K-means

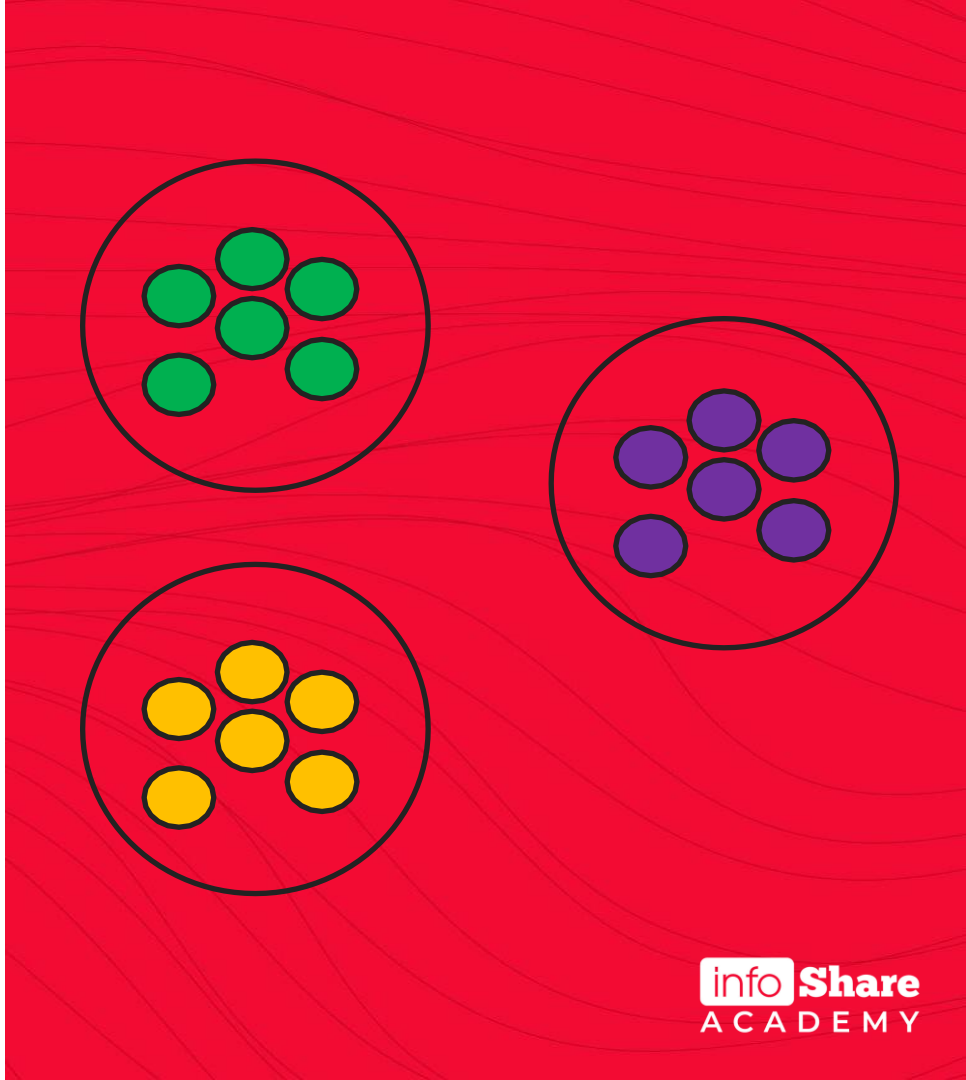
HELLO

Patryk Krysiński

- Absolwent Matematyki Finansowej na Politechnice Gdańskiej oraz Analizy Danych w Szkole Głównej Handlowej.
- Od ponad 6 lat zajmuję się analizą danych, data science i machine learning.
- Dotychczasowe dziedziny pracy: Bankowość, ryzyko kredytowe, windykacja i sprzedaż.
- Obecnie Data Science Practice Lead w Psignite.

Roadmapa

- Uczenie nienadzorowane
- Klasteryzacja
- K-Means
  - Algorytm
  - Ograniczenia metody
  - Wybór parametrów
  - Implementacja w sklearn



Podstawowa klasyfikacja metod uczenia

https://vas3k.com/blog/machine_learning/

Uczenie nadzorowane i nienadzorowane

- uczenie dzielimy na z nadzorem i bez nadzoru

każdy przykład ma określoną etykietkę (label, target)

$$p(y|\vec{x})$$

	sepal l	sepal w	petal l	petal w
0	5.1	3.5	1.4	0.2
1	4.9	3.0	1.4	0.2
2	4.7	3.2	1.3	0.2
3	4.6	3.1	1.5	0.2
4	5.0	3.6	1.4	0.2
...	...	...	...	...
145	6.7	3.0	5.2	2.3
146	6.3	2.5	5.0	1.9
147	6.5	3.0	5.2	2.0
148	6.2	3.4	5.4	2.3
149	5.9	3.0	5.1	1.8

	target
0	0
1	0
2	0
3	0
4	0
...	...
145	2
146	2
147	2
148	2
149	2

przykłady są tylko zbiorami cech; algorytm "uczy się" rozkładu prawdopodobieństwa, który wygenerował zbiór

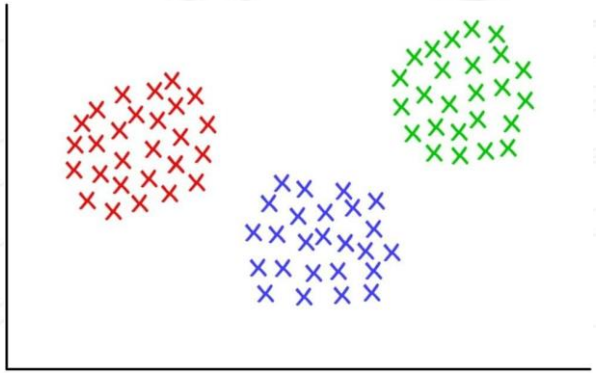
$$p(\vec{x})$$

	sepal l	sepal w	petal l	petal w
0	5.1	3.5	1.4	0.2
1	4.9	3.0	1.4	0.2
2	4.7	3.2	1.3	0.2
3	4.6	3.1	1.5	0.2
4	5.0	3.6	1.4	0.2
...	...	...	...	...
145	6.7	3.0	5.2	2.3
146	6.3	2.5	5.0	1.9
147	6.5	3.0	5.2	2.0
148	6.2	3.4	5.4	2.3
149	5.9	3.0	5.1	1.8

Klasteryzacja

Jest to proces dzielenia danych na grupy (klastry) bazując na wzorcach w danych.

Podział odbywa się w taki sposób, by podobne obiekty należały do tej samej grupy, a różne obiekty do różnych.



Przykłady zastosowań

- Segmentacja użytkowników
  - dedykowana komunikacja (marketing)
- Segmentacja grup klientów
  - dedykowane produkty
- Segmentacja obrazów
  - diagnostyka obrazowa w medycynie
- Grupowanie dokumentów
  - automatyczny podział podobnych do siebie dokumentów
- Grupowanie produktów
  - kategorie w sklepach internetowych
- Systemy rekomendacyjne
  - podobni do Ciebie kupili
  - podobne produkty

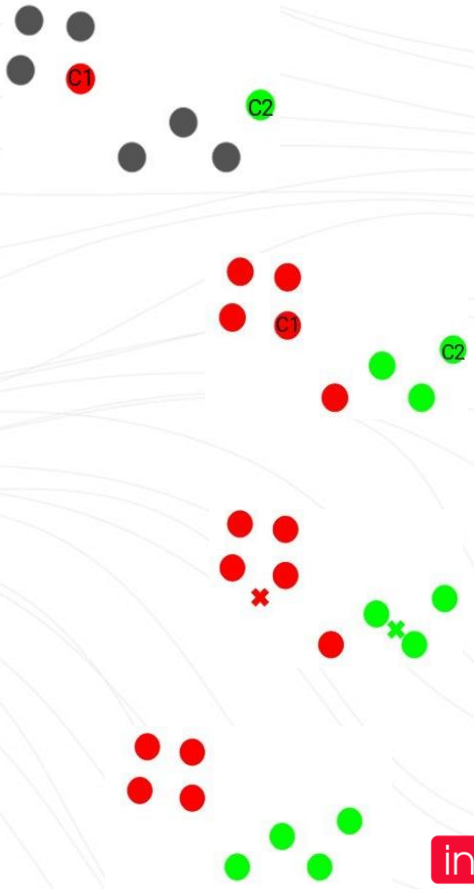
K-means

Algorytm k-średnich często jest algorytmem pierwszego wyboru ze względu na jego prostotę i łatwość interpretacji.

Interaktywna wizualizacja

Algorytm K-means

1. Wybierz liczbę klastrów K
2. Wybierz k losowych punktów z danych jako tzw. centroidy
3. Przypisz każdą obserwację do najbliższego centroidu (klastra)
4. Przelicz na nowo centroidy klastrów (środki)
5. Powtórz kroki 3 i 4 dla nowych centroidów do czasu aż:
    1. centroidy dla nowoutworzonych klastrów się nie zmieniają
    2. punkty pozostają w tym samym klastrze
    3. osiągnięto maksymalną liczbę iteracji



K-means w sklearn

Trenowanie modelu:

```
from sklearn.cluster import KMeans
model = KMeans(**parametry)
model.fit(dane_treningowe)
```

Aplikacja modelu:

```
model.predict(nowy_data_point)
```

- n_clusters: liczba klastrów, które zakładamy, że mamy w danych
- n_init: od początkowego położenia centroidów dużo zależy, więc chcemy wystartować algorytm n_init razy, żeby zniwelować wpływ losu. Najczęściej wystarczy 10 defaultowych runów.
- init: wybieramy “k-means++”, bo szybciej pozwala znaleźć optymalny wynik
- tol: jak bardzo powinny zmieniać swoje położenie środki klastrów, żeby algorytm dalej szukał optymalnych centroidów. Najczęściej zostawiamy default.
- max_iter: liczba iteracji algorytmów. W praktyce można ustawić z zapasem (np.: 1000 iteracji), bo zakładamy, że algorytm zostanie zatrzymany wcześniej przez tol.
- algorithm: najczęściej zostawiamy default. “full” jest tym, co implementowaliśmy od zera.
- random_state: środki klastrów są generowane losowo, więc w celu zapewnienia reprodukowalności wyników należy ustawić ziarno losowe.

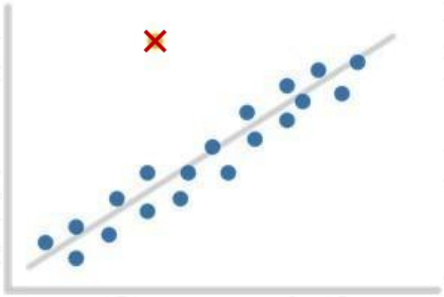
Normalizacja i standaryzacja

Standardisation (Z-score Normalization)	Max-Min Normalization
$x_{\text{stand}} = \frac{x - \text{mean}(x)}{\text{standard deviation } (x)}$	$x_{\text{norm}} = \frac{x - \min(x)}{\max(x) - \min(x)}$

Obserwacje odstające (outliery)

Dlaczego mogą być zagrożeniem?

- Zaburzają statystyki klasyczne (mean, max, min, std)
- Problem przy normalizacji / standaryzacji
- Negatywny wpływ na parametry modelu

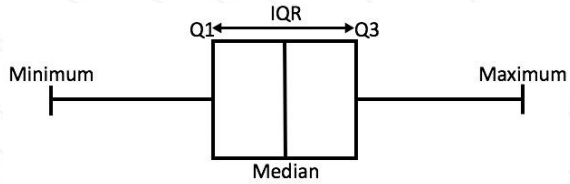


Co robić?

- Najłatwiej: usunąć. Szczególnie, gdy jest ich mało.
- Nie warto tego robić bezmyślnie. Warto zastanowić się biznesowo jaki charakter mają te obserwacje.

Jak znaleźć?

- Jedna z najpopularniejszych metod: 1.5IQR
- Inna bardziej zaawansowana metoda: IsolationForest



A box plot from source

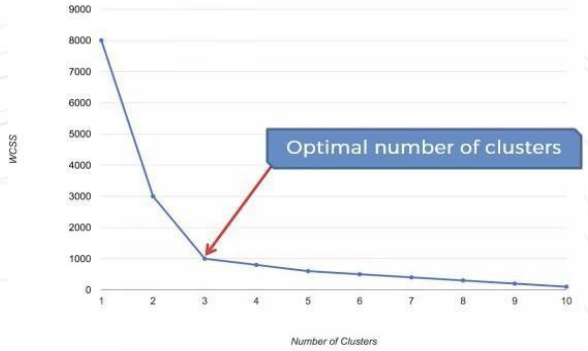
Założenia i ograniczenia

- metoda wrażliwa na początkową inicjalizację klastrów - losowy wybór punktów startowych
- metoda wrażliwa na różne skale danych
- różne wielkości klastrów i różne gęstości w klastrach

Wyznaczanie optymalnej liczby klastrów

Elbow Method

Nie ma jednoznacznej odpowiedzi, jaka liczba klastrów jest optymalna. Jedną z metod do ustalenie tego jest wykres sumy wariancji klastrów (inercji). Inercja wraz ze wzrostem podziału na coraz większą liczbę klastrów będzie malała.
Dzieje się tak, ponieważ klastry stają się coraz mniejsze.
Jednak optymalnej liczby klastrów poszukujemy w miejscu, gdzie suma wariancji przestaje gwałtownie maleć, a dokładanie kolejnych klastrów nie wprowadza już dużej poprawy.



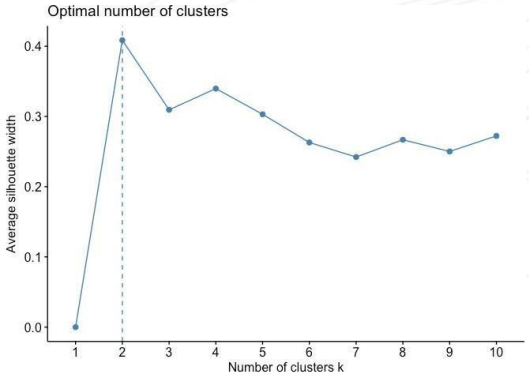
Wyznaczanie optymalnej liczby klastrów

Silhouette score

Silhouette score zawiera dwa składniki:

- Średni dystans od obserwacji do innych obserwacji w tym klastrze
- Średni dystans od obserwacji do obserwacji w innym najbliższym klastrze

Wartość score jest w przedziale [-1, 1], gdzie 1 oznacza, że obserwacja została przypisana do prawidłowego klastra.
Wybierana jest taka liczba klastrów, która daje najwyższy średni silhouette score.



$$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i, i \neq j} d(i,j)$$

średnia odległość punktu i wewnątrz klastra

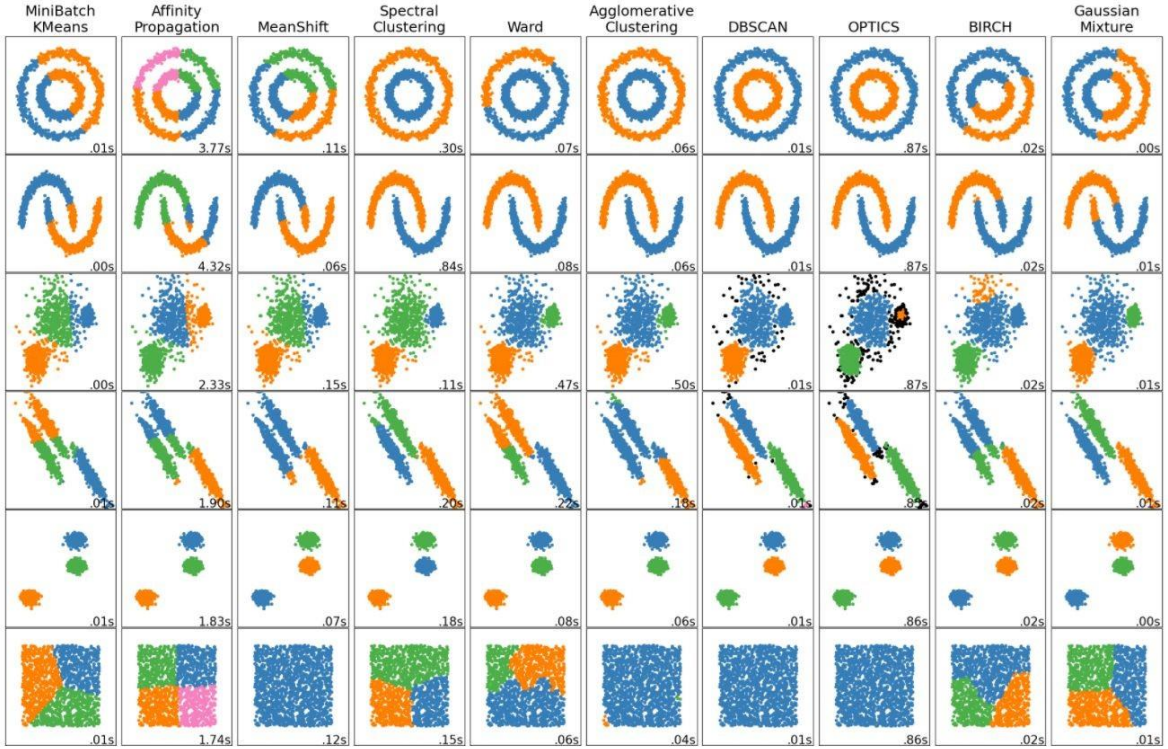
$$b(i) = \min_{k \neq i} \frac{1}{|C_k|} \sum_{j \in C_k} d(i,j)$$

średnia odległość punktu i do innego klastra

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \text{ if } |C_i| > 1$$

infoShare ACADEMY

Inne metody clusteringu



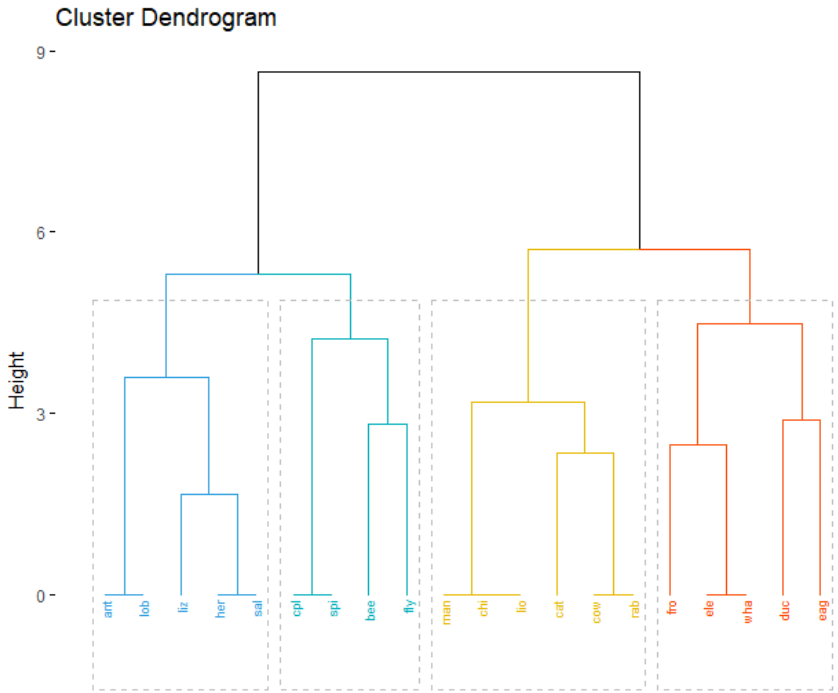
A comparison of the clustering algorithms in scikit-learn

Metody hierarchiczne

Grupa metod tworząca klasy na podstawie hierarchii elementów, czyli szukając podobieństw.
Każda obserwacja składająca się z n-elementów tworzy punkt w n-wymiarowej przestrzeni.

Metody aglomeracyjne – Początkowo każda obserwacja jest osobnym klastrem, w kolejnych krokach klastry są łączone.
Metody deglomeracyjne – Początkowo wszystkie obserwacje należą do jednego klastra, następnie są rozdzielane.

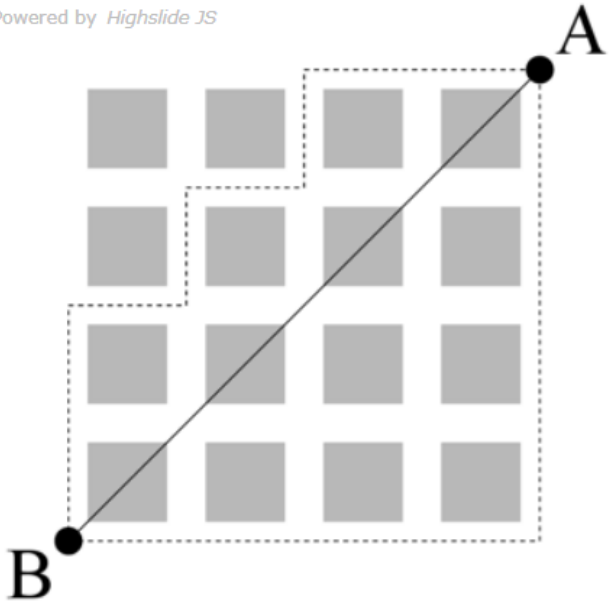
Dendogram



Źródło: https://pl.wikipedia.org/wiki/Grupowanie_hierarchiczne#/media/Plik:Dendrogram3.png

Metryki odległości

Nazwa	Wzór
Odległość Euklidesowa	$\|a-b\|_2 = \sqrt{\sum_i (a_i - b_i)^2}$
Kwadratowa Odległość Euklidesowa	$\|a-b\|_2^2 = \sum_i (a_i - b_i)^2$
Odległość Manhattan	$\|a-b\|_1 = \sum_i \|a_i - b_i\|$
Maksymalna Odległość	$\|a-b\|_\infty = \max_i \|a_i - b_i\|$



Źródła:
https://www.deltami.edu.pl/temat/matematyka/geometria/2019/03/25/Przestrzen_metryczna/
https://pl.wikipedia.org/wiki/Grupowanie_hierarchiczne

Dzięki!