(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}) = (8,14 \quad 8,86) - (27,65 \quad 34,06) = (-19,51 \quad -25,20)$$

$$(\bar{\mathbf{x}}_2 - \bar{\mathbf{x}}) = (27,57 \quad 18,86) - (27,65 \quad 34,06) = (-0,08 \quad -15,20)$$

$$(\bar{\mathbf{x}}_3 - \bar{\mathbf{x}}) = (38,56 \quad 59,89) - (27,65 \quad 34,06) = (-10,90 \quad -25,83)$$

PRZYKŁAD

$$\begin{aligned} \mathbf{M} = & 5 \left[\begin{pmatrix} -19,51 \\ -25,20 \end{pmatrix} (-19,51 \quad -25,20) \right] + \\ & + 7 \left[\begin{pmatrix} -0,08 \\ -15,20 \end{pmatrix} (-0,08 \quad -15,20) \right] + \\ & + 9 \left[\begin{pmatrix} -10,90 \\ -25,83 \end{pmatrix} (-10,90 \quad -25,83) \right] = \begin{pmatrix} 2973,6 & 5001,7 \\ 5001,7 & 10797 \end{pmatrix} \end{aligned}$$

PRZYKŁAD

4. Wyznaczamy macierz rozrzutu wewnątrzgrupowego **W**.

W tym celu obliczmy wartości \mathbf{X}_k^* (analogicznie z poprzednim przykładem). Dla klasy 1:

$$\mathbf{X}_1^* = \begin{pmatrix} 1,86 & 1,14 \\ 0,86 & -0,86 \\ 0,56 & 0,14 \\ -2,14 & -1,36 \\ -1,14 & 0,94 \end{pmatrix}.$$

PRZYKŁAD

Dla klasy 1 obliczamy wartość \mathbf{C}_1 :

$$\mathbf{C}_1 = \begin{pmatrix} 1,86 & 1,14 \\ 0,86 & -0,86 \\ 0,56 & 0,14 \\ -2,14 & -1,36 \\ -1,14 & 0,94 \end{pmatrix}^T \begin{pmatrix} 1,86 & 1,14 \\ 0,86 & -0,86 \\ 0,56 & 0,14 \\ -2,14 & -1,36 \\ -1,14 & 0,94 \end{pmatrix} = \begin{pmatrix} 10,39 & 3,29 \\ 3,29 & 4,79 \end{pmatrix}$$

Analogicznie macierz \mathbf{C}_2 , \mathbf{C}_3 :

$$\mathbf{C}_2 = \begin{pmatrix} 473,71 & 277,57 \\ 277,57 & 332,85 \end{pmatrix}, \mathbf{C}_3 = \begin{pmatrix} 194,22 & 159,56 \\ 159,56 & 302,89 \end{pmatrix}$$

PRZYKŁAD

Wyznaczymy macierz **W**:

$$\begin{aligned}\mathbf{W} = \mathbf{C}_1 + \mathbf{C}_2 + \mathbf{C}_3 &= \begin{pmatrix} 10,39 & 3,29 \\ 3,29 & 4,79 \end{pmatrix} + \begin{pmatrix} 473,71 & 277,57 \\ 277,57 & 332,85 \end{pmatrix} + \\ &+ \begin{pmatrix} 194,22 & 159,56 \\ 159,56 & 302,89 \end{pmatrix} = \begin{pmatrix} 678,33 & 440,43 \\ 440,43 & 640,54 \end{pmatrix}\end{aligned}$$

5. Macierz odwrotna do macierzy rozrzutu wewnątrzgrupowego:

$$\mathbf{W}^{-1} = \begin{pmatrix} 0,0027 & -0,0018 \\ -0,0018 & 0,0028 \end{pmatrix}$$

PRZYKŁAD

6. Rozwiązując równanie charakterystyczne oraz układ równań, znajdujemy wartości własne $\lambda_1 = 1,60$, $\lambda_2 = 18,46$ oraz wektory własne \mathbf{a}_k^T .

czyli funkcje dyskryminacyjne są następujące:

$$D_1 = -0,95 X_1 - 0,40 X_2$$

$$D_2 = -3,11 X_1 - 0,92 X_2$$

PRZYKŁAD

7. Ponieważ $n - K = 21 - 3 = 18$, to współczynniki dla funkcji klasyfikacyjnych są równe:

$$\mathbf{b}_1 = 18 \cdot (8,14 \quad 8,86) \cdot \begin{pmatrix} 0,0027 & -0,0018 \\ -0,0018 & 0,0028 \end{pmatrix} = (0,10 \quad 0,18)$$

$$\mathbf{b}_2 = 18 \cdot (27,57 \quad 18,86) \cdot \begin{pmatrix} 0,0027 & -0,0018 \\ -0,0018 & 0,0028 \end{pmatrix} = (0,70 \quad 0,05)$$

$$\mathbf{b}_3 = 18 \cdot (38,56 \quad 59,89) \cdot \begin{pmatrix} 0,0027 & -0,0018 \\ -0,0018 & 0,0028 \end{pmatrix} = (-0,12 \quad 1,77)$$

PRZYKŁAD

$$b_{10} = -\frac{1}{2}(0,10 \quad 0,18) \begin{pmatrix} 8,14 \\ 8,86 \end{pmatrix} = -1,2$$

$$b_{20} = -\frac{1}{2}(0,70 \quad 0,05) \begin{pmatrix} 27,57 \\ 18,86 \end{pmatrix} = -10$$

$$b_{30} = -\frac{1}{2}(0,12 \quad 1,77) \begin{pmatrix} 38,56 \\ 59,89 \end{pmatrix} = -50,56$$

Funkcje klasyfikacyjne:

$$h_1 = -1,2 + 0,10 X_1 + 0,18 X_2$$

$$h_2 = -10 + 0,70 X_1 + 0,05 X_2$$

$$h_3 = -50,56 - 0,12 X_1 + 1,77 X_2$$

PRZYKŁAD

8. Sprawdzamy do jakiej klasy należy obiekt $X_1 = 25, X_2 = 30$:

$$h_1 = -1,2 + 0,10 \cdot 25 + 0,18 \cdot 30 = 6,70$$

$$h_2 = -10 + 0,70 \cdot 25 + 0,05 \cdot 30 = 8,85$$

$$h_3 = -50,56 - 0,12 \cdot 25 + 1,77 \cdot 30 = -0,60$$

Ponieważ $h_2 > h_1$ oraz $h_2 > h_3$, to obiekt $X_1 = 25, X_2 = 30$ należy do klasy drugiej.

KLASYFIKACJA, METODA 2

Drugą, bardziej intuicyjną procedurą klasyfikacji, jest procedura, która opiera się na miarach odległości indywidualnego przypadku od centroidy klasy. Punkt należy do klasy, dla której ta odległość jest najmniejsza.

Najczęściej stosowana jest uogólniona miara odległości Mahalanobisa. Dla każdego obiektu obliczamy odległości Mahalanobisa od każdej z centroid klasowych, a następnie klasyfikujemy przypadek do grupy, od której odległość Mahalanobisa jest najmniejsza.

KLASYFIKACJA, METODA 3

Trzecią procedurą jest wykorzystanie prawdopodobieństwa przynależności do grupy.

Prawdopodobieństwo *a priori* jest prawdopodobieństwo tego, że dany obiekt należy do danej grupy.

Dotychczas milcząco zakładaliśmy, że każda grupa traktowana jest jednakowo. W praktyce tak nie zawsze musi być.

Przykładowo założmy, że mamy dwie klasy i 90% całej populacji należy do klasy pierwszej. Wówczas większość przypadków będziemy klasyfikować do klasy pierwszej, a do klasy drugiej tylko wtedy, gdy mamy na to mocne dowody.

KLASYFIKACJA, METODA 3

W wielu sytuacjach powinniśmy użyć prawdopodobieństw a priori, jest to szczególnie ważne, gdy klasy nie są dobrze rozdzielone i wiele obiektów będzie blisko linii granicznych między klasami.

Znając prawdopodobieństwo a priori, przykładowo możemy szacować oczekiwane prawdopodobieństwo błędnej klasyfikacji, a tym samym klasyfikować obiekty do tych grup, gdzie to prawdopodobieństwo jest najmniejsze. W tym celu można posługiwać się twierdzeniem Bayes'a.

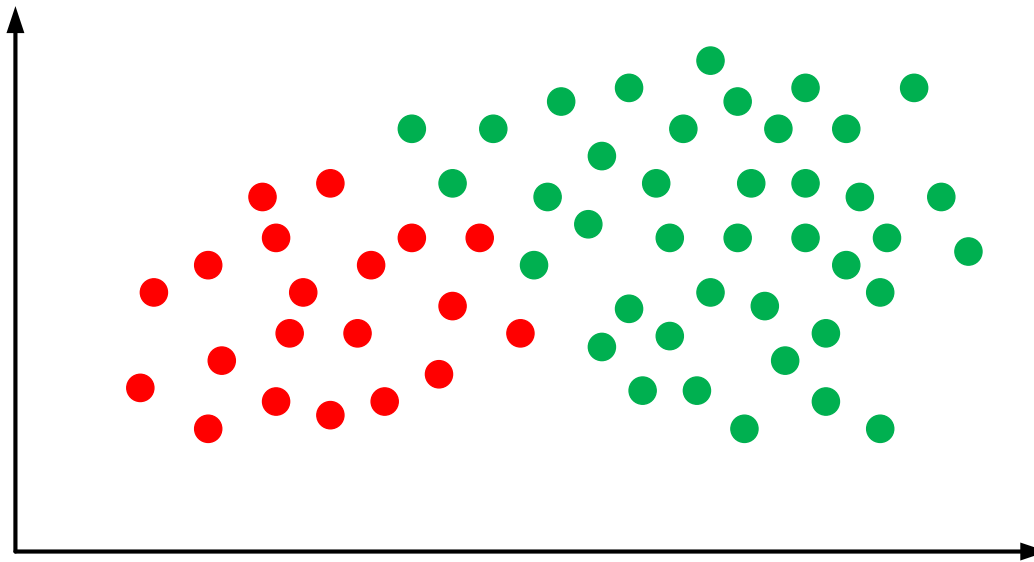
KLASYFIKACJA BAYES'A

Naiwny klasyfikator Bayesowski, bazujący na twierdzeniu Bayesa, nadaje się szczególnie do problemów o bardzo wielu wymiarach na wejściu.

Mimo prostoty metody, często działa ona lepiej od innych, bardzo skomplikowanych metod klasyfikujących.

KLASYFIKACJA BAYES'A

Dla ilustracji koncepcji Naiwnej metody Bayesa, rozpatrzmy przykład z rysunku. Jak widać, mamy tu obiekty zielone i czerwone. Naszym zadaniem będzie zaklasyfikowanie nowego obiektu, który może się tu pojawić.



KLASYFIKACJA BAYES'A

Ponieważ zielonych kólek jest dwa razy więcej niż czerwonych, rozsądnie będzie przyjąć, że nowy obiekt (którego jeszcze nie mamy) będzie miał dwa razy większe prawdopodobieństwo bycia zielonym niż czerwonym. W analizie Bayesowskiej, takie prawdopodobieństwa nazywane są prawdopodobieństwami *a priori*. Prawdopodobieństwa *a priori* wynikają z posiadanych, wcześniejszych (*a priori*) obserwacji.

W naszym przypadku, chodzi o procent zielonych względem czerwonych.

Prawdopodobieństwa *a priori* często służą do przewidywania klasy nieznanych przypadków, zanim one się pojawią. Możemy więc napisać: zob. kolejny slajd.

Założmy, że wszystkich obiektów jest 60, zielonych 40, a czerwonych 20

KLASYFIKACJA BAYES'A

$$\text{prawdopodobieństwo a priori zielonego} = \frac{\text{liczba obiektów zielonych}}{\text{całkowita liczba obiektów}}$$

$$\text{prawdopodobieństwo a priori czerwonego} = \frac{\text{liczba obiektów czerwonych}}{\text{całkowita liczba obiektów}}$$

Jako, że wszystkich obiektów jest 60, zielonych 40, a czerwonych 20, to prawdopodobieństwa a priori przynależności do odpowiednich klas:

$$\text{prawdopodobieństwo a priori zielonego} = \frac{40}{60}$$

$$\text{prawdopodobieństwo a priori czerwonego} = \frac{20}{60}$$

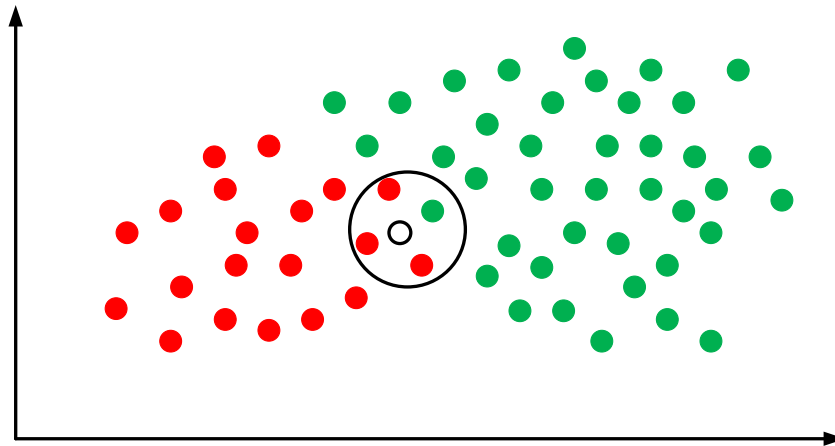
KLASYFIKACJA BAYES'A

Mając obliczone prawdopodobieństwa *a priori*, jesteśmy gotowi do zaklasyfikowania nowego obiektu (kółko *białe*).

Ponieważ obiekty są dobrze pogrupowane, sensownie będzie założyć, że im więcej jest *zielonych* (albo *czerwonych*) obiektów w pobliżu nowego obiektu, tym bardziej prawdopodobne jest, że obiekt ten ma kolor *zielony* (*czerwony*).

KLASYFIKACJA BAYES'A

Narysujmy więc okrąg wokół nowego obiektu, taki by obejmował wstępnie zadaną liczbę obiektów (niezależnie od ich klasy). Teraz będziemy mogli policzyć, ile wewnątrz okręgu jest *zielonych*, a ile *czzerwonych* kółek.



Skąd obliczymy wielkość, którą można nazwać szansą.

KLASYFIKACJA BAYES'A

$$\text{szansa, że } X \text{ będzie zielone} = \frac{\text{liczba zielonych w sąsiedztwie } X}{\text{całkowita liczba zielonych}}$$

$$\text{szansa, że } X \text{ będzie czerwone} = \frac{\text{liczba czerwonych w sąsiedztwie } X}{\text{całkowita liczba czerwonych}}$$

$$\text{szansa, że } X \text{ będzie zielone} = \frac{1}{40}$$

$$\text{szansa, że } X \text{ będzie czerwone} = \frac{3}{20}$$

Mimo, że prawdopodobieństwo *a priori* wskazuje, że X raczej będzie zielone (bo zielonych jest dwa razy więcej niż czerwonych), to szanse są odwrotne, ze względu na bliskość czerwonych.

KLASYFIKACJA BAYES'A

Końcowa klasyfikacja w analizie Bayesowskiej bazuje na obu informacjach, wg. reguły Bayesa (Thomas Bayes 1702-1761).

prawdopodobieństwo a posteriori zielonego =
= prawdopodobieństwo a priori zielonego · szansa, że X będzie zielone

prawdopodobieństwo a posteriori czerwonego =
= prawdopodobieństwo a priori czerwonego · szansa, że X będzie czerwone

$$\text{prawdopodobieństwo a posteriori zielonego} = \frac{4}{6} \cdot \frac{1}{40} = \frac{1}{60}$$

$$\text{prawdopodobieństwo a posteriori czerwonego} = \frac{2}{6} \cdot \frac{3}{40} = \frac{1}{40}$$

KLASYFIKACJA BAYES'A

W rezultacie klasyfikujemy X jako czerwone, gdyż większe jest prawdopodobieństwo *a posteriori* takiej właśnie przynależności.

Uwaga: Podane wyżej prawdopodobieństwa nie były normalizowane. Nie jest to konieczne przy klasyfikacji, gdyż czynnik normalizacyjny byłby ten sam dla wszystkich klas.

KLASYFIKACJA BAYES'A

Podany przykład miał charakter intuicyjny, jego celem było ułatwienie zrozumienia *naiwnej* metody klasyfikacyjnej Bayesa.

Naiwna metoda Bayesa jest w stanie analizować dowolną liczbę zmiennych niezależnych, ciągłych i skategoryzowanych.

Dla zmiennych $X = \{x_1, x_2, \dots, x_K\}$ prawdopodobieństwa *a posteriori* przypadku c_j , $j = 1, 2, \dots, d$ wg reguły Bayesa:

$$p(c_j | x_1, x_2, \dots, x_K) = p(x_1, x_2, \dots, x_K | c_j) \cdot p(c_j),$$

$p(x_1, x_2, \dots, x_K | c_j)$ prawdopodobieństwem *a priori* przynależności do klasy; $p(c_j)$ - szansa.

KLASYFIKACJA BAYES'A

Ponieważ naiwny klasyfikator Bayes'a zakłada, że warunkowe prawdopodobieństwa dla zmiennych niezależnych są wzajemnie, statystycznie niezależne, możemy *szansę* zapisać jako iloczyn

$$p(X|c_j) = \prod_{k=1}^K p(x_k|c_j)$$

a prawdopodobieństwo *a posteriori* w postaci

$$p(c_j|X) = p(c_j) \cdot \prod_{k=1}^K p(x_k|c_j).$$

Za pomocą reguły Bayes'a, nowy przypadek X etykietujemy nazwą klasy c_j , która ma największe prawdopodobieństwo *a posteriori*.

KLASYFIKACJA BAYES'A

Założenie o wzajemnej niezależności predyktorów (zmiennych niezależnych) nie zawsze jest całkiem ścisłe. Jednak upraszcza ono klasyfikację w zupełnie zasadniczy sposób, jako że można wtedy warunkowe gęstości prawdopodobieństwa dla klas $p(x_k | c_j)$ obliczyć osobno dla każdej zmiennej, co redukuje zadanie wielowymiarowe do szeregu jednowymiarowych. Przede wszystkim jednak, założenie to nie wydaje się mieć bardzo wielkiego wpływu na prawdopodobieństwa *a posteriori*, szczególnie w pobliżu granic decyzyjnych, czyli w sumie nie ma wielkiego wpływu na klasyfikację.

KLASYFIKACJA BAYES'A

Zalety:

- prosta postać analityczna;
- wysoka prędkość działania modeli na podstawie reguły Bayes'a;
- klasyfikator Bayes'a może być zbudowany na bazie próbki z brakującymi wartościami;
- wymagania do rozmiarów próbki zmniejszają się;

Wady:

model jest dobry przy spełnieniu założenia o niezależności; inaczej prawdopodobieństwa już nie są dokładne (nawet suma prawdopodobieństw może nie być równa 1, w związku z czym należy normalizować wyniki); jednak w praktyce zmniejszenie precyzji jest nieznaczne i nawet w przypadku dużej zależności między zmiennymi wynik działania klasyfikatora jest poprawny.

ALGORYTM K-NAJBLIŻSZYCH SĄSIADÓW (K-NN)

Metoda k-NN jest metodą automatycznej klasyfikacji obiektów.

Główną zasadą tej metody k-NN jest to, że obiekt będzie należał do tej klasy, która jest najbardziej rozpowszechniona wśród k sąsiedzi tego obiektu. Sąsiedzi są brane ze zbioru obiektów, klasy których już są znane.

Algorytm

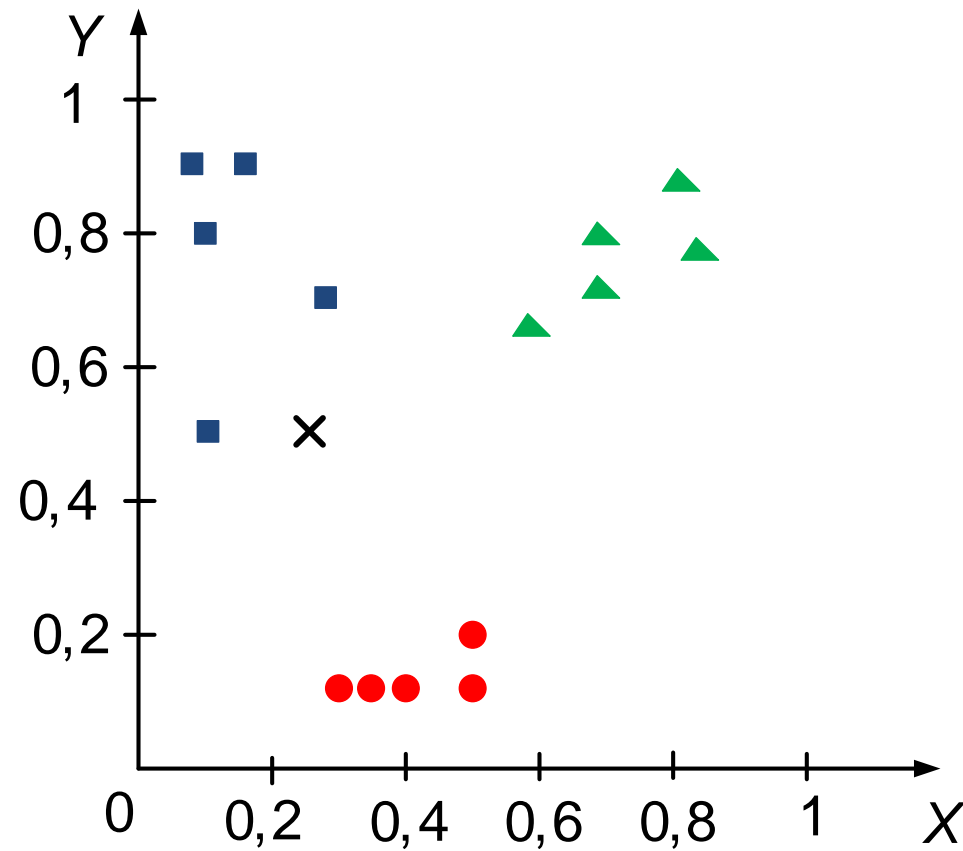
- obliczyć odległość do każdego z obiektów próbki uczącej;
- wybrać k obiektów próbki uczącej, odległość do których jest minimalna;
- klasą obiektu klasyfikującego jest klasa, która najczęściej spotykana jest wśród k sąsiedzi;
- jeśli wystąpił impas (dwie klasy miały taką samą liczbę głosów) rozwiązać problem losowo.

PRZYKŁAD

Mamy 3 obiekty, każdy z których ma 2 współrzędne. Także mamy punkt, współrzędne którego są znane. Należy określić do jakiego obiektu należy ten punkt.

Obiekt 1		Obiekt 2		Obiekt 3		Punkt nieokreślony	
X	Y	X	Y	X	Y	X	Y
0,3	0,1	0,7	0,8	0,1	0,8	0,25	0,5
0,5	0,2	0,9	0,75	0,2	0,9		
0,5	0,1	0,85	0,9	0,3	0,7		
0,4	0,1	0,55	0,65	0,1	0,9		
0,35	0,1	0,65	0,75	0,1	0,5		

PRZYKŁAD



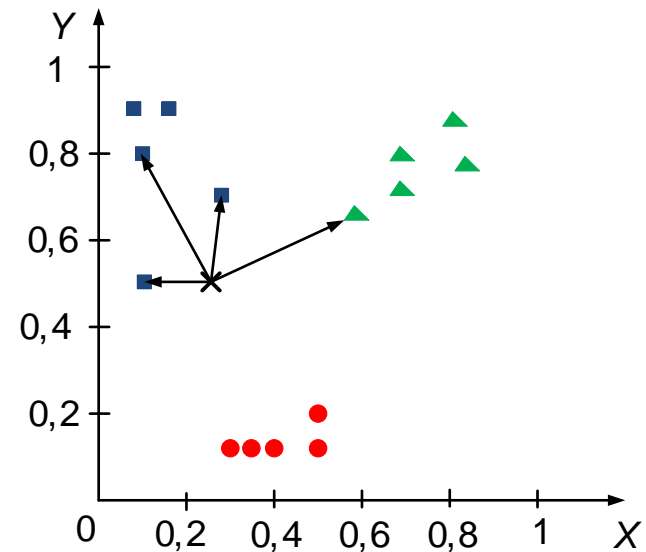
Np., dla Obiektu 1: $d = \sqrt{(0,25 - 0,3)^2 + (0,5 - 0,1)^2} = 0,4031$

PRZYKŁAD

Obiekt 1			Obiekt 2			Obiekt 3		
X	Y	d	X	Y	d	X	Y	d
0,3	0,1	0,4031	0,7	0,8	0,5408	0,1	0,8	0,3354
0,5	0,2	0,3905	0,9	0,75	0,6964	0,2	0,9	0,4031
0,5	0,1	0,4717	0,85	0,9	0,7210	0,3	0,7	0,2062
0,4	0,1	0,4272	0,55	0,65	0,3354	0,1	0,9	0,4272
0,35	0,1	0,4123	0,65	0,75	0,4717	0,1	0,5	0,1500

Najbliższe sąsiedzi

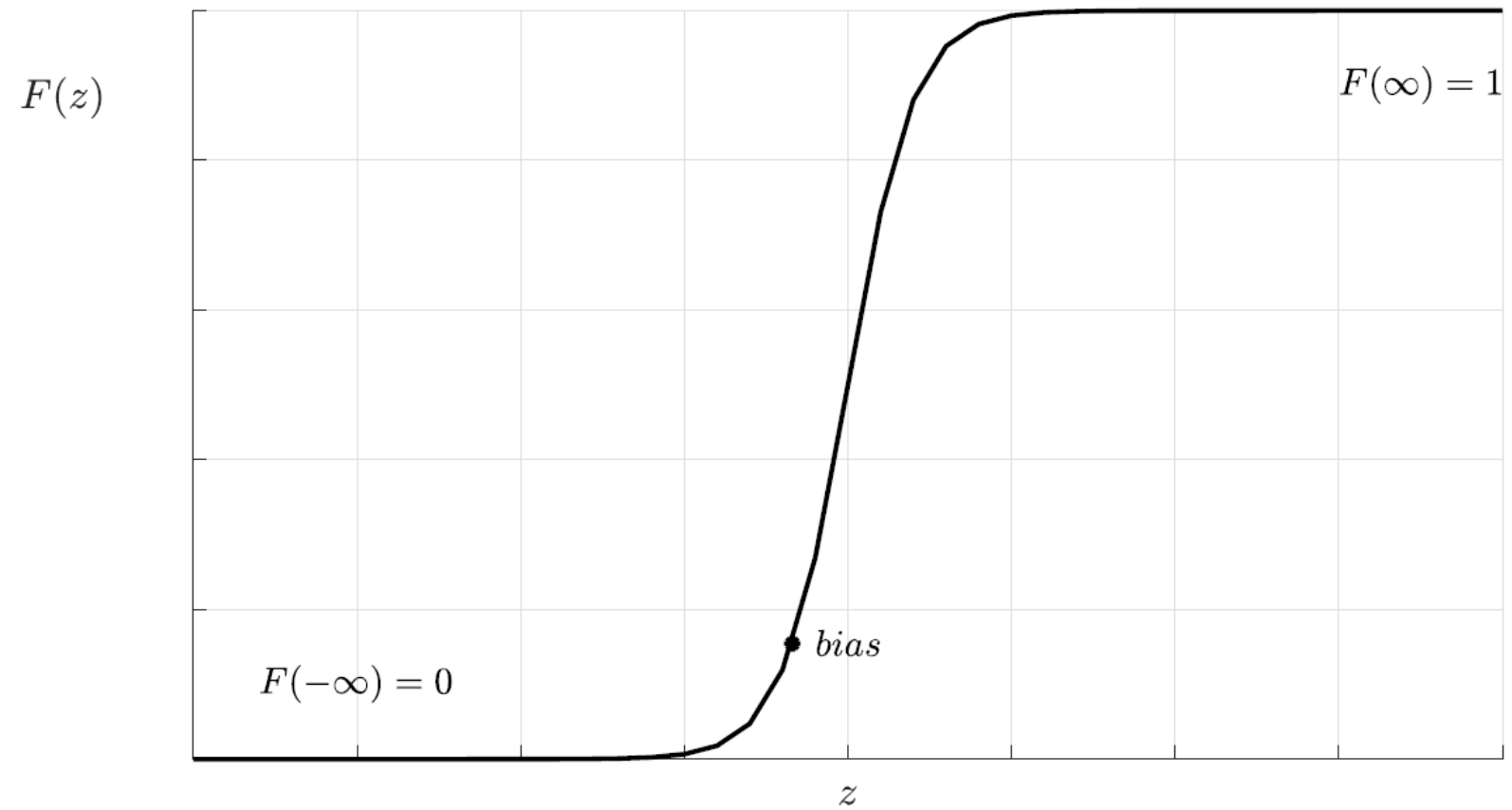
0,1	0,5	0,1500	Obiekt 3
0,3	0,7	0,2062	Obiekt 3
0,1	0,8	0,3354	Obiekt 3
0,55	0,65	0,3354	Obiekt 2



REGRESJA LOGISTYCZNA

Często w różnych naukach mamy do czynienia z sytuacją, gdy zmienna zależna jest typu dychotomicznego, czyli jest zmienną która przyjmuje tylko dwie wartości. Nie możemy w takich przypadkach wykorzystywać regresję liniową wieloparametryczną.

REGRESJA LOGISTYCZNA



REGRESJA LOGISTYCZNA

Funkcja $F(z)$ przyjmuje wartości pomiędzy 0 a 1.

Wartość funkcji $F(z)$ zmierza się do zera, gdy z dąży do $-\infty$, oraz zmierza się do jedynki, gdy z dąży do $+\infty$.

Kształt funkcji przypomina rozciągniętą literę S. Pokazuje on, że zmiany funkcji są minimalne, jeśli wartości zmiennych są mniejsze od pewnej wartości progowej *bias*.

Gdy ją przekroczą, wówczas wartość funkcji zaczyna gwałtownie rosnać do 1; prawdopodobieństwo utrzymuje się na wyjątkowo wysokim poziomie - blisko 1.

Za pomocą funkcji $F(z)$ można opisywać wartości prawdopodobieństwa.

REGRESJA LOGISTYCZNA

Doskonałym narzędziem do podobnego typu zagadnień jest ***regresja logistyczna***.

Zaletą tej regresji jest to, że analiza i interpretacja wyników jest podobna do analizy regresji liniowej. Model regresji logistycznej opiera się na dystrybuancie standaryzowanej funkcji logistycznej postaci:

$$F(z) = \frac{e^z}{1 + e^z} = \frac{1}{1 + e^{-z}}$$

z gęstością prawdopodobieństwa:

$$f(z) = \frac{e^z}{(1 + e^z)^2}$$

REGRESJA LOGISTYCZNA

Jeżeli zmienna zależna Y jest zmienną dychotomiczną (binarną), a zmienne niezależne X_1, X_2, \dots, X_m mogą być ilościowe i jakościowe, to zaznaczmy prawdopodobieństwo, że zmienna Y przyjmie wartość 1 dla wartości zmiennych niezależnych:

$$P(\mathbf{X}) = P(Y = 1 | X_1, X_2, \dots, X_m) = \frac{e^z}{1 + e^z} = \frac{1}{1 + e^{-z}},$$

gdzie $z = b_0 + \sum_{j=1}^m b_j X_j$, $b_j, j = 0, 1, \dots, m$ - współczynniki regresji.

Musimy oszacować współczynniki regresji b_j .

REGRESJA LOGISTYCZNA

W tym celu stosuje się metoda największej wiarygodności, która polega na maksymalizacji funkcji wiarygodności L , przy założeniu, że wszystkie obserwacje są od siebie niezależne.

Funkcja wiarygodności L jest iloczynem prawdopodobieństw pojawienia się poszczególnych obserwacji z próby przy danych parametrach modelu:

$$L = \prod_{i=1}^n P(y_i | b_0, b_1, \dots, b_m)$$

gdzie

n - liczba pomiarów (obserwacji); zalecane jest, aby $n > 10(m + 1)$

REGRESJA LOGISTYCZNA

Ponieważ $P(y_i = 0) = 1 - P(y_i = 1)$, to $P(y_i | b_0, b_1, \dots, b_m)$ oznaczmy w skrócie:

$$\begin{aligned} & p(y_i), \text{ gdy } y_i = 1 \\ & (1 - p(y_i)), \text{ gdy } y_i = 0, \end{aligned}$$

wtedy:

$$P(y_i | b_0, b_1, \dots, b_m) = p(y_i)^{y_i} \cdot (1 - p(y_i))^{1-y_i}$$

Wzór ten jest wygodny dla wykorzystania, ponieważ $y_i = 1$ zostaje tylko prawdopodobieństwo zajścia zdarzenia $p(y_i)$, gdy $y_i = 0$, to zostaje tylko prawdopodobieństwo niezajścia $(1 - p(y_i))$.

SZACOWANIE PARAMETRÓW

Zasadniczą ideą metody największej wiarygodności jest to, aby za oceny szacowanych parametrów $\mathbf{b} = (b_0 \ b_1 \ \dots \ b_m)^T$ brać wartości, dla których wiarygodność jest największa.

Bierze się to z założenia, że w wyniku wylosowania próby powinno zrealizować się zdarzenie o największym prawdopodobieństwie.

Ponieważ funkcja L osiąga maksimum w tych samych punktach, co jej logarytm, w praktyce wyznacza się maksimum funkcji $l = \ln L$

SZACOWANIE PARAMETRÓW

$$\begin{aligned} l = \ln L &= \ln \left(\prod_{i=1}^n P(y_i | b_0, b_1, \dots, b_m) \right) = \\ &= \ln \left(\prod_{i=1}^n \left(p(y_i)^{y_i} \cdot (1 - p(y_i))^{1-y_i} \right) \right) = \\ &= \ln \left(\left(\prod_{y_i=1} p(y_i)^{y_i} \right) \cdot \left(\prod_{y_i=0} (1 - p(y_i))^{1-y_i} \right) \right) = \\ &= \sum_{i=1}^n \left[y_i \ln p(y_i) + (1 - y_i) (1 - \ln p(y_i)) \right] \end{aligned}$$

SZACOWANIE PARAMETRÓW

Maksimum funkcji $I \rightarrow \max$ znajduje się metodami rachunku różniczkowego, przy rozwiązaniu układu równań:

$$\frac{\partial I}{\partial b_j} = 0$$

Najczęściej przy wyznaczaniu parametrów regresji logistycznej korzystają z algorytmu Newtona-Raphsona.

SZACOWANIE PARAMETRÓW

$$(\mathbf{b}^T)^{(t)} = (\mathbf{b}^T)^{(t-1)} - (\mathbf{H}^{-1} \nabla l),$$

\mathbf{H} - (Hessian) macierz drugich pochodnych po l ;

∇l - gradient funkcji l (macierz pierwszych pochodnych po l).

$$\nabla l = \frac{\partial l}{\partial \mathbf{b}_j} = \sum_{i=1}^n (p(y_i) - y_i) \mathbf{x}_i = \mathbf{X}(\mathbf{p} - \mathbf{y})^T$$

$$\mathbf{H} = \frac{\partial^2 l}{\partial \mathbf{b}_j^2} = \sum_{i=1}^n p(y_i)(1 - p(y_i)) \mathbf{x}_i^T \mathbf{x}_i = \mathbf{X}^T \mathbf{W} \mathbf{X}$$

gdzie \mathbf{W} - macierz diagonalna, na głównej przekątnej której znajdują się wartości $w_{ii} = p(y_i)(1 - p(y_i))$.

SZACOWANIE PARAMETRÓW

$$\begin{aligned}(\mathbf{b}^T)^{(t)} &= (\mathbf{b}^T)^{(t-1)} - (\mathbf{H}^{-1} \nabla l) = (\mathbf{b}^T)^{(t-1)} - (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T (\mathbf{p} - \mathbf{y})^T = \\ &= (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{u}^T\end{aligned}$$

gdzie:

$$\mathbf{u}^T = \mathbf{X} (\mathbf{b}^T)^{(t-1)} - \mathbf{W}^{-1} (\mathbf{p} - \mathbf{y})^T.$$

ALGORYTM

1. Wybór punktu startowego. Przykładowo punkt startowy można wybrać następująco:

$$\mathbf{b}^{(0)} = (b_0 \ 0 \ 0 \ \dots \ 0)^T, \text{ gdzie } b_0 = \ln \frac{\bar{y}}{1 - \bar{y}}.$$

2. Następujące kroki wykonują się iteracyjnie, zaczynając od liczby iteracji $t = 1$ i do momentu spełnienia warunku stopu:

ALGORYTM

- szukamy wektor $\mathbf{z} = (z_1 \ z_2 \ \dots \ z_n)^T$: $\mathbf{z} = \mathbf{X}\mathbf{b}^{(t-1)}$;
- szukamy wektor $\mathbf{p} = (p_1 \ p_2 \ \dots \ p_n)^T$: $p_i = \frac{1}{1 + e^{-z_i}}$;
- szukamy wektor \mathbf{u} : $u_i = z_i + \frac{y_i - p_i}{w_i}$, gdzie $w_i = p_i(1 - p_i)$ -

wektor wag wartości zmiennej zależnej;

- tworzymy macierz $\mathbf{W}_{n \times n}$, u której na głównej przekątnej są elementy w_i ;

- obliczamy nowe parametry regresji logistycznej
 $\mathbf{b}^{(t)} = (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{u}$;

- w przypadku spełnienia warunku stopu, kończymy obliczenia, inaczej wykonujemy kolejną iterację algorytmu dla $t = t + 1$, zaczynając od wyznaczenia \mathbf{z} .

ALGORYTM

3. Jako kryterium zatrzymania algorytmu możemy przyjąć sprawdzenie, czy względne poprawki dla wszystkich parametrów są mniejsze niż pewna wartość dopuszczalna.

LOGIT

Rozpatrzmy model z wyznaczonymi współczynnikami regresji **b**, w modelu tym zachodzą następujące zależności:

$$P(\mathbf{X}) = \frac{e^z}{1 + e^z},$$

$$1 - P(\mathbf{X}) = \frac{1}{1 + e^z},$$

$$\frac{P(\mathbf{X})}{1 - P(\mathbf{X})} = \frac{e^z}{1 + e^z} : \frac{1}{1 + e^z} = \frac{e^z}{1 + e^z} \cdot \frac{1 + e^z}{1} = e^z$$

$$\ln\left(\frac{P(\mathbf{X})}{1 - P(\mathbf{X})}\right) = \ln(e^z) = z = b_0 + \sum_{j=1}^m b_j \cdot X_j.$$

LOGIT

Model:

$$\ln\left(\frac{P(\mathbf{X})}{1-P(\mathbf{X})}\right) = b_0 + \sum_{j=1}^m b_j \cdot X_j$$

jest liniowy względem parametrów b_i i zmiennych X .

Wartość:

$$\ln\left(\frac{P(\mathbf{X})}{1-P(\mathbf{X})}\right)$$

nosi nazwę ***logit***.

LOGIT

Ponieważ $P(\mathbf{X})$ jest prawdopodobieństwem, że $Y = 1$, to logit jest logarytmem ilorazu prawdopodobieństw przyjęcia oraz nieprzyjęcia wartości 1 przez zmienną Y .

Jeśli prawdopodobieństwa są jednakowe $P(\mathbf{X}) = 0,5$; $1 - P(\mathbf{X}) = 0,5$, to logit jest równy zero:

$$\ln\left(\frac{P(\mathbf{X})}{1 - P(\mathbf{X})}\right) = \ln\frac{0,5}{0,5} = \ln 1 = 0$$

czyli dla $P(\mathbf{X}) > 0,5$ *logit* jest dodatni, dla $P(\mathbf{X}) < 0,5$ - ujemny.

SZANSA

W modelach logistycznych wartość $\frac{P(\mathbf{X})}{1 - P(\mathbf{X})}$ nazywa się **szansą**.

Jest to stosunek prawdopodobieństwa, że jakieś zdarzenie wystąpi do prawdopodobieństwa, że ten przypadek nie pojawi się.

Albo inaczej, przykładowo jeśli prawdopodobieństwo zdarzenia $P(\mathbf{X})$ jest równe 0,8, to szansa będzie równa $\frac{0,8}{1 - 0,8} = 4$.

Oznacza to, że prawdopodobieństwo pojawienia się przypadku jest cztery razy większe niż prawdopodobieństwa nie pojawienia się tego przypadku.

SZANSA

Uwaga: Szansa nie jest prawdopodobieństwem!

- Szansa nie jest symetryczna
- Przyjmuje wartości od 0 do $+\infty$ (a nie od 0 do 1)

ILORAZ SZANS

Dla dwóch grup porównywalnych A i B definiuje się stosunek szansy wystąpienia A do szansy wystąpienia B , który nosi nazwę **iloraz szans** :

$$OR_{A \times B} = \frac{P(A)}{1 - P(A)} : \frac{P(B)}{1 - P(B)}.$$

Jeżeli $OR_{A \times B} > 1$, to szansa wystąpienia pewnego zdarzenia w grupie A jest większa, niż w grupie B .

Ponieważ regresja logistyczna często używana w zagadnieniach klasyfikacji binarnej, tzn. podziału zbiorowości na dwie klasy, to iloraz szans OR często jest obliczany jako stosunek iloczynu poprawnie zaklasyfikowanych przypadków do iloczynu niepoprawnie zaklasyfikowanych przypadków.

MASZYNA WEKTORÓW NOŚNYCH (SVM)

Maszyna wektorów nośnych (*ang.* Support Vector Machines) - abstrakcyjny concept maszyny, która działa jak klasyfikator, a której nauka ma na celu wyznaczenie hiperpłaszczyzny rozdzielającej z maksymalnym marginesem obiekty należące do dwóch klas.

SVM: MODEL LINIOWY

Zbiór uczący to zbiór wartości $\mathbf{X}_{n \times m}$, w którym każdy i -ty wiersz oznaczmy $\mathbf{x}_i = \mathbf{X}(i,:)$ - są to obiekty każdy z których jest przepisany do jednej z dwóch klas.

Także do próby uczącej należy wektor $\mathbf{y} = (y_1, y_2, \dots, y_n)$ taki, że $y_i \in \{-1, 1\}$, czyli wartość y_i jest równa 1 w przypadku, gdy \mathbf{x}_i należy do pierwszej klasy oraz wartość y_i jest równa -1 , gdy \mathbf{x}_i należy do drugiej klasy.

SVM: MODEL LINIOWY

Należy znaleźć taki wektor $\mathbf{b} = (b_1, b_2, \dots, b_m)$, aby dla pewnej granicznej wartości b_0 oraz nowego obiektu \mathbf{x}_i spełniony został warunek:

$$\mathbf{x}_i \mathbf{b}^T > b_0 \Rightarrow y_i = 1$$

$$\mathbf{x}_i \mathbf{b}^T < b_0 \Rightarrow y_i = -1$$

Równanie $\mathbf{x}_i \mathbf{b}^T = b_0$ opisuje hiperpłaszczyznę, która dzieli przestrzeń na dwie klasy. Możemy wybrać dowolną hiperpłaszczyznę, więc skorzystamy z tego w celu polepszania klasyfikacji.

SVM: MODEL LINIOWY

Wyznamy płaszczyznę rozdzielającą w taki sposób, aby jak można dalej stała ona od najbliższych do niej punktów klas, tzn. znajdziemy taki wektor \mathbf{b} oraz taką liczbę b_0 , aby dla pewnej wartości $\varepsilon > 0$ spełniony został warunek:

$$\mathbf{x}_i \mathbf{b}^T > b_0 + \varepsilon \Rightarrow y_i = 1$$

$$\mathbf{x}_i \mathbf{b}^T < b_0 - \varepsilon \Rightarrow y_i = -1$$

Algorytm klasyfikacji nie zmieni się, jeżeli \mathbf{b} i b_0 jednocześnie pomnożyć razy pewną wartość stałą. Wybierzmy stałą w taki sposób, aby dla wszystkich wartości brzegowych (tj. najbliższych do rozdzielającej hiperpłaszczyzny) spełniony został warunek

$\mathbf{x}_i \mathbf{b}^T - b_0 = y_i$. W tym celu pomnóżmy nierówności razy $\frac{1}{\varepsilon}$ i wybierzmy $\varepsilon = 1$.

SVM: MODEL LINIOWY

Wtedy dla wszystkich wektorów ze zbioru uczącego \mathbf{X} zachodzi warunek:

$$\mathbf{x}_i \mathbf{b}^T - b_0 \geq 1 \text{ gdy } y_i = 1$$

$$\mathbf{x}_i \mathbf{b}^T - b_0 \geq -1 \text{ gdy } y_i = -1$$

co oznacza, że warunek:

$$-1 \leq \mathbf{x}_i \mathbf{b}^T - b_0 \leq 1$$

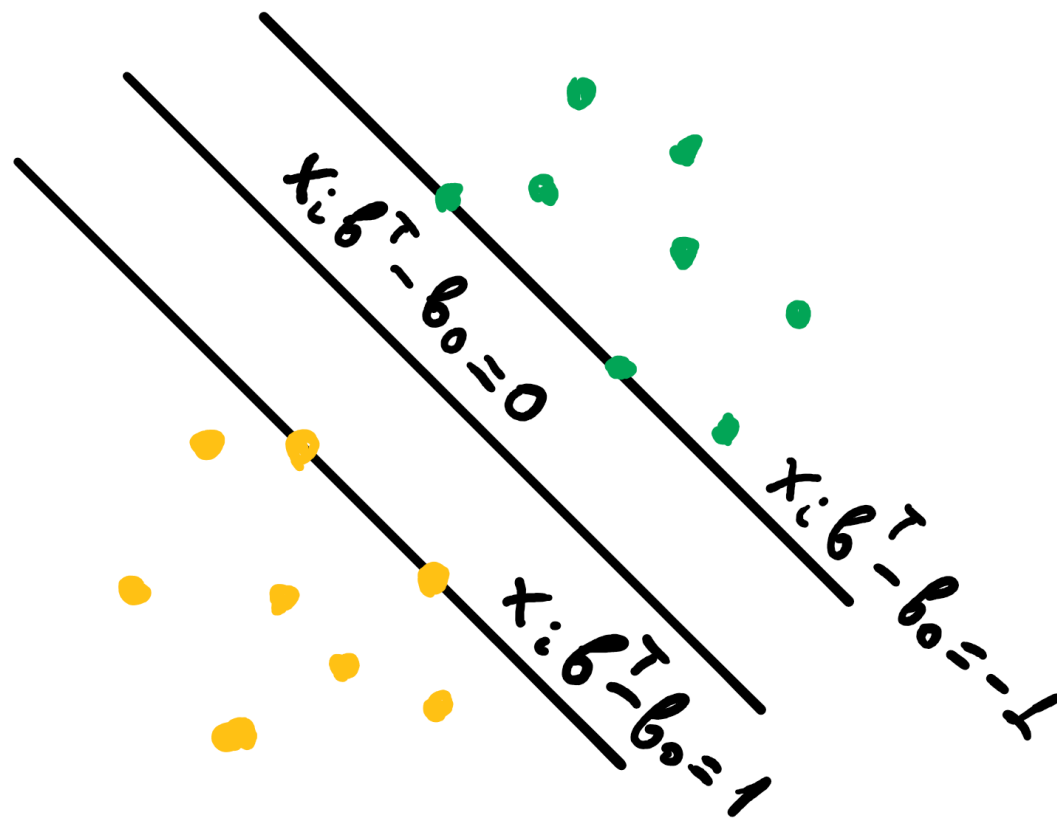
zadaje margines który dzieli punkty próby uczącej na klasy.

Żaden punkt zbioru uczącego nie może znajdować się wewnątrz tego marginesu. Szerokość marginesu jest równa:

$$\frac{2}{\|\mathbf{b}\|}$$

gdzie $\|\mathbf{b}\|$ - norma wektora \mathbf{b} .

SVM: MODEL LINIOWY



SVM: MODEL LINIOWY

Zadanie odzyskania optymalnego marginesu polega na odnalezieniu takich wartości \mathbf{b} oraz b_0 , dla których spełnione jest ograniczenie:

$$y_i (\mathbf{x}_i \mathbf{b}^T - b_0) \geq 1$$

oraz norma wektora \mathbf{b} była jak można najmniejsza (aby szerokość marginesu była największa), czyli należy minimalizować funkcję:

$$\|\mathbf{b}\|^2 \rightarrow \min$$

Wada tej metody jest, że, jeżeli z zbiorze uczącym jest błąd tzn. nie dobrze został zaklasyfikowany jeden lub kilka punktów, to margines może być wyznaczony błędnie.

SVM: MODEL LINIOWY

Niech algorytm „dopuszcza” błędy na zbiorze uczącym. Wprowadźmy zbiór dodatkowych zmiennych $\xi_i \geq 0$, charakteryzujących wartość błędu na obiektach \mathbf{x}_i .

Wtedy:

$$y_i \cdot (\mathbf{x}_i \mathbf{b}^T - b_0) \geq 1 - \xi_i$$

Jeśli:

- $\xi_i = 0$, to błędu brak,
- $\xi_i > 1$, to w punkcie \mathbf{x}_i jest błąd,
- $0 < \xi_i \leq 1$, to obiekt trafi do marginesu, ale algorytm przypisze go do właściwej klasy.

ZADANIE OPTYMALIZACJI

Zadanie optymalizacji wygląda następująco:

$$\|\mathbf{b}\|^2 + C \sum_{i=1}^n \xi_i \rightarrow \min.$$

Współczynnik C - parametr dopasowania metody, który pozwala regulować relacje pomiędzy maksymalizacją szerokości marginesu i minimalizacją błędu sumarycznego. Parametr ten wybiera się ręcznie (o tym później).

ZADANIE OPTYMALIZACJI

Do rozwiązania zadania optymalizacji z ograniczeniami skorzystamy z metody Lagrange'a, która pozwala zadanie poszukiwania minimum lokalnego (zadanie z ograniczeniami) doprowadzić do zadania poszukiwania minimum globalnego (zadanie bez ograniczeń). Funkcja celu wygląda następująco:

$$\Phi = \frac{1}{2} \|\mathbf{b}\|^2 + C \sum_{i=1}^n \xi_i - \sum_{i=1}^n \lambda_i \left(y_i (\mathbf{x}_i \mathbf{b}^T - b_0) - 1 \right),$$

gdzie λ_i - współczynniki Lagrange'a, $\lambda_i \geq 0$.

Wartość $\frac{1}{2}$ wprowadza się sztucznie do funkcji Φ w celu łatwiejszego (wygodniejszego) rozwiązania zadania optymalizacji, nie zmienia ona rozwiązania.

ZADANIE OPTYMALIZACJI

Funkcję Φ należy minimalizować po \mathbf{b} , b_0 oraz maksymalizować po λ_i :

$$\begin{cases} \frac{\partial \Phi}{\partial \mathbf{b}} = \mathbf{b} - \sum_{i=1}^n \lambda_i y_i \mathbf{x}_i = 0 \\ \frac{\partial \Phi}{\partial b_0} = \sum_{i=1}^n \lambda_i y_i = 0 \\ \frac{\partial \Phi}{\partial \lambda_i} = \left(y_i (\mathbf{x}_i \mathbf{b}^T - b_0) - 1 \right) = 0 \end{cases}$$

ZADANIE OPTIMALIZACJI

Z równania pierwszego wynika:

$$\mathbf{b} = \sum_{i=1}^n \lambda_i y_i \mathbf{x}_i$$

Oznacza to, że wektor \mathbf{b}^T musi być kombinacją liniową wektorów uczących, przy czym tylko takich dla których $\lambda_i \neq 0$.

Jeśli $\lambda_i > 0$, to \mathbf{x}_i zbioru uczącego nazywa się **wektorem nośnym**.

Wektory nośne - wektory przez które przechodzą graniczne hiperpłaszczyzny, rozdzielające przestrzeń na klasy.

ZADANIE OPTYMALIZACJI

Z równania trzeciego układu wynika, że na podstawie dowolnego \mathbf{x}_s ze zbioru uczącego:

$$b_0 = \mathbf{x}_s \mathbf{b}^T - \frac{1}{y_s}$$

ZADANIE OPTIMALIZACJI

W przypadku, gdy w próbie uczącej brak błędów, czyli wszystkie $\xi_i = 0$, przy podstawieniu \mathbf{b}^T do Φ oraz przy uwzględnieniu, że

$\|\mathbf{b}\|^2 = \mathbf{b}\mathbf{b}^T$ i $\sum_{i=1}^n \lambda_i y_i = 0$ w przestrzeni $\lambda \geq 0$, otrzymamy:

$$\begin{aligned}\Phi &= \frac{1}{2} \mathbf{b}\mathbf{b}^T - \sum_{i=1}^n \lambda_i y_i \mathbf{x}_i \mathbf{b}^T - b_0 \sum_{i=1}^n \lambda_i y_i + \sum_{i=1}^n \lambda_i = \\ &= \frac{1}{2} \mathbf{b}\mathbf{b}^T - \mathbf{b}\mathbf{b}^T + b_0 \cdot 0 + \sum_{i=1}^n \lambda_i = \sum_{i=1}^n \lambda_i - \frac{1}{2} \mathbf{b}\mathbf{b}^T = \\ &= \sum_{i=1}^n \lambda_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j y_i y_j (\mathbf{x}_i \mathbf{x}_j^T)\end{aligned}$$

ZADANIE OPTYMALIZACJI

Otrzymaliśmy nowe zadanie optymalizacji:

$$\Phi = \sum_{i=1}^n \lambda_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j y_i y_j (\mathbf{x}_i \mathbf{x}_j^T) \rightarrow \max$$

$$\sum_{i=1}^n \lambda_i y_i = 0 \text{ w przestrzeni } \lambda \geq 0.$$

PRZYKŁAD

Wykonać podział na klasy metodą wektorów nośnych na podstawie próby uczącej:

$$\mathbf{X} = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 2 & 3 \end{pmatrix}, \mathbf{y}^T = \begin{pmatrix} 1 \\ -1 \\ -1 \end{pmatrix}$$

PRZYKŁAD

Funkcja Φ jest równa:

$$\begin{aligned}\Phi(\lambda) &= \sum_{i=1}^n \lambda_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j y_i y_j (\mathbf{x}_i \mathbf{x}_j^T) = \\ &= \sum_{i=1}^3 \lambda_i - \frac{1}{2} \sum_{i=1}^3 \sum_{j=1}^3 \lambda_i \cdot \lambda_j \cdot y_i \cdot y_j \cdot (\mathbf{x}_i \cdot \mathbf{x}_j^T) = \\ &= \lambda_1 + \lambda_2 + \lambda_3 - \frac{1}{2} \cdot \left(\lambda_1 \lambda_1 y_1 y_1 (\mathbf{x}_1 \mathbf{x}_1^T) + \lambda_1 \lambda_2 y_1 y_2 (\mathbf{x}_1 \mathbf{x}_2^T) + \dots + \lambda_3 \lambda_3 y_3 y_3 (\mathbf{x}_3 \mathbf{x}_3^T) \right) = \\ &= \lambda_1 + \lambda_2 + \lambda_3 - \frac{1}{2} \cdot \left(\lambda_1 \lambda_1 \cdot 1 \cdot 1 \left(\begin{pmatrix} 1 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right) + \lambda_1 \lambda_2 \cdot 1 \cdot (-1) \begin{pmatrix} 1 & 2 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \end{pmatrix} + \dots \right) = \\ &= \lambda_1 + \lambda_2 + \lambda_3 - \frac{1}{2} \cdot (2 \cdot \lambda_1^2 + 5 \cdot \lambda_2^2 + 13 \cdot \lambda_3^2 - 6 \cdot \lambda_1 \cdot \lambda_2 - 10 \cdot \lambda_1 \cdot \lambda_3 + 16 \cdot \lambda_2 \cdot \lambda_3) \end{aligned}$$

PRZYKŁAD

Ponieważ $\sum_{i=1}^n \lambda_i y_i = 0$, to:

$$\lambda_1 - \lambda_2 - \lambda_3 = 0 \Rightarrow \lambda_3 = \lambda_1 - \lambda_2$$

Podstawiając to do Φ otrzymamy:

$$\Phi(\lambda_1, \lambda_2) = 2 \cdot \lambda_1 - 2,5 \cdot \lambda_1^2 - \lambda_2^2 + 3 \cdot \lambda_1 \cdot \lambda_2$$

Szukamy wartości λ_i :

$$\begin{cases} \frac{\partial \Phi(\lambda_1, \lambda_2)}{\partial \lambda_1} = 0 \\ \frac{\partial \Phi(\lambda_1, \lambda_2)}{\partial \lambda_2} = 0 \end{cases} \Leftrightarrow \begin{cases} 2 - 5 \cdot \lambda_1 + 3 \cdot \lambda_2 = 0 \\ -2 \cdot \lambda_2 + 3 \lambda_1 = 0 \end{cases} \Leftrightarrow \begin{cases} \lambda_1 = 4 \\ \lambda_2 = 6 \end{cases}$$

PRZYKŁAD

Ponieważ $\lambda_3 = \lambda_1 - \lambda_2$, to $\lambda_3 = \lambda_1 - \lambda_2 = 4 - 6 = -2 < 0$, zbadamy funkcję $\Phi(\lambda_1, \lambda_2)$ na granicach $\lambda \geq 0$:

$\lambda_1 = 0$ $\lambda_3 = -\lambda_2$ $\Lambda = (0 \quad 0 \quad 0)$ $\Phi(\Lambda) = 0$	$\lambda_2 = 0$ $\lambda_1 = \lambda_3 = \lambda$ $\Lambda = (\lambda \quad 0 \quad \lambda)$ $\Phi(\Lambda) = 2 \cdot \lambda - 2,5 \cdot \lambda^2$ $\Phi'(\Lambda) = 0 \text{ dla } \lambda = \frac{2}{5}$ $\Lambda = \left(\frac{2}{5} \quad 0 \quad \frac{2}{5} \right)$ $\Phi(\Lambda) = \frac{2}{5}$	$\lambda_3 = 0$ $\lambda_1 = \lambda_2 = \lambda$ $\Lambda = (\lambda \quad \lambda \quad 0)$ $\Phi(\Lambda) = 2 \cdot \lambda - 0,5 \cdot \lambda^2$ $\Phi'(\Lambda) = 0 \text{ dla } \lambda = 2$ $\Lambda = (2 \quad 2 \quad 0)$ $\Phi(\Lambda) = 2$
---	--	---

PRZYKŁAD

Największą wartość $\Phi(\Lambda)$ przyjmuje dla $\Lambda = (2 \ 2 \ 0)$, więc:

$$\left\{ \begin{array}{l} \mathbf{b} = \sum_{i=1}^3 \lambda_i \cdot y_i \cdot \mathbf{x}_i = 2 \cdot \mathbf{x}_1 - 2 \cdot \mathbf{x}_2 = 2 \cdot 1 \cdot (1 \ 1) - 2 \cdot (1)(1 \ 2) = (0 \ -2) \\ b_0 = \mathbf{x}_1 \mathbf{b}^T - \frac{1}{y_1} = (1 \ 1) \begin{pmatrix} 0 \\ -2 \end{pmatrix} - \frac{1}{(-1)} = -3 \end{array} \right.$$

PRZYKŁAD

Wzór prostej rozdzielającej przestrzeń na dwie klasy dla dowolnego obiektu $\mathbf{x} = (x_1 \ x_2)$:

$$f(\mathbf{x}) = \mathbf{x}\mathbf{b}^T - b_0 = (x_1 \ x_2) \begin{pmatrix} 0 \\ -2 \end{pmatrix} - (-3) = -2 \cdot x_2 + 3$$

Jeśli $f(\mathbf{x})$, czyli $-2 \cdot x_2 + 3 = 0$, to $x_2 = 1,5$.

Margines wynosi:

$$\frac{2}{\|\mathbf{b}\|} = \frac{2}{\sqrt{0 + (-2)^2}} = 1$$

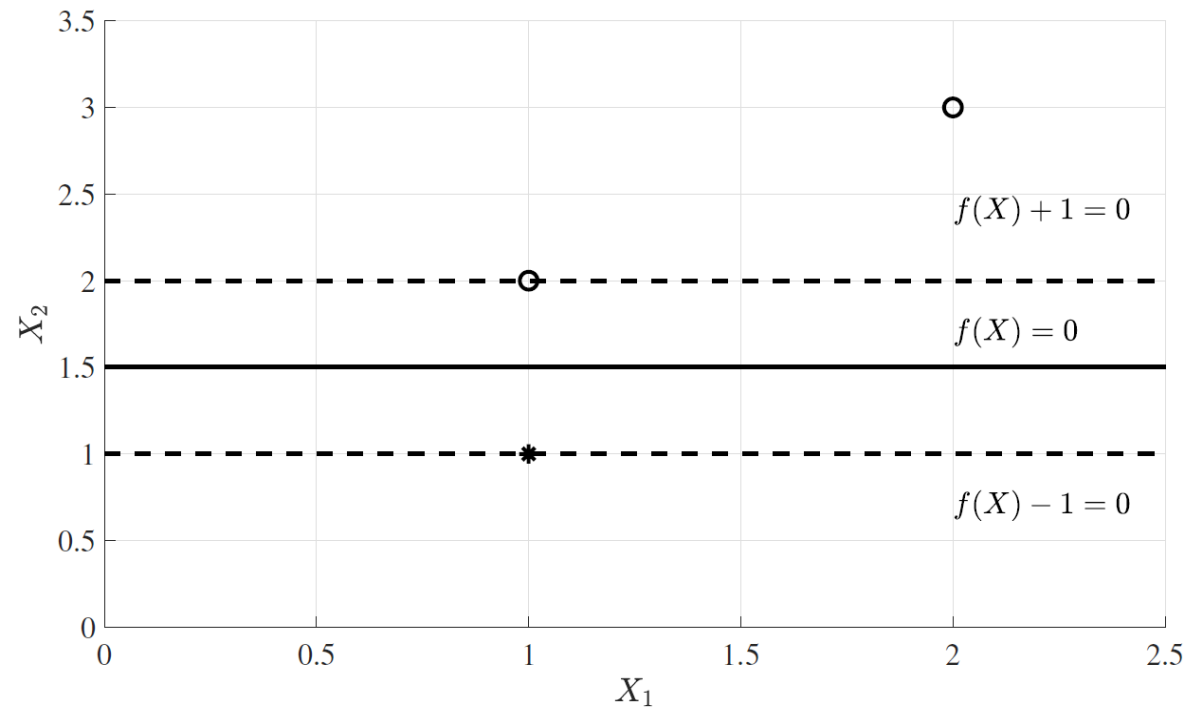
czyli proste $f(\mathbf{x}) \pm 1$ są granicami:

$$f(\mathbf{x}) + 1 = -2 \cdot x_2 + 3 + 1 = 0 \Rightarrow x_2 = 2$$

$$f(\mathbf{x}) - 1 = -2 \cdot x_2 + 3 - 1 = 0 \Rightarrow x_2 = 1$$

PRZYKŁAD

Na rys. pokazany rzut na płaszczyznę X_1 - X_2 obiektów próby uczącej (gwiazdką oznaczona klasa pierwsza, kółkami - klasa druga), prosta rozdzielająca $f(\mathbf{x}) = 0$ oraz granice decyzyjne:



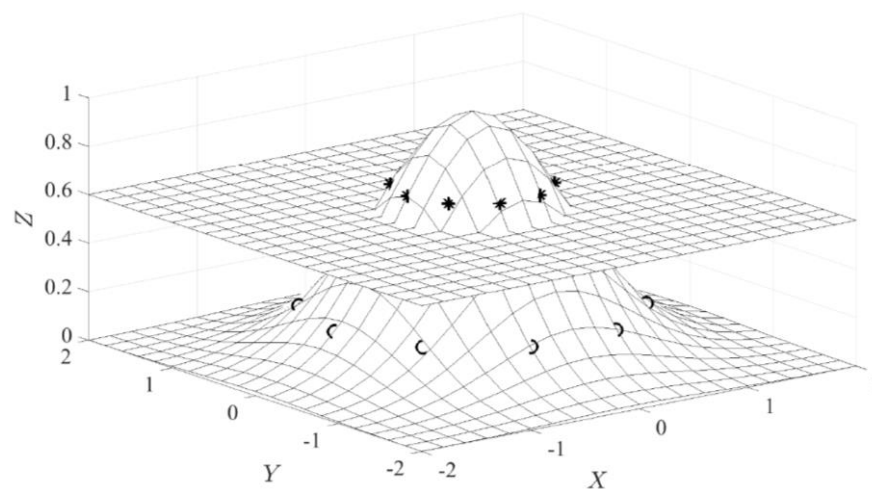
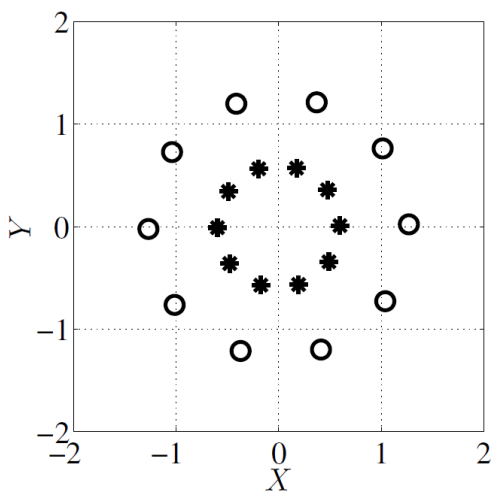
SVM: MODEL NIELINIOWY

W roku 1992 B. Boser, I. Guyon i W. Wapnik zaproponowali sposób adaptacji maszyny wektorów nośnych dla nieliniowego podziału na klasy. W przypadku takim należy przestrzeń R^n "włożyć" do przestrzeni H większej wymiarowości za pomocą odwzorowania:

$$\varphi : R^n \rightarrow H$$

SVM: MODEL NIELINIOWY

Ideę można wytłumaczyć na następującym przykładzie. Na rys.1 w przestrzeni dwuwymiarowej ółeczka i gwiazdki nie są podzielone, jednak jeżeli przeniesiemy te punkty w przestrzeń trójwymiarową rys.2, to punkty te zostaną podzielone za pomocą płaszczyzny, tzn. jeżeli wygiąć płaszczyznę za pomocą pewnego odwzorowania $\varphi(\mathbf{x})$ to łatwo można znaleźć rozdzielającą hiperpłaszczyznę.



SVM: MODEL NIELINIOWY

Przy wykorzystaniu odwzorowania $\varphi: R^n \rightarrow H$ metoda wektorów nośnych rozpatruje się dla obrazów $\varphi(\mathbf{x})$ próby uczącej i doprowadza zadanie do przypadku liniowego, tzn. funkcja graniczna jest postaci:

$$f(\mathbf{x}) = \varphi(\mathbf{x})\mathbf{b}^T - b_0,$$

gdzie

$$\mathbf{b} = \sum_{i=1}^n \lambda_i y_i \varphi(\mathbf{x}_i), \quad b_0 = \varphi(\mathbf{x}_s)\mathbf{b}^T - \frac{1}{y_s}$$

i wartości λ_i zależą od y_i oraz od

$$K = \varphi(\mathbf{x}_i)\varphi(\mathbf{x}_j)^T.$$

W rzeczywistości nie tyle należy znać postać odwzorowania $\varphi(\mathbf{x})$ ile postać K , która nazywa **jądrem**.

JĄDRO

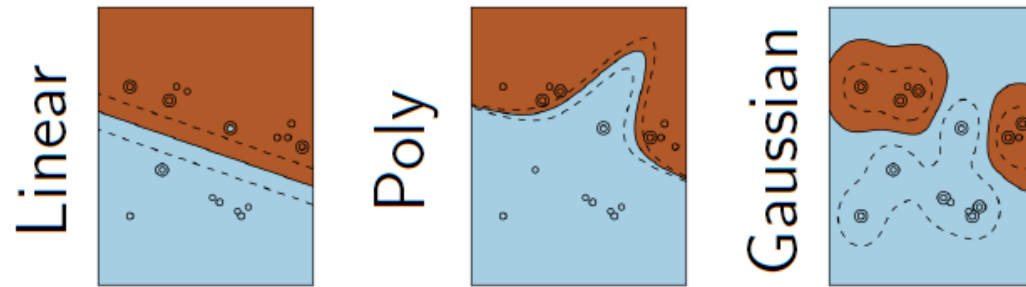
Funkcja $K(\mathbf{x}_i, \mathbf{x}_j)$ jest jądrem wtedy i tylko wtedy, gdy jest ona symetryczna oraz nieujemnie zdefiniowana dla dowolnych wektorów \mathbf{x}_i . Nie istnieje ogólnego algorytmu wyboru jądra w przypadku liniowej niepodzielności klas.

Najbardziej popularne jądra są następujące:

- $K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i \mathbf{x}_j^T + \gamma$ - jądro liniowe;
- $K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i \mathbf{x}_j^T + \gamma)^r$ - jądro wielomianowe;
- $K(\mathbf{x}_i, \mathbf{x}_j) = e^{-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2}$ - jądro radialnych funkcji bazowych (RBF);
- $K(\mathbf{x}_i, \mathbf{x}_j) = \tanh(\beta \mathbf{x}_i \mathbf{x}_j^T - \gamma)$ - jądro sigmoidalne,

gdzie β, γ - stałe współczynniki liczbowe, r - stopień wielomianu.

JĄDRO



Dla przykładu skorzystamy z własności, że dowolne $r + 1$ wektory mogą zostać podzielone na dwie klasy za pomocą odwzorowania jednomianowego stopnia nie więcej niż r .

Jeśli $\varphi : \mathbf{x} \rightarrow \{\mathbf{x}_1^{i_1} \dots \mathbf{x}_n^{i_n}\}$, $i_1 + \dots + i_n \leq r$ jest takim odwzorowaniem, to odpowiadające mu jądro można szukać postaci:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \varphi(\mathbf{x}_i) \varphi(\mathbf{x}_j)^T = (\mathbf{x}_i \mathbf{x}_j^T + 1)^r$$

Jądro takie gwarantuje podział dowolnych $r + 1$ wektorów na dwie klasy.

ZADANIE OPTYMALIZACJI

Zadanie podziału na klasy jest zadaniem optymalizacji:

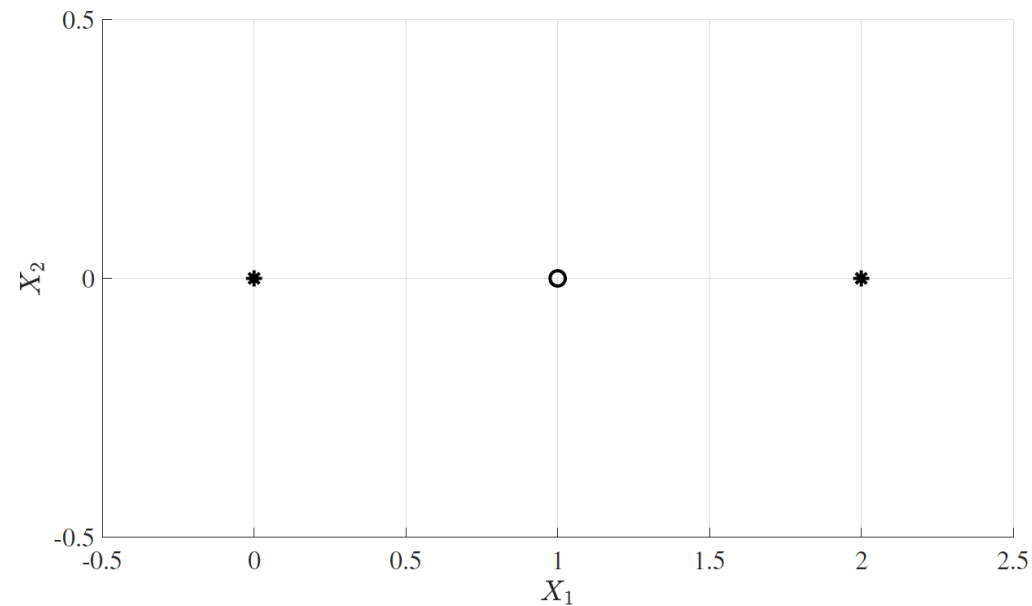
$$\Phi = \sum_{i=1}^n \lambda_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \rightarrow \max, \quad (1)$$

pod $\sum_{i=1}^n \lambda_i y_i = 0$ w przestrzeni $\lambda \geq 0$.

PRZYKŁAD

Wykonać podział na klasy metodą wektorów nośnych na podstawie próby uczącej:

$$\mathbf{X} = \begin{pmatrix} 0 & 0 \\ 2 & 0 \\ 1 & 0 \end{pmatrix}, \mathbf{y}^T = \begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix}$$



PRZYKŁAD

Ponieważ liczba wektorów próby uczącej jest równa 3, to $r = 2$ i w postaci jądra weźmiemy funkcję:

$$K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i \mathbf{x}_j^T + 1)^2$$

Podstawiając wartości próby uczącej do funkcji celu otrzymamy:

$$\Phi(\lambda_1, \lambda_2) = 2\lambda_1 + 2\lambda_2 - 0,5(3\lambda_1^2 + 11\lambda_2^2 - 10\lambda_1\lambda_2),$$

gdzie $\lambda_3 = \lambda_1 + \lambda_2$.

Szukamy wartości λ_i :

$$\begin{cases} \frac{\partial \Phi}{\partial \lambda_1} = 0 \\ \frac{\partial \Phi}{\partial \lambda_2} = 0 \end{cases} \Leftrightarrow \begin{cases} 2 - 3\lambda_1 + 5 \cdot \lambda_2 = 0 \\ 2 - 11\lambda_2 + 5 \cdot \lambda_1 = 0 \end{cases} \Leftrightarrow \begin{cases} \lambda_1 = 4 \\ \lambda_2 = 2 \\ \lambda_3 = 6 \end{cases}$$

PRZYKŁAD

Można pokazać, że w punkcie $\Lambda = (4 \ 2 \ 6)^T$ będzie osiągnięto maksimum funkcji $\Phi(\Lambda)$.

Funkcja rozdzielająca ma postać:

$$f(\mathbf{x}) = 2x_1^2 - 4x_1 + 1$$

W przypadku, gdy $f(\mathbf{x}) = 0$ $x_1 = 1 \pm \frac{\sqrt{2}}{2}$.

Margines:

$$f_1(\mathbf{x}) = f(\mathbf{x}) + 1 = (2x_1^2 - 4x_1 + 1) + 1 = (x_1 - 1)^2$$

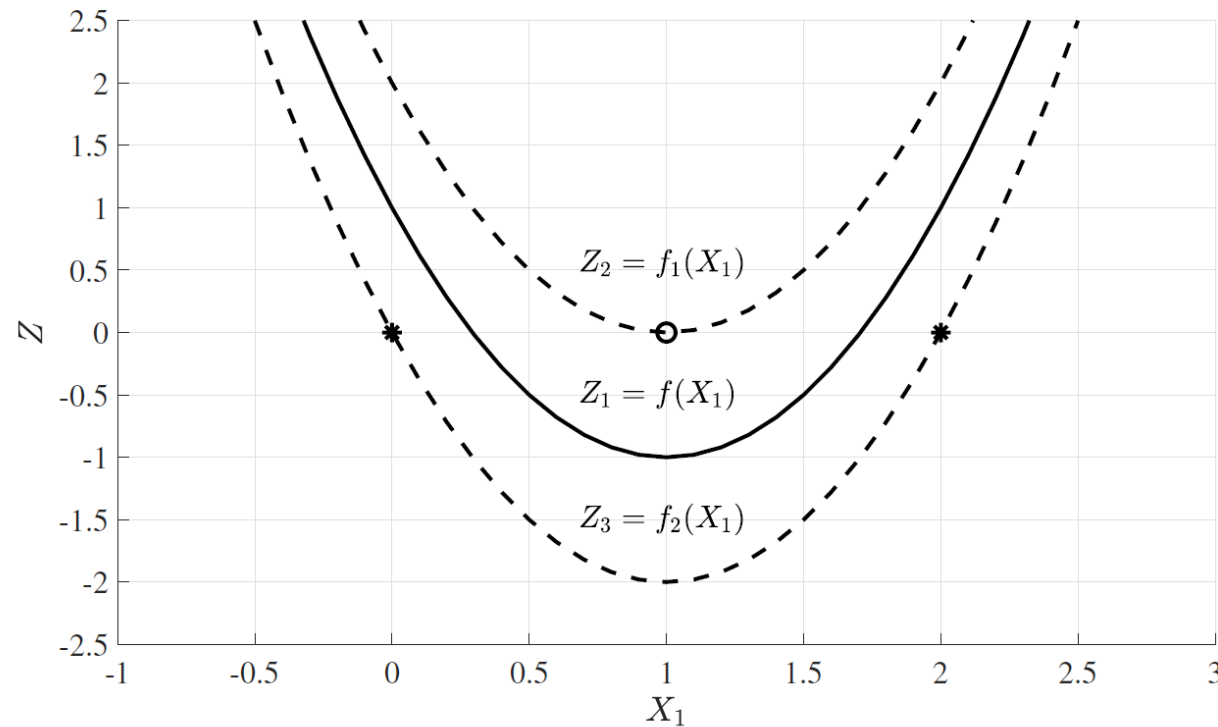
$$f_2(\mathbf{x}) = f(\mathbf{x}) - 1 = (2x_1^2 - 4x_1 + 1) - 1 = 2x_1(x_1 - 2)$$

$$f_1(\mathbf{x}) = 0 \text{ gdy } x_1 = 1$$

$$f_2(\mathbf{x}) = 0 \text{ gdy } x_1 = 0 \wedge x_1 = 2.$$

PRZYKŁAD

Można zilustrować podział elementów próby uczącej wykresami funkcji $Z_1 = f(x_1), Z_2 = f_1(x_1), Z_3 = f_2(x_1)$ pokazując rzut na płaszczyznę X_1OZ .



ZALETY I WADY

Zaletami metody SVM jest to, że metoda ta zawsze doprowadza zadania do jednoznacznego rozwiązania, jest najszybszą metodą odnalezienia funkcji decydujących a także pozwala na znalezienie marginesu maksymalnej szerokości, co doprowadza do pewniejszej klasyfikacji.

Jednak jest ona nieodporna na szum i standaryzację danych oraz brak jest ogólnego algorytmu wyboru jądra w przypadku liniowej niepodzielności klas.

ZALETY I WADY

Ponieważ niezależnie od zastosowanego jądra i rodzaju zadania obliczenia SVM sprowadza się do rozwiązania zadania programowania kwadratowego z ograniczeniami liniowymi, to problemem staje się duża ilość danych uczących, co związane jest z nieraz ogromną ilością optymalizowanych zmiennych (mnożników Lagrange'a). Pojawiają się problemy z pamięcią i złożonością obliczeniową, co eliminuje możliwość zastosowania klasycznych metod programowania kwadratowego. Dla rozwiązania tego problemu stosuje się dekompozycję zbioru uczącego na szereg podzbiorów oraz strategię aktywnych ograniczeń wynikających z równości, zaniedbując te nieaktywne ze znakiem silnie nierówności; wykorzystuje się również różne wersje algorytmu programowania sekwencyjnego SMO (*ang.* Sequential Minimal Optimization) lub metoda optymalizacji heurystyki Platta oraz suboptymalną metodę Joachimsa. Dalej pokażemy metodę SMO.

SDT

Algorytm ***pojedynczego drzewa decyzyjnego*** (*ang.* single decision tree, SDT) został po raz pierwszy przedstawiony w pracy (Hunt et al, 1966) a następnie niezależnie w (Friedman, 1977) oraz (Breiman et al, 1984).

SDT jest hierarchiczną strukturą przedstawianą w formie schematu blokowego (grafu) używaną do wspomagania procesu decyzyjnego.

W kontekście algorytmu uczenia, SDT można rozpatrywać jako model predykcyjny, który dokonuje odwzorowania wejście-wyjście na podstawie zadanych rekordów z oczekiwanym wyjściem pożądanym.

STRUKTURA SDT

Jeżeli oczekiwane wyjście przyjmuje postać dyskretną (klasa, etykieta), SDT jest **drzewem klasyfikacji**.

W przypadku, gdy wyjście rekordu jest liczbą rzeczywistą, SDT traktowane jest jako **drzewo regresji**.

W strukturze SDT można wyróżnić trzy rodzaje elementów:

➤ **węzły**: reprezentują **podział** danych na podstawie wybranej cechy (zmiennej); korzeń jest węzłem umieszczanym na szczycie drzewa;

➤ **gałęzie**: reprezentują wartości poszczególnych cech (zmiennych, atrybutów);

➤ **liście**: reprezentują etykiety klas (klasyfikacja) albo wartości zmiennoprzecinkowe (regresja).

STRUKTURA SDT

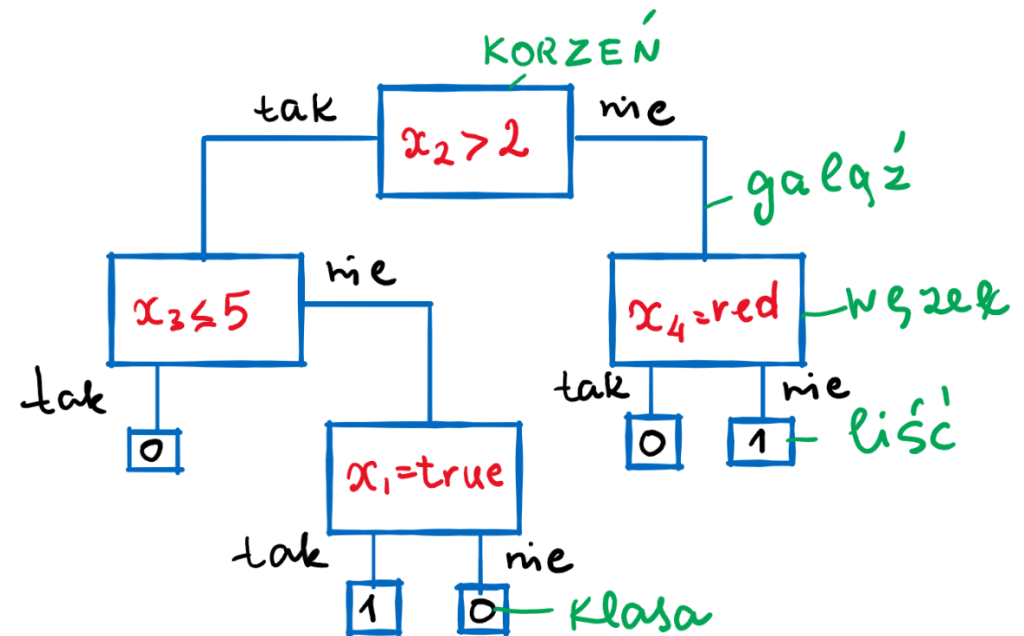
Podział danych w oparciu o wybraną cechę realizowany jest w inny sposób dla atrybutów dyskretnych i ciągłych.

Dla „cechy dyskretne” przypisywana jest pojedyncza gałąź dla każdej możliwej jej wartości (lub grupy wartości). W przypadku „cechy ciągłe” wprowadza się próg liczbowy, który tworzy binarny podział dla danej cechy. Poszczególne węzły wybierane są na podstawie różnych kryteriów (miar informacji), które wyznaczane są przez miary stopnia niejednorodności danych.

Będziemy dalej rozpatrywać drzewa klasyfikacyjne.

STRUKTURA SDT

Algorytm konstruowania drzewa decyzyjnego jest realizowany na zasadzie „dziel i zwyciężaj” i polega na funkcji rekurencyjnej, która jako argument przyjmuje bazę reguł. Baza reguł składa się z atrybutów (zmiennych) oraz przypisanych im wartości.



Proces tworzenia drzew polega na odpowiednim wyborze zmiennych (cech, atrybutów), jako kandydatów na węzły.

W kolejnych rekurencyjnych przebiegach algorytmu węzły dzielone są na podwęzły, aż do uzyskania maksymalnego drzewa.

STRUKTURA SDT

Standardowo podział odbywa się zachłannie, a więc wybierany jest potencjalnie najlepszy podział pod względem wartości wybranego kryterium podziału. Dobre kryterium podziału powinno jak najmniej różnicować obiekty (pod względem ich klasy decyzyjnej) w każdym potomku węzła. W momencie kiedy w węźle wszystkie obiekty należą do jednej klasy decyzyjnej, węzeł ten staje się liściem.

W efekcie drzewo decyzyjne reprezentuje proces podziału obiektów (pod względem wartości ich zmiennych) ze zbioru danych na jednorodne klasy. Kryterium podziału powinien minimalizować błąd klasyfikacji przypadków ze zbioru testowego.

Wybór podziału danych w każdym węźle jest najtrudniejszym i najbardziej złożonym etapem konstruowania drzew decyzyjnych. Zastosowanie konkretnego kryterium zależne jest od stosowanego algorytmu lub nawet konkretnych zastosowań algorytmu.

OGÓLNY SCHEMAT SDT

1. Drzewo zaczyna od pojedynczego węzła.
2. Jeżeli wszystkie przykłady należą do jednej klasy decyzyjnej, to zbadany węzeł staje się liściem i jest on etykietowany tą decyzją.

W przeciwnym przypadku oblicza się kryterium podziału do wyboru zmiennej (atrybutu), która najlepiej dzieli zbiór przykładów.

3. Dla każdego wyniku tworzy się jedno odgałęzienie i przypadki są odpowiednio rozdzielone do nowych węzłów.
4. Algorytm działa dalej w rekurencyjny sposób dla zbiorów przykładów przydzielonych do poddrzew.
5. Algorytm kończy się, gdy kryterium stopu jest spełnione.

ALGORYTMY SDT

1. CART (*ang.* classification and regression trees) (Breiman et al, 1984): drzewa decyzyjne tworzone przez ten algorytm są binarne, czyli zawierają dwie gałęzie w każdym z węzłów. Może być stosowane zarówno dla danych ciągłych jak i dyskretnych.
2. ID3 (Quinlan, 1986) - cechuje się prostotą, wymaga kompletnych reguł, nie pozwala na szum w danych; zakłada, że atrybuty przyjmują wartości dyskretne, a nie ciągłe.
3. C4.5 (Quinlan, 1993) jest rozszerzeniem algorytmu ID3; rozwiązuje większość problemów algorytmu ID3, np.: braki w danych, wartości ciągłe atrybutów, możliwość przycinania zbyt rozrośniętych struktur (drzew).
4. Inne podejścia: C5, ID4, ID5, CHAID, CLS . . .

ILOŚĆ INFORMACJI

Proces tworzenia drzewa polega na odpowiednim wyborze zmiennych jako kandydatów na węzły; finalne tworzenie węzłów realizowane jest w oparciu o pewne kryterium informacji.

Prawdopodobieństwo wystąpienia przypadku:

$$p(c_k, \mathbf{X}) = \frac{n_k}{n}$$

gdzie n_k oznacza liczbę przypadków w zbiorze \mathbf{X} , które należą do klasy $c_k, k = 1, 2, \dots, K$, a $n = |\mathbf{X}|$ - liczność zbioru \mathbf{X} .

Ilość przekazywanej informacji:

$$I(c_k, \mathbf{X}) = -\log_2 p(c_k, \mathbf{X}) = -\log_2 \left(\frac{n_k}{n} \right).$$

Istnieje wiele różnych odmian algorytmu CART, wszystkie one tworzą podział na podstawie indeksu Giniego lub entropii.

ILOŚĆ INFORMACJI

Algorytmy ID3 oraz C4.5 do wyboru zmiennej na węzła wykorzystuje kryterium zysku, które z kolei uwzględnia prawdopodobieństwo wystąpienia przypadku oraz ilość przekazywanej informacji.

Aby zbudować SDT, należy podać zysk jaki daje podział zbioru X na podstawie cechy f . W tym celu należy wyznaczyć:

ZYSK

1. **średnią ilość informacji** o przynależności danych do wszystkich klas w zbiorze \mathbf{X} za pomocą wyznaczenia **entropii**:

$$H(\mathbf{X}) = - \sum_{k=1}^K p(c_k, \mathbf{X}) \log_2 (p(c_k, \mathbf{X}));$$

2. informację jaka jest wymagana do podziału danych w zbiorze \mathbf{X} na podstawie J możliwych wartości cechy f :

$$H_f(\mathbf{X}) = - \sum_{j=1}^J \frac{|\mathbf{X}_j|}{|\mathbf{X}|} \cdot H(\mathbf{X}_j),$$

gdzie \mathbf{X}_j jest j -tym podzbiorem, przyjmującym te same wartości dla zmiennej f ;

ZYSK

3. zysk ΔH_f uzyskany dzięki podziałowi zbioru \mathbf{X} na podstawie zmiennej f :

$$\Delta H_f = H(\mathbf{X}) - H_f(\mathbf{X}).$$

Wzór ten definiuje kryterium zysku przyniesionego z podziału danych względem zmiennej. Dzięki temu kryterium wybierana jest zmienna, która maksymalizuje zysk na informacje.

Kroki 1-3 są podstawą tworzenia SDT w algorytmie ID3 Quinlana.

Mimo, że ID3 daje całkiem dobre wyniki, użycie wyłącznie kryterium zysku do selekcji zmiennych ma poważną wadę: obserwuje się skłonność algorytmu do wyboru zmiennych, które przyjmują wiele różniących się wartości, co doprowadza do rozrastania się drzewa. Wada ta została usunięta w „następcy” ID3, tj. C4.5.

NORMALIZACJA ZYSKU

Quinlan wprowadził normalizację zysku:

$$\Delta H_f^{norm} = \frac{\Delta H_f}{H_{Jf}},$$

gdzie:

$$H_{Jf} = \sum_{j=1}^J \frac{|\mathbf{x}_j|}{|\mathbf{x}|} \log_2 \left(\frac{|\mathbf{x}_j|}{|\mathbf{x}|} \right)$$

reprezentuje ilość informacji uzyskana przez podział zbioru \mathbf{X} na J podzbiorów względem zmiennej f .

PRZYKŁAD: DANE

Dany jest zbiór $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$ rekordów uczących o 4 cechach:

1. płeć={kobieta, mężczyzna}
2. posiadanie samochodu={0, 1, 2}
3. odległość={krótka, średnia, duża}
4. dochód={niski, średni, wysoki}

Dla każdego rekordu $i = 1, 2, \dots, n$ istnieją pary (\mathbf{x}_i, c_k) , gdzie $c_k = \{\text{samochód, pociąg, samolot}\}$ - klasa decyzyjna.

PRZYKŁAD: DANE C.D.

Celem jest stworzenie SDT, które posiada przepis na wybór transportu.

\mathbf{x}_1	$f = 1$	$f = 2$	$f = 3$	$f = 4$	C_k
\mathbf{x}_1	mężczyzna	0	krótka	niski	samochód
\mathbf{x}_2	mężczyzna	1	krótka	średni	samochód
\mathbf{x}_3	kobieta	1	krótka	średni	pociąg
\mathbf{x}_4	kobieta	0	krótka	niski	samochód
\mathbf{x}_5	mężczyzna	1	krótka	średni	samochód
\mathbf{x}_6	mężczyzna	0	średnia	średni	pociąg
\mathbf{x}_7	kobieta	1	średnia	średni	pociąg
\mathbf{x}_8	kobieta	1	duża	wysoki	samolot
\mathbf{x}_9	mężczyzna	2	duża	średni	samolot
\mathbf{x}_{10}	kobieta	2	duża	wysoki	samolot

PRZYKŁAD: PRAWDOPODOBIEŃSTWA

$K = 3$, więc $c_1 = \text{samochód}$, $c_2 = \text{pociąg}$, $c_3 = \text{samolot}$.

Prawdopodobieństwo przynależności danych do każdej klasy wynosi:

$$p_1 = p(c_1, \mathbf{X}) = \frac{n_1}{n} = \frac{4}{10},$$

$$p_2 = p(c_2, \mathbf{X}) = \frac{n_2}{n} = \frac{3}{10},$$

$$p_3 = p(c_3, \mathbf{X}) = \frac{n_3}{n} = \frac{3}{10}.$$

Warto również zauważyć, że dla $f = 1$ (płeć) $J = 2$, dla pozostałych wartości f $J = 3$.

I ITERACJA ALGORYTMU: ENTROPIA

Znajdywanie pierwszego węzła dla drzewa decyzyjnego:

1. Entropia zbioru **X**:

$$\begin{aligned} H(\mathbf{X}) &= -p_1 \log_2(p_1) - p_2 \log_2(p_2) - p_3 \log_2(p_3) = \\ &= -\frac{4}{10} \log_2\left(\frac{4}{10}\right) - \frac{3}{10} \log_2\left(\frac{3}{10}\right) - \frac{3}{10} \log_2\left(\frac{3}{10}\right) = \\ &= 0,5288 + 0,5211 + 0,5211 = 1,570 \end{aligned}$$

I ITERACJA ALGORYTMU: INFORMACJA

2. Informacja do podziału \mathbf{X} na podstawie wartości zmiennej „płeć”, $J = 2$:

➤ prawdopodobieństwa przynależności danych do każdej klasy (samochód, pociąg, samolot) w podzbiorze „mężczyzna”:

$$p_1 = 3/5, p_2 = 1/5, p_3 = 1/5$$

$$H(\mathbf{X}_1) = -\frac{3}{5}\log_2\left(\frac{3}{5}\right) - \frac{1}{5}\log_2\left(\frac{1}{5}\right) - \frac{1}{5}\log_2\left(\frac{1}{5}\right) = 1,371;$$

➤ w podzbiorze „kobieta”:

$$p_1 = 1/5, p_2 = 2/5, p_3 = 2/5$$

$$H(\mathbf{X}_2) = -\frac{1}{5}\log_2\left(\frac{1}{5}\right) - \frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{2}{5}\log_2\left(\frac{2}{5}\right) = 1,522$$

$$H_{f=1}(\mathbf{X}) = -\sum_{j=1}^J \frac{|\mathbf{X}_j|}{|\mathbf{X}|} \cdot H(\mathbf{X}_j) = \frac{5}{10} \cdot 1,371 + \frac{5}{10} \cdot 1,522 = 1,446.$$

I ITERACJA ALGORYTMU: ZYSK

3. Zysk uzyskany dzięki podziałowi zbioru \mathbf{X} na podstawie zmiennej „płeć”:

$$\Delta H_{f=1} = H(\mathbf{X}) - H_{f=1}(\mathbf{X}) = 1,570 - 1,446 = 0,124.$$

(na tym się kończy algorytm ID3).

4. Ilość informacji, uzyskanej po podziale zbioru \mathbf{X} na $J = 2$ podzbiory względem zmiennej „płeć”:

$$H_{Jf=1} = \sum_{j=1}^J \frac{|\mathbf{X}_j|}{|\mathbf{X}|} \log_2 \left(\frac{|\mathbf{X}_j|}{|\mathbf{X}|} \right) = -\frac{5}{10} \log_2 \frac{5}{10} - \frac{5}{10} \log_2 \frac{5}{10} = 1$$

5. Zysk znormalizowany:

$$\Delta H_{f=1}^{norm} = \frac{\Delta H_f}{H_{Jf=1}} = \frac{0,124}{1} = 0,124.$$

I ITERACJA ALGORYTMU: KOLEJNE CECHY

2. Informacja, zmienna „posiadanie samochodu”, $J = 3$:

➤ w podzbiorze „0”: $p_1 = 2/3, p_2 = 1/3, p_3 = 0/3$

$$H(\mathbf{X}_1) = -\frac{2}{3}\log_2\left(\frac{2}{3}\right) - \frac{1}{3}\log_2\left(\frac{1}{3}\right) - \frac{0}{3}\log_2\left(\frac{0}{3}\right) = 0,918;$$

➤ w podzbiorze „1”: $p_1 = 2/5, p_2 = 2/5, p_3 = 1/5$

$$H(\mathbf{X}_2) = -\frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{1}{5}\log_2\left(\frac{1}{5}\right) = 1,522;$$

➤ w podzbiorze „2”: $p_1 = 0/2, p_2 = 0/2, p_3 = 2/2$

$$H(\mathbf{X}_3) = -\frac{0}{2}\log_2\left(\frac{0}{2}\right) - \frac{0}{2}\log_2\left(\frac{0}{2}\right) - \frac{2}{2}\log_2\left(\frac{2}{2}\right) = 0;$$

$$H_{f=2}(\mathbf{X}) = -\sum_{j=1}^J \frac{|\mathbf{X}_j|}{|\mathbf{X}|} \cdot H(\mathbf{X}_j) = \frac{3}{10} \cdot 0,918 + \frac{5}{10} \cdot 1,522 + \frac{2}{10} \cdot 0 = 1,036$$

I ITERACJA ALGORYTMU: KOLEJNE CECHY C.D.

3. Zysk, zmienna „posiadanie samochodu”:

$$\Delta H_{f=1} = H(\mathbf{X}) - H_{f=1}(\mathbf{X}) = 1,570 - 1,036 = 0,534.$$

4. Ilość informacji, zmienna „posiadanie samochodu”:

$$H_{Jf=2} = -\frac{3}{10} \log_2 \frac{3}{10} - \frac{5}{10} \log_2 \frac{5}{10} - \frac{2}{10} \log_2 \frac{2}{10} = 1,486$$

5. Zysk znormalizowany:

$$\Delta H_{f=2}^{norm} = \frac{\Delta H_f}{H_{Jf=2}} = \frac{0,534}{1,486} = 0,359.$$

I ITERACJA ALGORYTMU: KOLEJNE CECHY C.D.

Analogicznie obliczamy zysk znormalizowany dla podziału na podstawie zmiennej „odległość”:

$$\Delta H_{f=3}^{norm} = \frac{\Delta H_f}{H_{Jf=3}} = \frac{1,209}{1,661} = 0,728$$

oraz dla zmiennej „dochód”:

$$\Delta H_{f=4}^{norm} = \frac{\Delta H_f}{H_{Jf=4}} = \frac{0,695}{1,371} = 0,507.$$

Jak można zaobserwować, dla zmiennej:

- „płeć”: $\Delta H_{f=1}^{norm} = 0,124$;
- „posiadanie samochodu”: $\Delta H_{f=2}^{norm} = 0,359$;
- „odległość”: $\Delta H_{f=3}^{norm} = 0,728$
- „dochód”: $\Delta H_{f=4}^{norm} = 0,507$

I ITERACJA ALGORYTMU: 1 WĘZŁ

Największa wartość współczynnika zysku znormalizowanego wyznaczona jest dla zmiennej „odległość”.

Na podstawie tej zmiennej rozpoczyna się budowa drzewa decyzyjnego.

Ze względu na to, że “odległość” wnosi najwięcej informacji w procesie decyzyjnym, zmienna ta staje się pierwszym węzłem, tzw. korzeniem w strukturze SDT.

Zmienna odległość przyjmuje 3 wartości

$\text{odległość} = \{\text{krótka}, \text{średnia}, \text{duża}\}$

przez co z węzła tego powinny wychodzić 3 gałęzie.

I ITERACJA ALGORYTMU: UAKTUALNIENIE ZBIORU X

Sprowadza się to do podziału danych na podgrupy na podstawie cechy o największym zysku informacji (odległość).

	$f = 1$	$f = 2$	$f = 3$	$f = 4$	C_k
mężczyzna	0		krótka	niski	samochód
mężczyzna	1		krótka	średni	samochód
kobieta	1		krótka	średni	pociąg
kobieta	0		krótka	niski	samochód
mężczyzna	1		krótka	średni	samochód
mężczyzna	0		średnia	średni	pociąg
kobieta	1		średnia	średni	pociąg
kobieta	1		duża	wysoki	samolot
mężczyzna	2		duża	średni	samolot
kobieta	2		duża	wysoki	samolot

I ITERACJA ALGORYTMU: OBSERWACJE

Obserwacja 1: wartość **średnia** cechy “odległość” powiązana jest wyłącznie z klasą **pociąg** (i odwrotnie). Wynika stąd jednoznaczna reguła:

Jeżeli “odległość” jest **średnia**, to rodzajem transportu jest **pociąg**.

Obserwacja 2: wartość **duża** cechy “odległość” powiązana jest wyłącznie z klasą **samolot**. Daje to kolejną regułę:

Jeżeli “odległość” jest **duża**, to rodzajem transportu jest **samolot**.

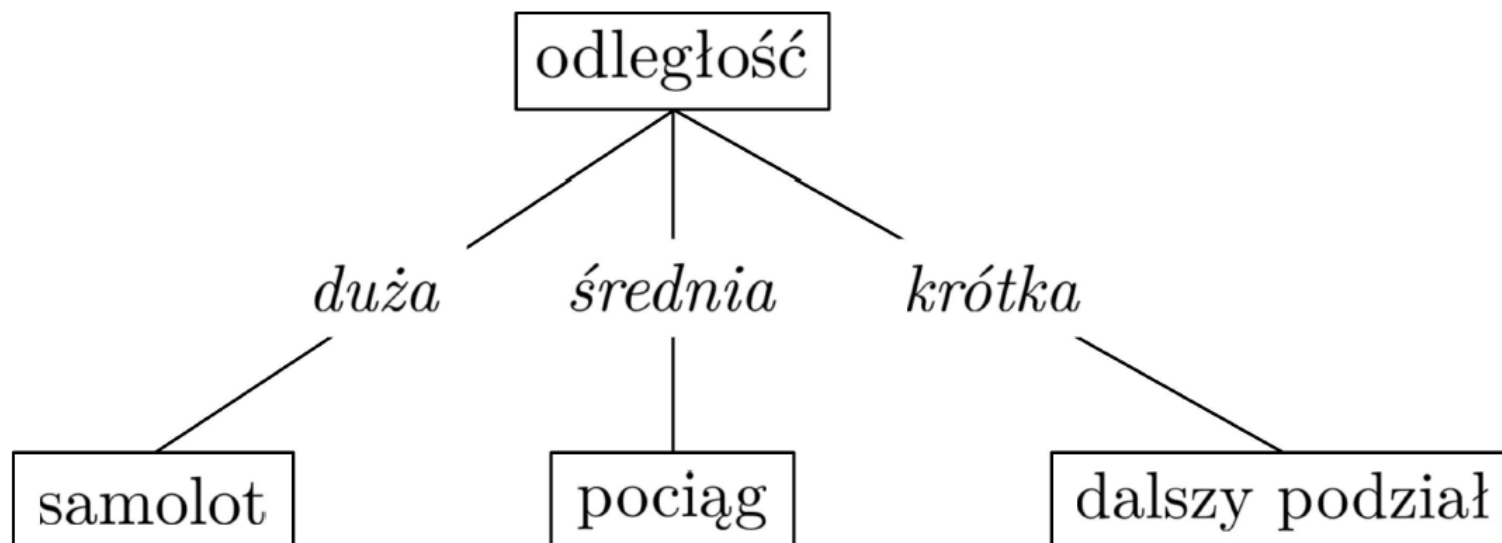
W przypadku gdy klasa powiązana jest wyłącznie z jedną wartością cechy (węzła w drzewie), klasa ta staje się węzłem końcowym - liściem.

I ITERACJA ALGORYTMU: OBSERWACJE C.D.

Obserwacja 3: wartość *krótka* cechy “odległość” powiązana jest z dwoma klasami: samochód oraz pociąg.

Konieczny jest dalszy podział danych.

Na podstawie dotychczasowych wyników można narysować pierwszy fragment struktury SDT:



II ITERACJA ALGORYTMU: DANE

Ustalenie kolejnego węzła na podstawie wybranej zmiennej –
usunięcie rekordów powiązanych z wartościami **duża** oraz
średnia zmiennej “odległość” oraz **usunięcie tej zmiennej**

Pozostał następujący zbiór danych:

$f = 1$		$f = 2$	$f = 4$	C_k
mężczyzna	0	niski	samochód	
mężczyzna	1	średni	samochód	
kobieta	1	średni	pociąg	
kobieta	0	niski	samochód	
mężczyzna	1	średni	samochód	

II ITERACJA ALGORYTMU: INFORMACJA

Średnia ilość informacji - entropia uaktualnionego zbioru **X**:

$$H(\mathbf{X}) = -\frac{4}{5}\log_2\left(\frac{4}{5}\right) - \frac{1}{5}\log_2\left(\frac{1}{5}\right) = 0,722.$$

Po analogicznych obliczeniach:

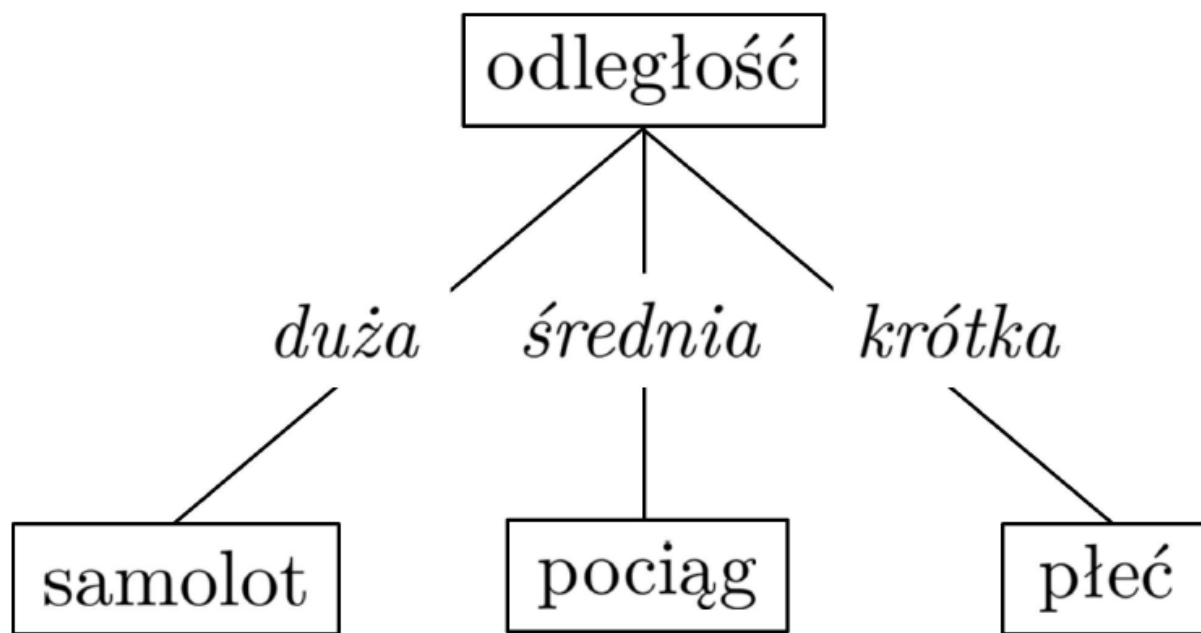
- „płeć”: $\Delta H_{f=1}^{norm} = 0,332$;
- „posiadanie samochodu”: $\Delta H_{f=2}^{norm} = 0,176$;
- „dochód”: $\Delta H_{f=4}^{norm} = 0,176$

Największa wartość współczynnika zysku wyznaczona jest dla zmiennej „płeć”, czyli zmienna ta staje się kolejnym węzłem w drzewie decyzyjnym.

II ITERACJA ALGORYTMU: 2 WĘZŁ

Zmienna płeć={kobieta, mężczyzna} , przez co z węzła powinny wychodzić 2 gałęzie.

Do struktury SDT można dodać kolejny węzeł:



II ITERACJA ALGORYTMU: UAKTUALNIENIE ZBIORU X

Kolejne kroki to uaktualnienie zbioru **X** i podział danych na podgrupy na podstawie zmiennej o największej wartości współczynnika zysku ('płeć')

$f = 1$	$f = 2$	$f = 4$	C_k	
mężczyzna	0	niski	samochód	
mężczyzna	1	średni	samochód	
mężczyzna	1	średni	samochód	
kobieta	1	średni	pociąg	
kobieta	01	niski	samochód	

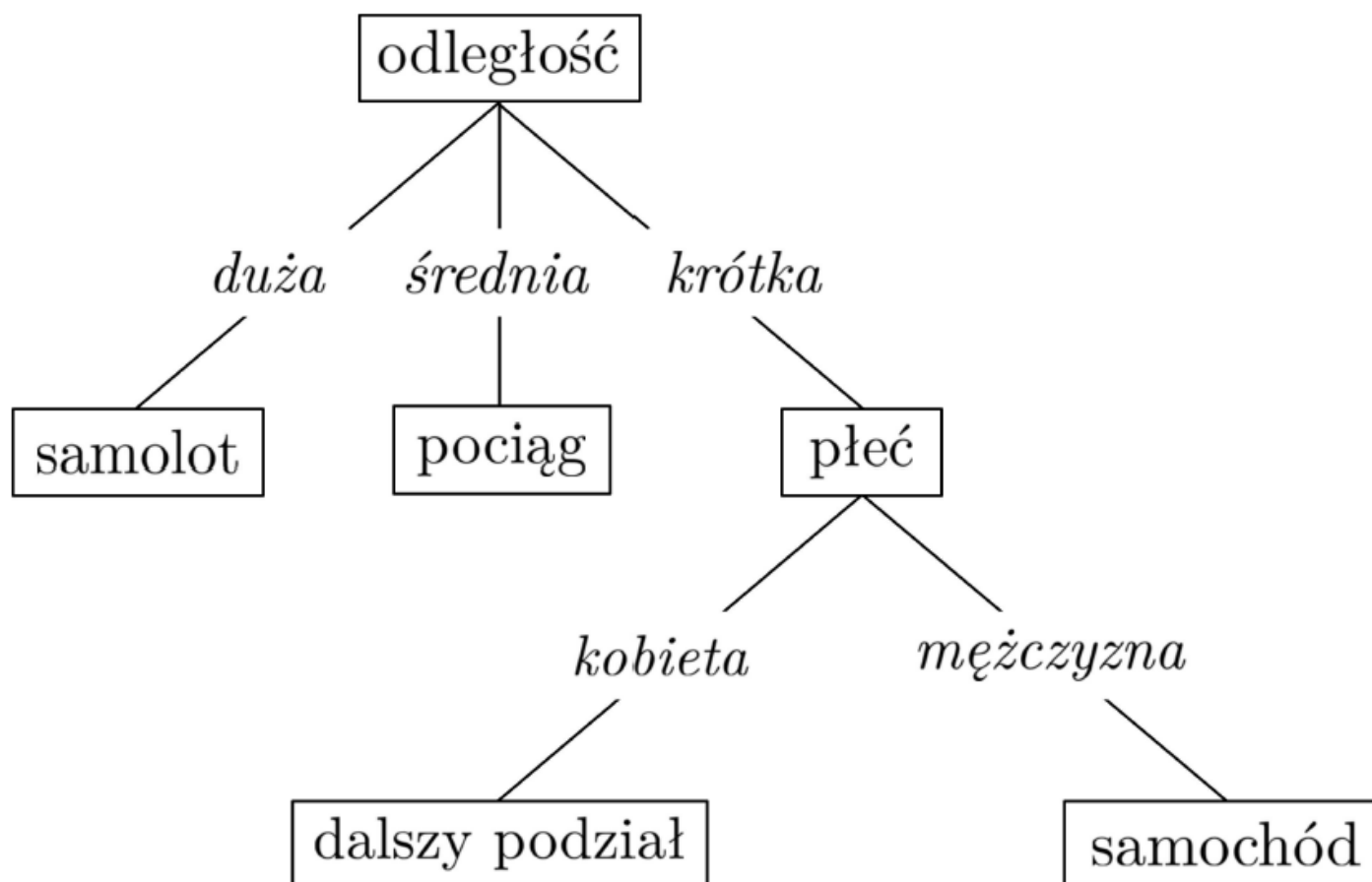
II ITERACJA ALGORYTMU: OBSERWACJE

Obserwacja 1: wartość *mężczyzna* zmiennej “płeć” powiązana jest wyłącznie z klasą *samochód* (i odwrotnie). Wynika stąd reguła: Jeżeli “płeć” jest *mężczyzna*, to rodzajem transportu jest *samochód*.

Obserwacja 2: wartość *kobieta* cechy “płeć” powiązana jest z dwoma klasami: pociąg oraz samochód.

Regułę: “Jeżeli płeć jest mężczyzna, to rodzajem transportu jest samochód” można od razu uwzględnić w strukturze SDT.

II ITERACJA ALGORYTMU: OBSERWACJE C.D.



III ITERACJA ALGORYTMU

Wszystkie rekordy posiadające wartość męczyzna dla cechy płeć można usunąć ze zbioru \mathbf{X} .

Konieczne jest również usunięcie zmiennej „płeć”.

Pozostaje następujący zbiór danych:

$f = 2$	$f = 4$	C_k
1	średni	pociąg
01	niski	samochód

Średnia ilość informacji - entropia uaktualnionego zbioru \mathbf{X} :

$$H(\mathbf{X}) = -p_1 \log_2(p_1) - p_2 \log_2(p_2) = -\frac{1}{2} \log_2\left(\frac{4}{5}\right) - \frac{1}{2} \log_2\left(\frac{1}{5}\right) = 1.$$

III ITERACJA ALGORYTMU C.D.

Po analogicznych obliczeniach:

➤ „posiadanie samochodu”: $\Delta H_{f=2}^{norm} = 1$;

➤ „dochód”: $\Delta H_{f=4}^{norm} = 1$

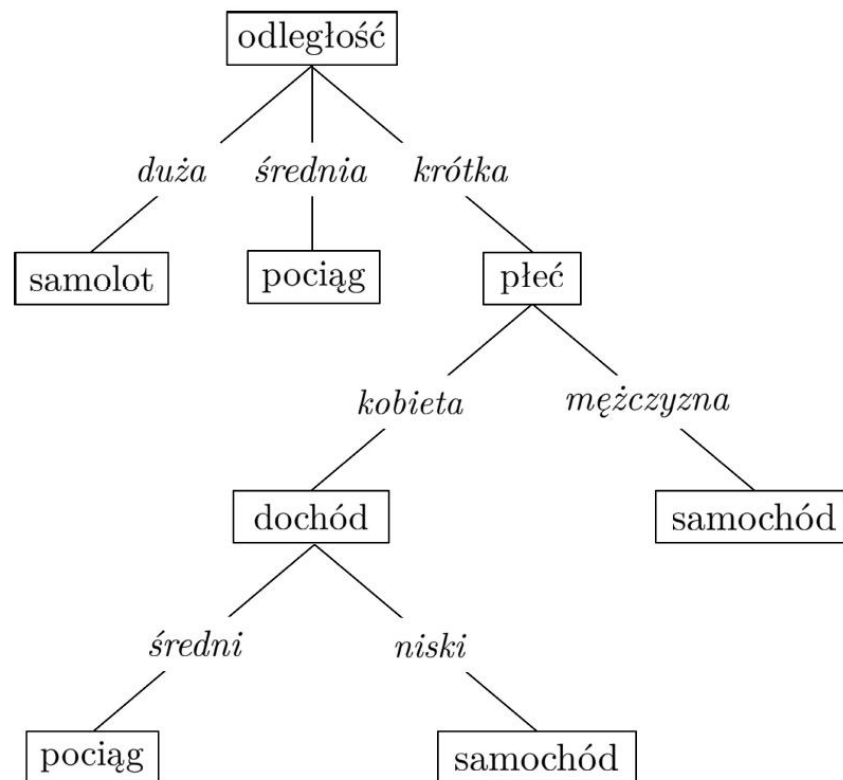
Wskaźnik zysku jest zatem równy dla obydwu zmiennych. Wynika to z faktu, że każda cecha posiada inną wartość dla poszczególnej klasy.

Zatem, wszystko jedno, która z cech będzie węzłem w SDT - można ją wybrać w sposób losowy.

III ITERACJA ALGORYTMU C.D.

Jeżeli węzłem tym będzie zmienna “dochód”, to wyjdą od niego dwie gałęzie: **średni** oraz **niski**.

Węzeł ten będzie ostatnim rozgałęzieniem w strukturze drzewa.



REGUŁY

Na podstawie stworzonego SDT można wygenerować zestaw pięciu następujących reguł:

- **Jeżeli** “odległość” = *duża* **to** *samolot*.
- **Jeżeli** “odległość” = *średnia* **to** *pociąg*.
- **Jeżeli** “odległość” = *krótka* i “płeć” = *mężczyzna* **to** *samochód*.
- **Jeżeli** “odległość” = *krótka* i “płeć” = *kobieta* i “dochód” = *średni* **to** *pociąg*.
- **Jeżeli** “odległość” = *krótka* i “płeć” = *kobieta* i dochód = *niski* **to** *samochód*.

Powyższe reguły mogą posłużyć do klasyfikacji danych wykorzystujących te same cechy.

WARTOŚCI CIĄGŁE ZMIENNYCH

W omawianym przykładzie nie uwzględniono cech, które przyjmują wartości ciągłe.

Przykładowo: zmienna “dochód” zamiast kategorii {niski, średni, wysoki} może przyjmować wartości rzeczywiste, np.: 2340 lub 4550.

Pytanie: czy liczba gałęzi dla węzła “dochód” równa będzie wówczas liczbie możliwych wartości, jakie przyjmuje ta zmienna (maksymalnie)?

Odpowiedź: można takie drzewo stworzyć, ale jego budowa będzie skomplikowana co doprowadzać będzie do niskiej predykcji.

WARTOŚCI CIĄGŁE ZMIENNYCH

Quinlan w roku 1993 zaproponował dla zbioru posortowanych rosnąco wartości danej zmiennej ciągłej $\{x_1, x_2, \dots, x_j, x_{j+1}, \dots, x_J\}$ wprowadzić wartość progową:

$$t = \frac{x_j + x_{j+1}}{2}$$

i dzielić zmienną ciągłą na dwie kategorie, np., mały ($\leq t$) i duży ($> t$).

DANE Z BRAKUJĄCYMI WARTOŚCIAMI ZMIENNYCH

W algorytmie C4.5, rozwiązanie problemu braku danych polega na przeskalowaniu wskaźnika zysku.

Przeskalowanie to bazuje na wyznaczeniu częstości występowania cechy z wartościami brakującymi '?'.
Formalnie:

Niech P oznacza część rekordów, u których istnieją '?' dla wybranej zmiennej f .

Wówczas zysk uzyskany dzięki podziałowi zbioru \mathbf{X} na podstawie cechy f korygowany jest do postaci:

$$\Delta H_f = P \cdot [H(\mathbf{X}) - H_f(\mathbf{X})].$$

PRZECINANIE DRZEW DECYZYJNYCH

Rekurencyjna metoda podziału w procesie tworzenia SDT wykonuje się dla przyjętego kryterium stopu.

Kryteria stopu najczęściej spotykane:

1. wszystkie rekordy w podzbiorze X_j należą do tej samej klasy;
2. osiągnięty jest minimalny rozmiar węzła - w podzbiorze X_j znajduje się mniej rekordów niż założony próg;
3. osiągnięta jest maksymalna głębokość drzewa (liczba poziomów węzłów idąc od korzenia w dół).

Wykonanie pełnej metody podziału może się jednak przyczynić do konstrukcji rozrośniętej struktury drzewa.

W takich wypadkach proces podziału danych można zmodyfikować, aby stworzyć prostsze drzewa, czyli wykonać przycinanie drzew.

PRZECINANIE DRZEW DECYZYJNYCH

Pytanie: **kiedy przycinać?**

Najprostsza odpowiedź: ocenić każdy podział z punktu widzenia istotności statystycznej, zysku informacji lub redukcji błędu.

Jeśli ta ocena jest poniżej przyjętego progu, odrzuca się podział względem rozpatrywanej cechy.

PRZECINANIE DRZEW DECYZYJNYCH C.D.

Istnieją dwa podejścia do przycinania struktur SDT:

1. Przycinanie na bieżąco: wstrzymanie dalszego podziału, jeśli przyczynia się on do rozbudowanej struktury.
2. Przycinanie stworzonego SDT: wybrane gałęzie usuwane są z całej struktury w celu poprawy jakości klasyfikacji.

Jak przycinać? Najczęściej poprzez odrzucenie jednego/kilku poddrzew i zastąpienie ich liśćmi.

Popularne podejście: przycinanie na podstawie błędu (ang. error based pruning):

W C4.5 przycinanie realizowane jest przez zastąpienie poddrzewa liściem albo gałęzią, która zmniejsza błąd klasyfikacji.

ZALETY DRZEW DECYZYJNYCH

- ich graficzna prezentacja jest przejrzysta i łatwa w zrozumieniu;
- mogą posłużyć do klasyfikacji i regresji;
- generują zbiór łatwo interpretowanych reguł (biała skrzynka);
- atrybuty danych użytych do budowy drzewa mogą przyjmować wartości dowolnego typu (numeryczne i kategoryczne);
- ze względu na prostotę obliczeń, rozwiązanie znajduwane jest szybko, nawet dla dużej liczby danych.

WADY DRZEW DECYZYJNYCH

- są podatne na przetrenowanie - w przypadku danych nieliniowych posiadają bardzo skomplikowaną strukturę;
- algorytmy generowania drzew decyzyjnych bazują na heurystykach - nie zawsze znajdowane jest optymalne rozwiązanie podczas znajdowania węzłów.