Statystyczna Analiza Danych – laboratorium

Wprowadzenie do uczenia maszynowego, klasyfikacja

Dorota Celińska-Kopczyńska

Uniwersytet Warszawski

Zajęcia 6
13 kwietnia/14 kwietnia 2023

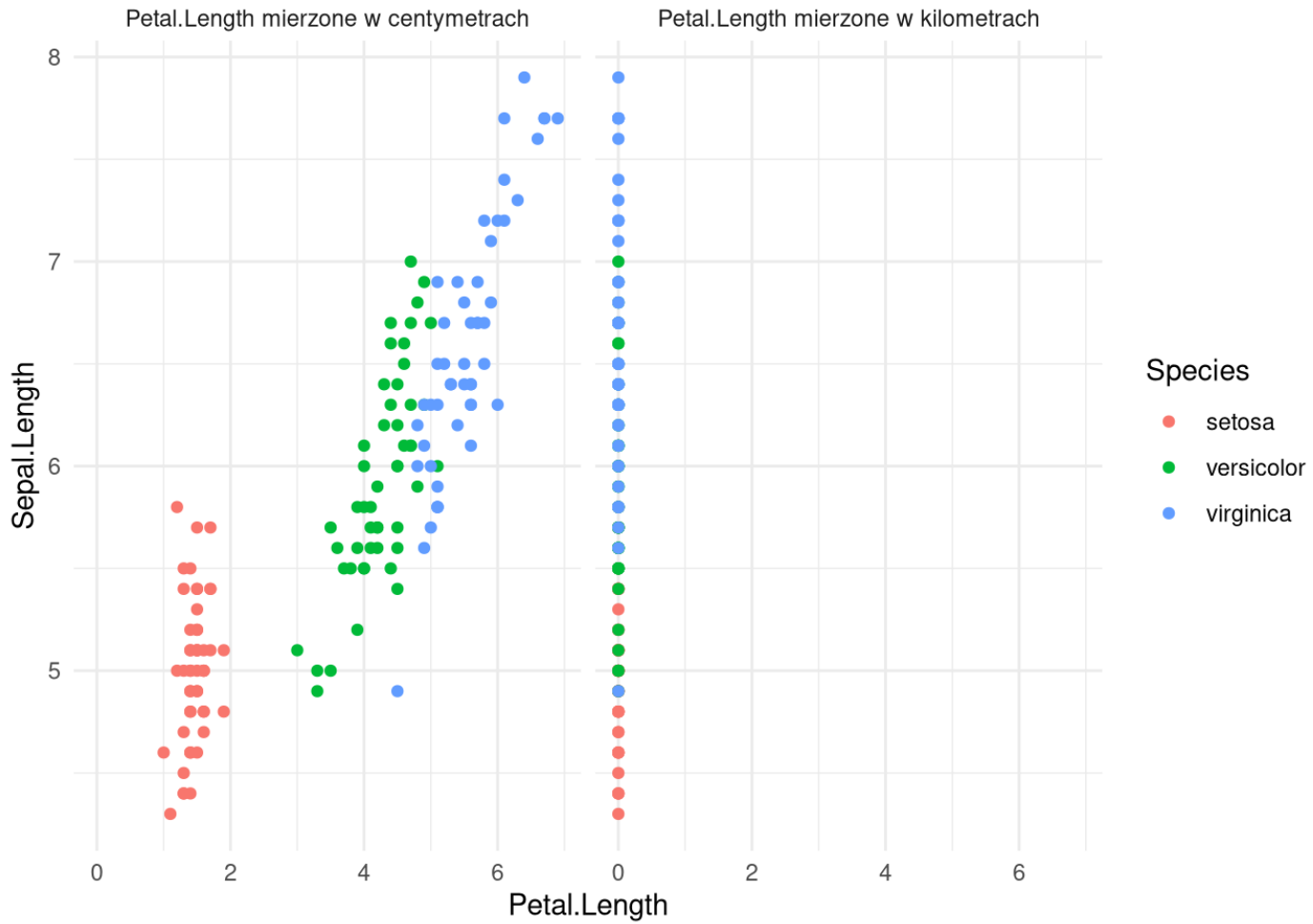
Idea zajęć – co i po co będziemy robić?

- Wchodzimy w tematykę uczenia maszynowego
- Zajmiemy się problemem **klasyfikacji** – na podstawie obserwowalnych cech (predyktorów) będziemy chcieli przyporządkować obiekty do k rozłącznych klas (wiemy, co to za klasy)
- Pokażemy również kilka miar oceniania jakości naszych wyników

Odwzorowanie punktów

- Obiekt opisany jest za pomocą n zmiennych $X_1, \ldots, X_n$ i przedstawiony jako punkt $x = (x_1, \ldots, x_n)$ w przestrzeni n-wymiarowej
- Dążymy do tego, żeby obiekty podobne (reprezentowane przez punkty znajdujące się blisko siebie w przestrzeni) znalazły się w jednej grupie, a niepodobne w różnych

Efekt jednostek

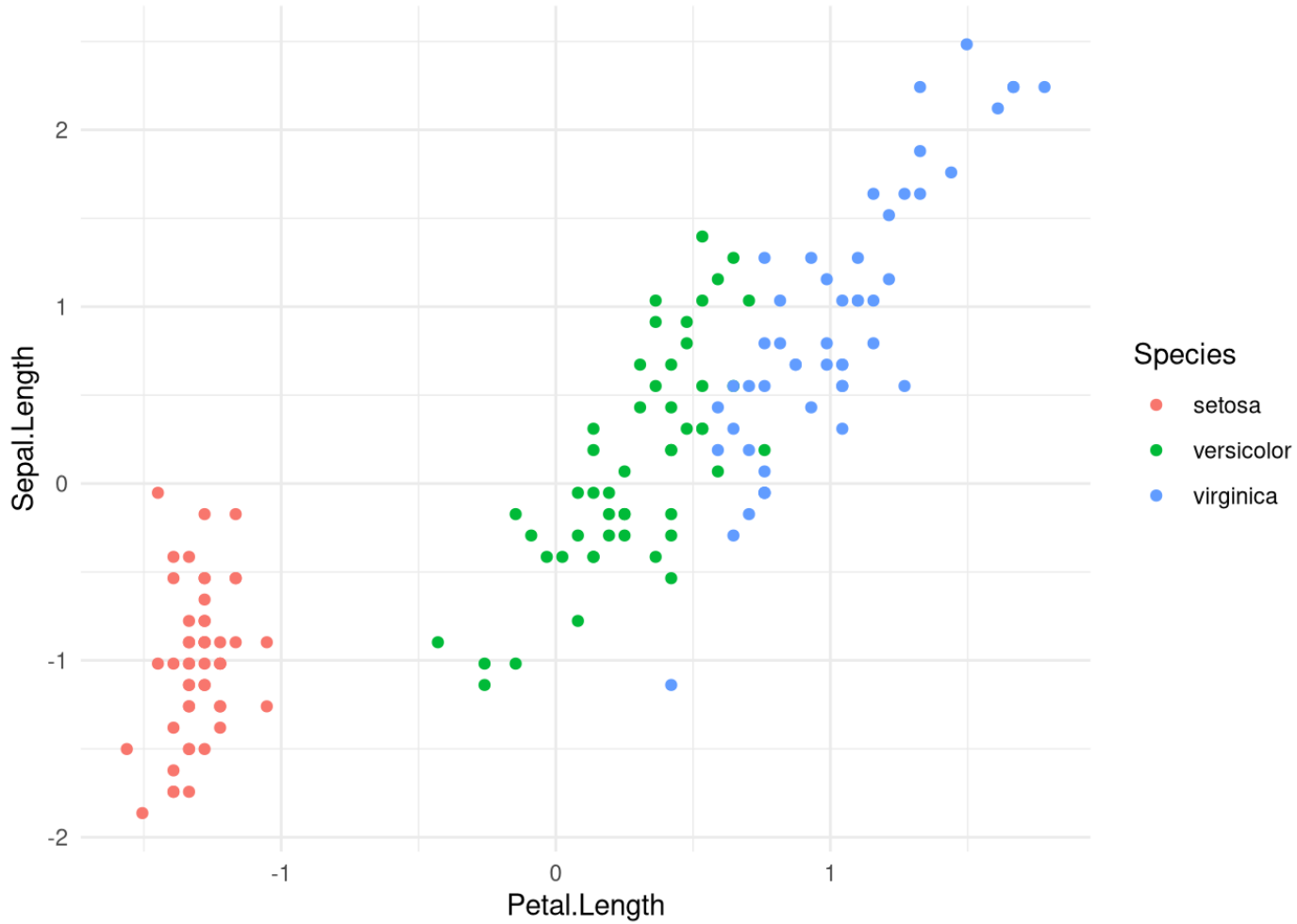


Przekształcenia liniowe zmiennych

- Klasyfikatory mogą niewłaściwie działać, jeśli zmienne podane są w różnych jednostkach
- Centrowanie: odjęcie od każdej kolumny wartości A
- Wyskalowanie: podzielenie każdej kolumny przez jej B
- Podstawowe metody radzenia sobie z problemem różnych jednostek:
  - Standaryzacja: $A = \bar{X}$, $B = S_X$
  - Normalizacja: $A = \min(X)$, $B = \max(X) - \min(X)$

```
# dla standaryzacji:
# centrowanie
iris[,1:4] <- apply(iris[,1:4], 2, function(x) x - mean(x))
# skalowanie
iris[,1:4] <- apply(iris[,1:4], 2, function(x) x / sd(x))
```

Efekt standaryzacji



Zadanie 1

- Wczytaj dane z pliku wine.csv
- Dokonaj standaryzacji wszystkie kolumny predyktorów z danych wine. Możesz w tym celu wykorzystać albo apply(), albo funkcję scale()
- Zrzutuj zmienną Quality na typ factor za pomocą funkcji as.factor()

Idea klasyfikatora KNN

- Jeden z najprostszych klasyfikatorów
- Przypisujemy danej obserwacji taką klasę, jaka pojawia się najczęściej wśród jej k najbliższych sąsiadów
- Użytkownik decyduje, jakie k wybrać i jaką miarę odległości
- Dane zawierające prawdziwe klasy to dane treningowe. Dane, które chcemy przyporządkować to dane testowe.

Przykładowe miary odległości

- Odległość Minkowskiego: $d(O_1, O_2) = (\sum_{i=1}^{n} |x_{1i} - x_{2i}|^p)^{1/p}$
  - Odległość miejska ($p = 1$): $d(O_1, O_2) = \sum_{i=1}^{n} |x_{1i} - x_{2i}|$
  - Odległość euklidesowa ($p = 2$): $d(O_1, O_2) = \sqrt{\sum_{i=1}^{n} (x_{1i} - x_{2i})^2}$
  - Odległość Czebyszewa ($p = \infty$): $d(O_1, O_2) = \max |x_{1i} - x_{2i}|$
- Odległość Mahalanobisa: $\sqrt{(\mathbf{X} - \mathbf{Y})^T \mathbf{C}^{-1} (\mathbf{X} - \mathbf{Y})}$, macierz $\mathbf{C}$ symetryczna, dodatnio określona

Zadanie 2

- Korzystając z klasyfikatora kNN, spróbuj przewidzieć jakość wina o parametrach (dla wystandaryzowanych danych), użyj metryki euklidesowej :
  0.42, 0.03, -0.90, 0.15, -1.25, -0.15, -0.01, 0.73, 0.90, -0.82, -0.69
- Czy wyniki zmienią się, jeśli skorzystamy z różnych k?
- Czy wyniki zmienią się, jeśli użyjemy innej metryki odległości?

Błąd treningowy, błąd testowy

- Aby ocenić właściwości klasyfikatora używamy zbioru testowego
- Duża różnica pomiędzy wartością błędu na zbiorze treningowym a testowym może sugerować *przeuczenie* klasyfikatora (reaguje dobrze na znane fakty, ale nie radzi sobie z nowymi)
- Jednokrotne sprawdzenie na zbiorze testowym nie jest polecane. Walidacją krzyżową zajmiemy się w przyszłości

Podział zbioru

```
# losowanie indeksow bez zwracania
indeksy_testowe <- sample(1:nrow(wine), 480, replace=F)
zbior_testowy <- wine[indeksy_testowe, ]
zbior_treningowy <- wine[-indeksy_testowe, ]  # Indeksowanie ujemne wiele ułatwia!
```

Accuracy

- Proporcja poprawnie zaklasyfikowanych obserwacji
- Jakie może być accuracy dla danych rozmiaru 100, w których 99 obserwacji jest typu A, a jedna typu B? Rozważ różne klasyfikatory

Macierz konfuzji, precision i recall

- Tablica klasyfikacyjna: wiersze odpowiadają prawdziwym klasom, a kolumny klasom zwróconym przez klasyfikator. Komórka $m_{i,j}$ zawiera liczbę obserwacji z klasy i zaklasyfikowaną jako klasa j
- Precision dla klasy i:

$$p_i = \frac{m_{i,i}}{\sum_{j=1}^{k} m_{j,i}}$$

- Recall dla klasy i:

$$r_i = \frac{m_{i,i}}{\sum_{j=1}^{k} m_{i,j}}$$

Zadanie 5

- Utwórz macierz konfuzji dla wyników klasyfikatora KNN na zbiorze testowym z danych wine
- Oblicz precision i recall dla każdej klasy
- Porównaj wyniki dla 3 wybranych wartości parametru k