KNN (k nearest neighbours)

HELLO

Katarzyna Zdon

Senior Data Scientist

katarzynazdon@gmail.com

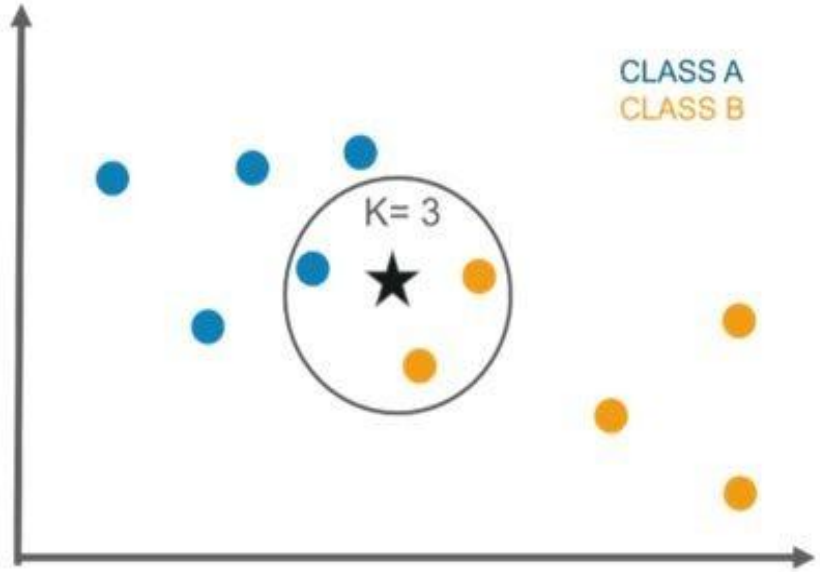
https://www.linkedin.com/in/katarzyna-zdon/

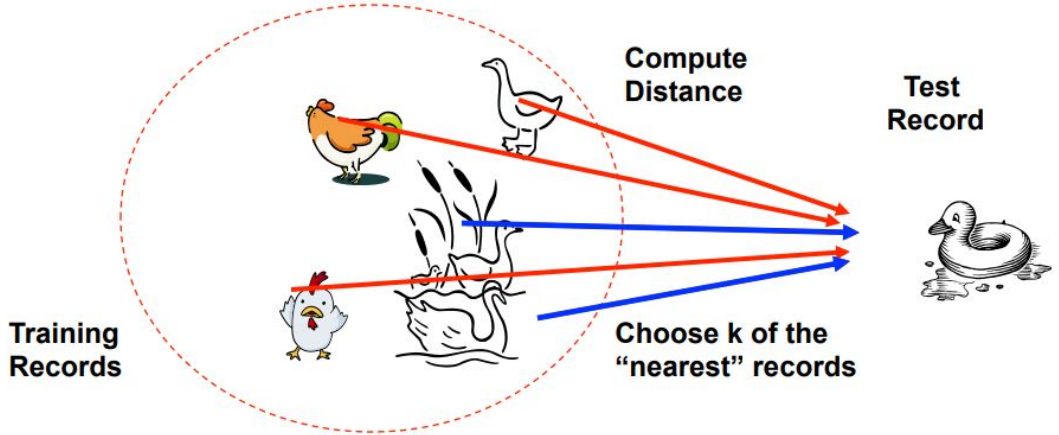
infoShare ACADEMY

O czym będzie mowa?

Intuicja tzw. Nearest Neighbor Classifiers (bardzo prosta):

- "If it walks like a duck, quacks like a duck, then it's probably a duck"





Roadmapa

1. **Przestrzeń Metryczna**, czyli jak mierzyć odległość
   a. Definicje
   b. Rodzaje metryk

2. **Normalizacja i standaryzacja**, czyli jak poradzić sobie z różnymi wielkościami i jednostkami zmiennych

3. Metoda klasyfikacji **KNN**

Przestrzeń metryczna

Niech **X** będzie niepustym zbiorem, np. $\mathbb{R}$ (oś liczbowa), $\mathbb{R}^2$(układ współrzędnych na płaszczyźnie), $\mathbb{R}^3$(układ współrzędnych w przestrzeni)
Metrykę w zbiorze **X** nazywamy funkcję $d: X \times X \to [0,\infty)$,
spełniającą dla dowolnych elementów a, b, c ze zbioru **X** następujące warunki:

- Identyczność $d(a,b) = 0 \Leftrightarrow a = b$
- Symetria $d(a,b) = d(b,a)$
- Nierówność trójkąta $d(a,b) \leq d(a,c) + d(c,b)$

Mówimy wtedy, że para **(X, d)** jest **przestrzenią metryczną**

infoShare ACADEMY

Pojęcie kuli

Kula w przestrzeni metrycznej **(*X, d*)** o środku w punkcie ***O*** i promieniu ***r*** to zbiór wszystkich elementów, których odległość od środka jest mniejsza od długości promienia. Gdy dodamy do niej zbiór wszystkich punktów odległych dokładnie o **r**, otrzymamy kulę domkniętą.

Przykładowo kulą w $\mathbb{R}$ może być odcinek, w $\mathbb{R}^2$ koło, a w $\mathbb{R}^3$ kula.

infoShare ACADEMY

Przykłady popularnych metryk

Metryka euklidesowa

Metryka euklidesowa – jest to „naturalny” sposób mierzenia odległości w przestrzeniach.

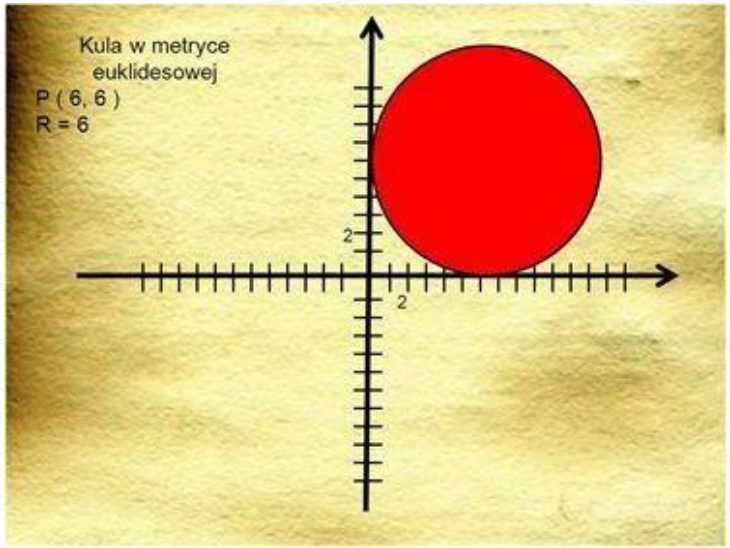
$$d_e(x, y) = \sqrt{(y_1 - x_1)^2 + \ldots + (y_n - x_n)^2}$$

Kula w tej przestrzeni to wszystkie takie punkty x spełniające nierówność:

$$d_e(x, x_0) \leq r$$

W tym przypadku to:

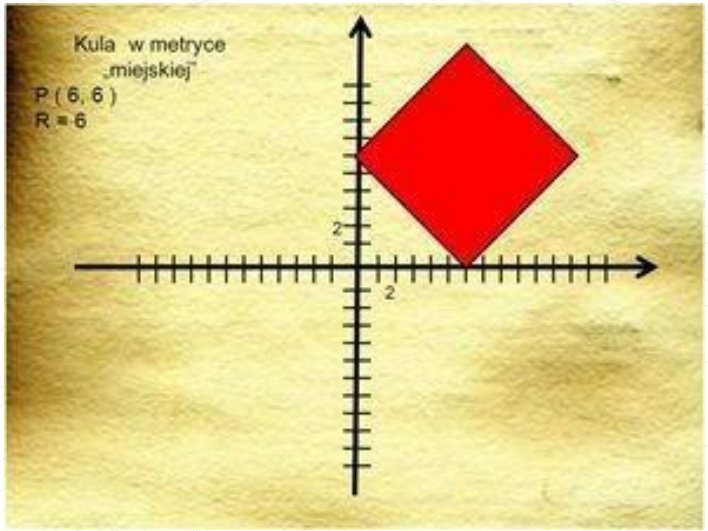
$$(x - x_0)^2 + (y - y_0)^2 + \ldots \leq r^2$$



infoShare ACADEMY

Metryka taksówkowa

Zwana również **Manhattan** lub **miejską**. Między dwoma punktami poruszamy się tylko prosto wschód-zachód i północ-południe.



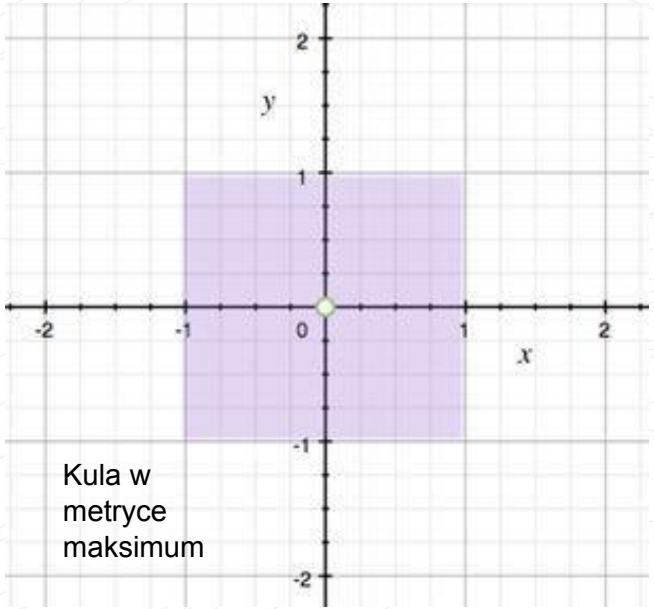
$$d_e(x,y) = |x_1 - y_1| + \ldots + |x_n - y_n|$$

$$|x_1 - x_2| + |y_1 - y_2| \leqslant r.$$

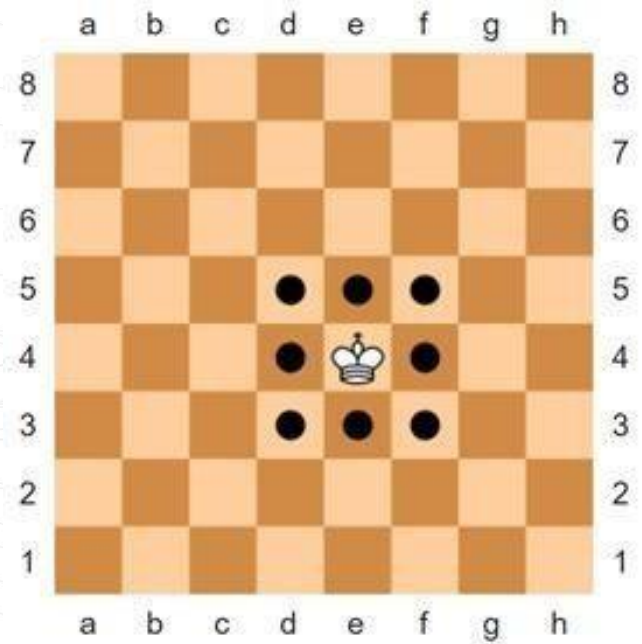
infoShare ACADEMY

Metryka maksimum

Zwana również **Czebyszewa**, **nieskończoność**, **szachową**.



Kula w metryce maksimum



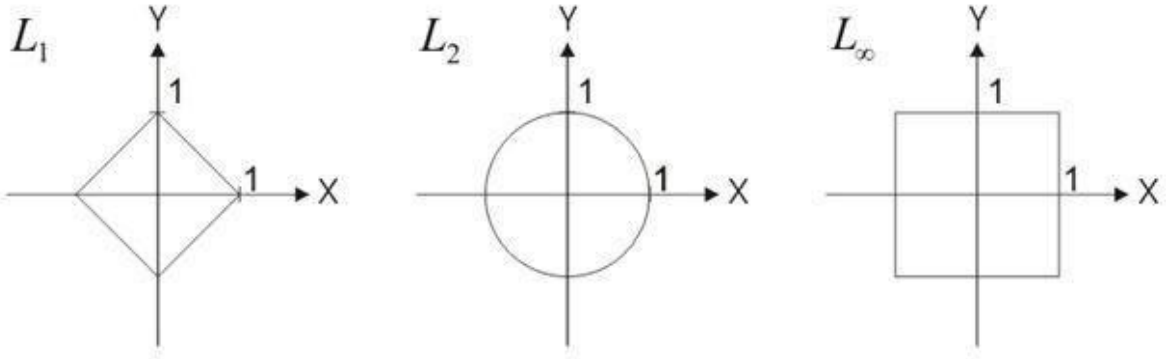
$$d_e(x,y) = \max_{k=1,..n} |x_k - y_k|$$

infoShare ACADEMY

Metryka Minkowskiego

Jest to uogólniona miara między punktami w przestrzeni euklidesowej. Nazywana również metryką L_m.

$$L_m(x,y) = \left(\sum_{i=1}^{n} |x_i - y_i|^m \right)^{\frac{1}{m}}$$



01.Distance_Normalization.ipynb

infoShare ACADEMY

Inne metryki

- Odległość na powierzchni kuli (haversine) - używana do mierzenia odległości punktów na powierzchni ziemi

  https://en.wikipedia.org/wiki/Haversine_formula

- Odległość Levenshteina - opisuje ile przekształceń potrzeba by z jednego napisu otrzymać inny. Używana przy porównywaniu słów, tekstów.

  https://en.wikipedia.org/wiki/Levenshtein_distance

info Share ACADEMY

Normalizacja i standaryzacja

Normalizacja i standaryzacja

Normalizacja to proces wstępnego przygotowania danych, aby możliwe było porównywanie zmiennych o różnych wielkościach wyrażonych w różnych jednostkach.

Najczęściej stosowanymi przekształceniami są:

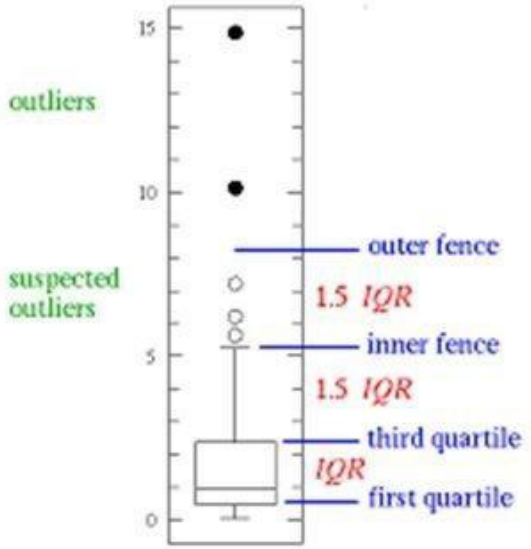
- Standaryzacja Z (Z-score) $$z = \frac{x - \mu}{\sigma}$$
- Normalizacja Min-max $$z = \frac{x - \min x}{\max x - \min x}$$

infoShare ACADEMY

Wartości odstające (outliers)

Często w danych występują wartości, które znacznie odstają od większości obserwacji.
Przy standaryzacji, takie wartości mogą zaburzyć statystyki (std, avg, min, max), na podstawie których będziemy skalować zbiór danych.

Również takie wartości mogą istotnie wpłynąć na parametry modelu. W celu zabezpieczenia się przed błędami, często usuwamy wartości odstające z danych.



infoShare ACADEMY

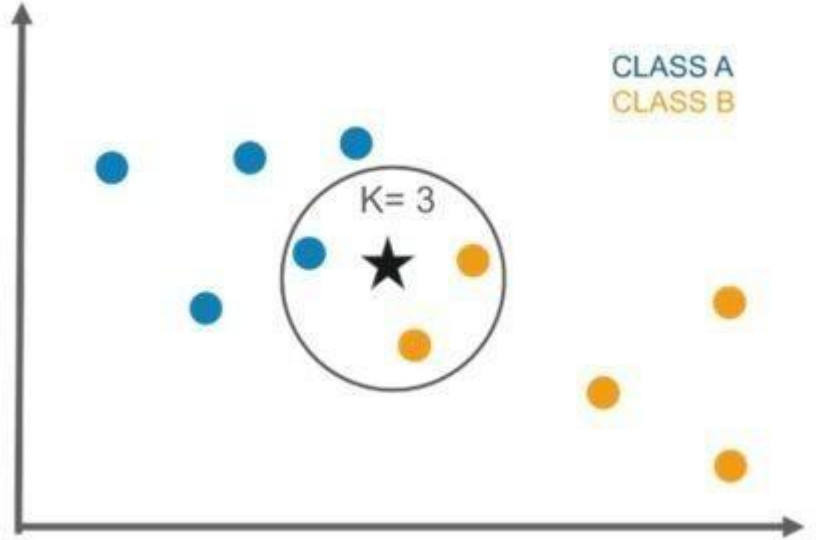
KNN - Metoda K najbliższych sąsiadów (K nearest neighbors)

K najbliższych sąsiadów

KNN jest metodą uczenia maszynowego z nadzorem.

Polega na wyszukaniu obserwacji będących w najbliższym sąsiedztwie w przestrzeni.

Może być używana zarówno do klasyfikacji jak i regresji.

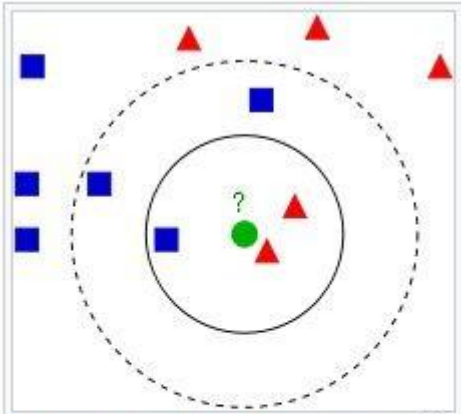


infoShare ACADEMY

Algorytm KNN

1. Zapamiętujemy położenie wszystkich punktów w zbiorze uczącym.
2. Dla obserwacji, której dokonujemy predykcji wyliczamy odległości do wszystkich punktów ze zbioru uczącego.
3. Wybieramy K obserwacji znajdujących się najbliżej tej obserwacji.
4. W przypadku klasyfikacji predykcją będzie klasa najczęściej występująca
5. W przypadku regresji, będzie to średnia wartość zmiennej zależnej z K najbliższy obserwacji.

Możemy też opcjonalnie również użyć odległości jako wag.



W przypadku k=3 (mniejszy okrąg), zielona kropka zostanie zakwalifikowana do czerwonych trójkątów. W przypadku k=5 (większy okrąg) – do niebieskich kwadratów

02.KNN.ipynb

infoShare ACADEMY

Rekomendacje

Algorytm ten stosuje się nie tylko do regresji czy klasyfikacji - dzięki niemu możemy znaleźć najbliższe sąsiedztwo w przestrzeni wielowymiarowej.

Można zastosować go do różnego rodzaju rekomendacji.

Inni klienci oglądali również ...

Podobne oferty ...

Rekomendowane dla Ciebie ...

Zobacz również ...

info Share ACADEMY

THANK YOU FOR YOUR ATTENTION

infoShareAcademy.com