Inżynieria Mechatroniczna

Podstawy Sztucznej Inteligencji i Uczenia Głębokiego:

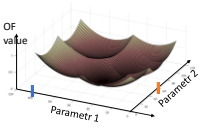
3: Uczenie (maszynowe) na podstawie danych

Ziemowit Dworakowski

AGH w Krakowie

1

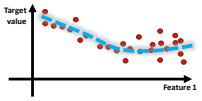
Co już wiemy?

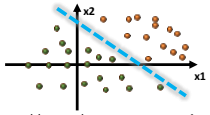


Jak dotąd, wiemy jak znaleźć współrzędne minimum globalnego funkcji celu

Wiemy czym jest regresja i jak zaprojektować liniowe i nieliniowe modele regresyjne

Wiemy czym jest klasyfikacja i potrafimy zaproponować liniowy model klasyfikacyjny





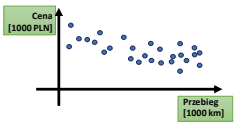
*Dzisiaj nieco **uogólnimy**, spojrzymy na te problemy od strony **przestrzeni cech** i poznamy kilka **nowych narzędzi i pojęć** które przydadzą się po drodze*

2

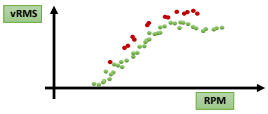
Przestrzeń parametrów a przestrzeń cech

Jeśli jakąś wartość **mierzymy** (lub z góry otrzymujemy) – jest to **cecha (ang. feature)**. Jeśli mamy nad nią kontrolę i aktywnie poszukujemy jej konkretnej wartości - jest to **parametr (ang. parameter).**

Regresja:
„Podaj spodziewaną cenę samochodu na podstawie przebiegu"



Klasyfikacja:
„Podaj spodziewany stan turbiny wiatrowej na podstawie pomiarów RPM i vRMS"



3

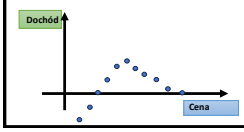
Przestrzeń parametrów a przestrzeń cech

Jeśli jakąś wartość **mierzymy** (lub z góry otrzymujemy) – jest to **cecha (ang. feature)**. Jeśli mamy nad nią kontrolę i aktywnie poszukujemy jej konkretnej wartości - jest to **parametr (ang. parameter).**

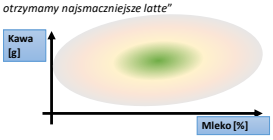
Dana zmienna może być parametrem lub cechą – w zależności od kontekstu

Optymalizację wykonujemy w przestrzeni parametrów. Regresję i klasyfikację w przestrzeni cech

Optymalizacja:
Znajdź taką cenę za samochód, by zmaksymalizować dochód:

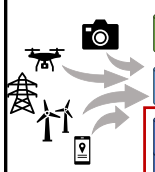


Optymalizacja:
„Podaj dla jakich proporcji kawy i mleka otrzymamy najsmaczniejsze latte"



4

Organizacja danych



Te podzbiory muszą być rozłączne!

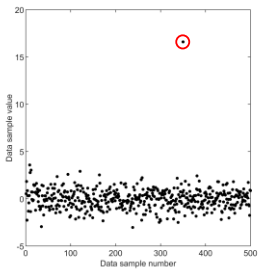
Na podstawie podzbioru **testowego** oceniamy spodziewaną docelową skuteczność metod

Na podstawie podzbioru **walidacyjnego** konfigurujemy modele i sprawdzamy czy wzorce są ogólne

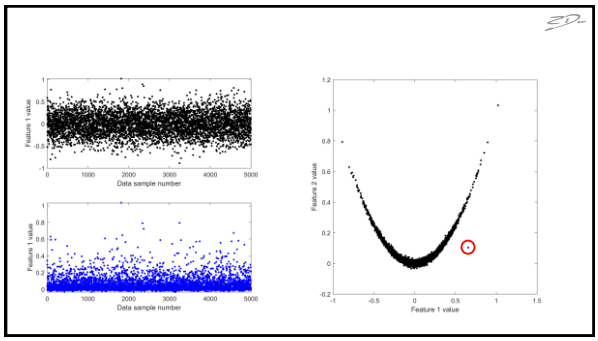
Na podstawie podzbioru **treningowego** uczymy się rozpoznawania wzorców w danych

Dzisiaj skupimy się wyłącznie na tym początkowym kroku, finalną konfigurację i weryfikację zostawiając na kolejne okazje...

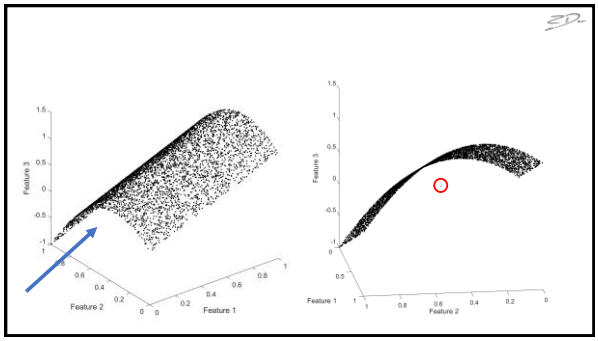
5



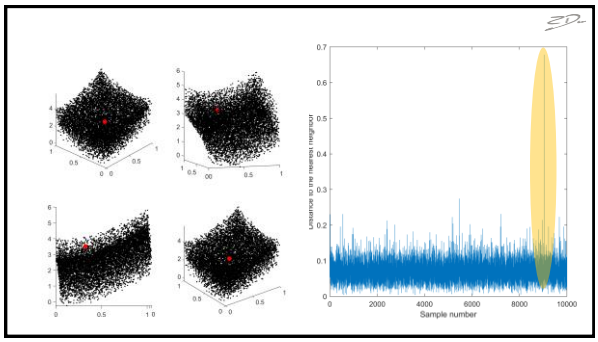
6


Feature 1 value
Data sample number
Feature 1 value
Data sample number
Feature 2 value
Feature 1 value

7


Feature 3
Feature 2
Feature 1
Feature 3
Feature 1
Feature 2

8


Sample number

9

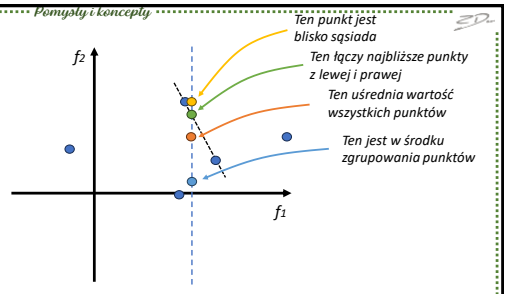
Wizualizacje danych mogą być mylące…

Więc – chcemy zaprojektować takie metody reprezentowania danych, by być w stanie automatycznie rozpoznawać wzorce w wielowymiarowej przestrzeni cech

- *Zbudować reguły określające gdzie można się spodziewać obecności danych*
- *Zorganizować dane poprzez rozpoznanie zależności między nimi i ew. przynależności do podkategorii*

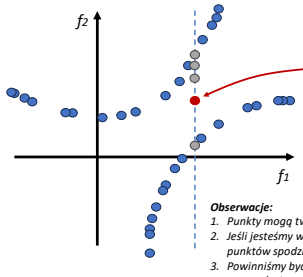
10

Pomysły i koncepty



11

Pomysły i koncepty



*Ten **wciąż** jest w środku wszystkich punktów. A dodatkowo jest dokładnie pomiędzy dwójką najbliższych sąsiadów w przestrzeni cech ORAZ przestrzeni decyzyjnej*

Więc dlaczego nie wygląda jak dobry wybór?

Obserwacje:

1. *Punkty mogą tworzyć makro-struktury w przestrzeni*
2. *Jeśli jesteśmy w stanie przewidywać gdzie można się punktów spodziewać a gdzie nie – to chyba dobrze?*
3. *Powinniśmy być w stanie aktualizować nasz model wraz z dostępem do nowych danych*

12

Pomysły i koncepty

uczenie się z danych

Koncepcja „dystansów w przestrzeni”

- *Im bliżej są punkty, tym mocniej są powiązane*
- *Przestrzeń cech przejawia różne właściwości w różnych obszarach*
- *Przestrzeń cech powinna być traktowana jak **Model***

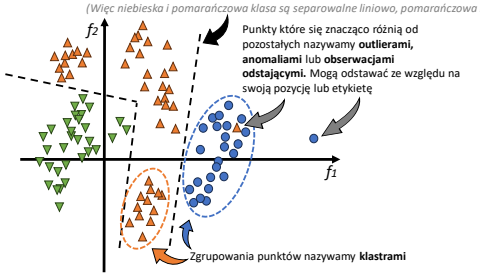
Koncepcja „modelu”

- *Relacje punktów są określone przez wzorce do których należą*
- *Identyfikacja wzorców (reguł decydujących o punktach) pozwala na lepsze wnioskowanie*
- ***Model** jest konstruktem umieszczonym w przestrzeni cech*

13

Jeśli możemy rozdzielić dwie klasy za pomocą linii (lub hiperpłaszczyzny), mówimy, że klasy są **separowalne liniowo.**W przeciwnym wypadku są **nieseparowalne liniowo**

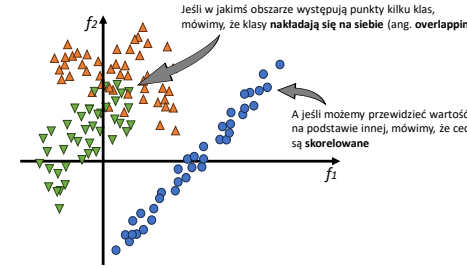
(Więc niebieska i pomarańczowa klasa są separowalne liniowo, pomarańczowa i zielona nie)



Punkty które się znacząco różnią od pozostałych nazywamy **outlierami, anomaliami** lub **obserwacjami odstającymi.** Mogą odstawać ze względu na swoją pozycję lub etykietę

Zgrupowania punktów nazywamy **klastrami**

14

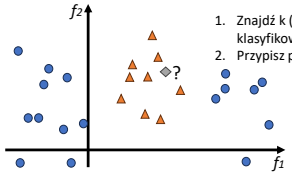


Jeśli w jakimś obszarze występują punkty kilku klas, mówimy, że klasy **nakładają się na siebie** (ang. **overlapping)**

A jeśli możemy przewidzieć wartość jednej cechy na podstawie innej, mówimy, że cechy lub dane są **skorelowane**

15

Metoda k-najbliższych sąsiadów (k Nearest Neighbor classifier, kNN)



1. Znajdź k (nieparzyste) najbliższych sąsiadów klasyfikowanego punktu
2. Przypisz punktowi najpopularniejszą etykietę

Jak metoda zareaguje na anomalie?

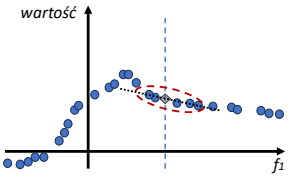
k powinno być >= 3

+ Intuicyjna
+ Prosta w implementacji
+ Prosta w konfiguracji

- Wymaga dużo pamięci
- Wymaga dużo mocy obliczeniowej (powolna)
- Źle skaluje się w wysokowymiarową przestrzeń cech

16

Regresja z zastosowaniem metody kNN

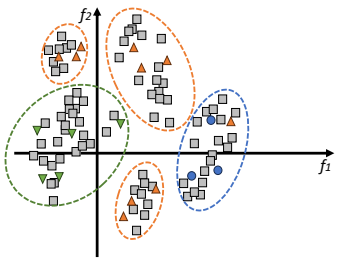


... znamy już coś bardzo podobnego!

(Prawie dokładnie tak samo działa estymator lokalnie liniowy)

17

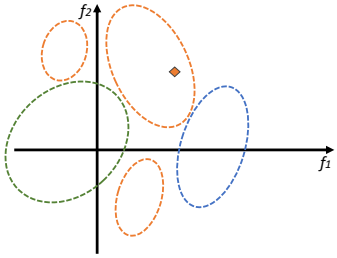
Rozwiązania bazujące na klasteryzacji



1. Zacznijmy od nieoetykietowanych danych *(to metoda **nienadzorowana**)*
2. Znajdźmy klastry danych
3. „Pokolorujmy" klastry na podstawie wybranych oetykietowanych punktów *(np. wybierając większościową etykietę)*

18

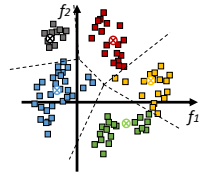
Rozwiązania bazujące na klasteryzacji



1. Zacznijmy od nieoetykietowanych danych *(to metoda **nienadzorowana**)*
2. **Znajdźmy klastry danych**
3. „Pokolorujmy" klastry na podstawie wybranych oetykietowanych punktów *(np. wybierając większościową etykietę)*
4. Dla nowych danych sprawdźmy do których klastrów należą punkty i klasyfikujmy je odpowiednio

19

Algorytm centroidów (K-means clustering, kMC)



Założymy ilość klastrów, wybierzmy ich centroidy losowo

↓

Przypiszmy punkty do klastrów bazując na ich dystansie do centroidu

↓

Przeliczmy położenia centroidów na podstawie przypisanych do nich punktów

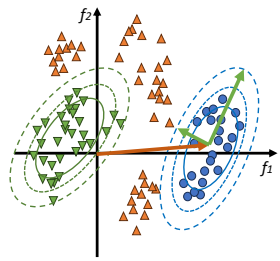
↓

Powtarzajmy, dopóki położenia centroidów się zmieniają

+ Intuicyjny
+ Jedna z najprostszych metod klasteryzacji

- Musimy wiedzieć, ile klastrów jest w danych
- Granice klastrów nie mają dobrego uzasadnienia
- Metoda źle skaluje się w przestrzeń wielowymiarową

20



- Większość rzeczywistych rozkładów danych przypomina rozkład gaussa (normalny)

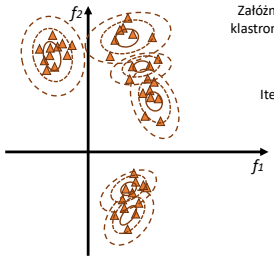
*- Więc często rozsądne jest założenie z góry, iż **rozkład danych jest normalny***

*- Wtedy możemy dopasować do danych **wielowymiarowy rozkład normalny** który najlepiej wytłumaczy ich położenie*

*...A definiujemy go za pomocą **wektora średniej** i **macierzy kowariancji***

21

Gaussian Mixture Model (pl: *Mieszanina rozkładów Gaussa)*



Załóżmy ilość rozkładów gaussa (odpowiadających klastrom) które wymagane są do wyjaśnienia danych *(powiedzmy: 6)*

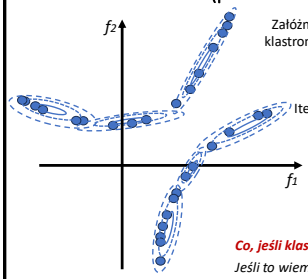
Iteracyjnie dopasowujmy parametry rozkładów by dopasować je jak najlepiej do danych

Co, jeśli założymy zbyt wiele rozkładów?

Nic bardzo groźnego się nie zdarzy – po prostu mamy problem trudniejszy w optymalizacji (dopasowaniu rozkładów)

22

Gaussian Mixture Model (pl: *Mieszanina rozkładów Gaussa)*



Załóżmy ilość rozkładów gaussa (odpowiadających klastrom) które wymagane są do wyjaśnienia danych *(powiedzmy: 6)*

Iteracyjnie dopasowujmy parametry rozkładów by dopasować je jak najlepiej do danych

Co, jeśli założymy zbyt wiele rozkładów?

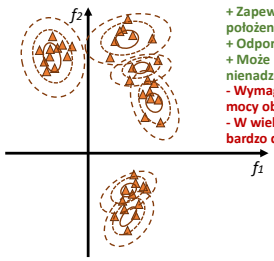
Nic bardzo groźnego się nie zdarzy – po prostu mamy problem trudniejszy w optymalizacji (dopasowaniu rozkładów)

Co, jeśli klastry również nie mają rozkładu normalnego?

Jeśli to wiemy – możemy zmodyfikować używany rozkład
Jeśli nie – wciąż zwykle jesteśmy dość blisko rzeczywistości

23

Gaussian Mixture Model (pl: *Mieszanina rozkładów Gaussa)*



+ Zapewnia miarę gęstości prawdopodobieństwa położenia punktów
+ Odporny na anomalie
+ Może być używany do klasyfikacji lub do nienadzorowanej klasteryzacji
- Wymaga albo znajomości ilości klastrów lub dużej mocy obliczeniowej
- W wielowymiarowej przestrzeni cech wymaga bardzo dużo danych

24

Pytania do powtórki:

1. Przedstaw jak dane eksperymentalne dzielone są na podzbiory, nazwij te podzbiory i wyjaśnij do czego są używane
2. Wyjaśnij separowalność liniową i jej brak, anomalie, klastry i korelację danych. Zaprezentuj graficzne ilustracje w.w. pojeć
3. Wyjaśnij jak działa metoda kNN w klasyfikacji (podaj też wady i zalety)
4. Wyjaśnij na czym polega klasteryzacja danych, wyjaśnij metodę centroidów z jej wadami i zaletami
5. Wyjaśnij jak działa metoda mieszaniny rozkładów Gaussa – z graficznym przykładem Gaussianów dopasowanych do klastrów danych. Przedstaw wady i zalety tej metody

25