Inteligencja obliczeniowa

Laboratorium 4: Korelacja i Analiza Głównych Składowych

Tematem laboratoriów jest badanie trzech gatunków irysów.

Iris setosa

Iris versicolor

Iris virginica

Każdemu z trzech gatunków mierzono długość i szerokość płatka „petal” i działki kielicha kwiatu „sepal” (to te mniejsze płatki pomiędzy większymi, skierowane do góry). Mamy więc 4 atrybuty o wartościach rzeczywistych. Stworzono małą bazę danych (załączony plik lub komenda `iris` w R).

Zadanie 1 *(0,5 pkt)*
Czy długość płatka i szerokość płatka są skorelowane? A może im mniejsza szerokość działki kielicha tym dłuższy płatek? Twoim pierwszym celem jest zbadanie wzajemnej korelacji wszystkich czterech atrybutów.

a) Wczytaj załączoną bazę danych irysów i zapisz je pod nazwą `iris.original`. Sprawdź jak wyglądają dane.
b) Znajdź komendę badającą korelację i dokonaj porównania każdej numerycznej kolumny z każdą inną numeryczną (4 kolumny, co daje łącznie 6 porównań).
c) Czy w każdym z przypadków korelacja była dodatnia czy ujemna? Czy była silna (0.7-1), słaba (0.2-0.7) czy w ogóle jej brakło (0-0.2)? Zinterpretuj wyniki.

Zanim przejdziemy dod alszych zadań zwróć uwagę, że wszystkie trzy irysy są do siebie podobne. Jak je odróżnić? Naszym celem będzie teraz przeprowadzenie analizy głównych składowych. W skrócie: chcemy 4 parametry kwiatów zamienić na mniejszą ilość parametrów, które nadal będą dobrze charakteryzowały kwiaty. Gdy zredukujemy ilość atrybutów np. do dwóch (dwie główne składowe), to nie dość, że baza danych będzie „odchudzona” to w dodatku będzie można kwiaty umieścić na wykresie 2D, gdzie wyodrębnią się pewne grupy.

Zadanie 2 *(0,5 pkt)*
Dokonaj preprocessingu danych i zapisz końcowa tabelę pod nazwą : `iris.preproc.`
a) Pod nazwą `iris.numeric` zapisz tabelę bez ostatniej kolumny (czyli bez gatunków)
b) Zlogarytmuj wszystkie dane liczbowe (funkcja log)
c) Dokonaj standaryzacji wszystkich kolumn (funkcja scale).

Funkcje logarytm i standarize często „ładniej” rozmieszczają dane na wykresie, ale bywają sytuacje gdy ich stosowanie jest niewskazane. Możesz później wykonać wersję zadań bez tych dwóch funkcji i zobaczyć jakie otrzymasz wyniki.

Zadanie 3 *(1 pkt)*
Zajmiemy się teraz analizą głównych składowych dla tabeli.

a) Skorzystaj z funkcji `prcomp`, by obliczyć główne składowe i rezultat zapisz pod nazwą `iris.pca`.

b) Wyświetl za pomocą komendy `iris.pca` lub `print(iris.pca)` dane dotyczące głównych składowych. Jakie odchylenia standardowe mają główne składowe? W analizie głównych składowych należy (choć nie zawsze) zostawić główne składowe tylko z największą wariancją. Wygooglaj jak się ma odchylenie standardowe do wariancji oraz które składowe powinniśmy zostawić, a które odrzucić. Co zwracają komendy `iris.pca[1]` i `iris.pca[2]`?

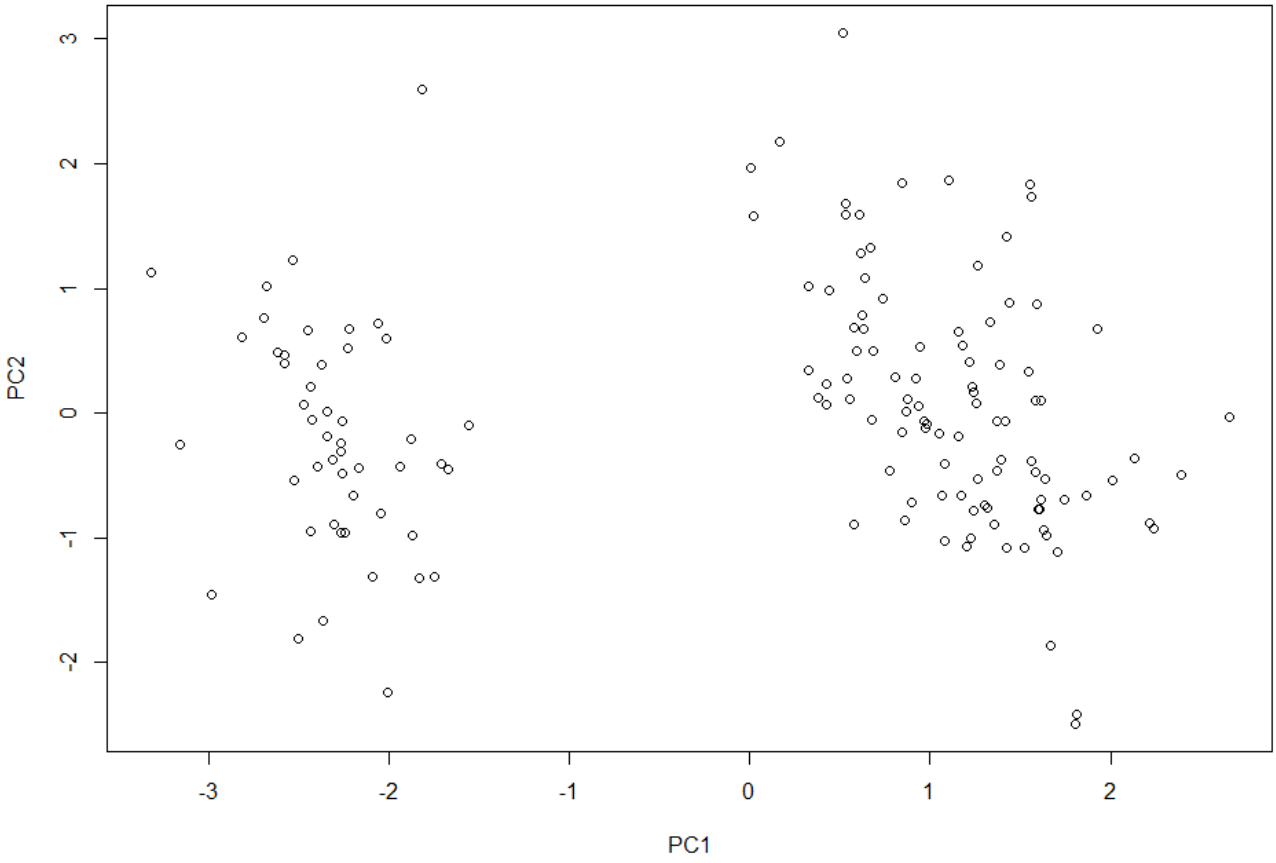
c) Kombinacja liniowa z jakimi współczynnikami tworzy główne składowe? Jaką tabelę zwraca komenda `predict(iris.pca)`? Sprawdź czy ta nowa tabela danych (z głównym składowymi) jest poprawnie stworzoną kombinacją liniową danych wejściowych.
Wskazówka: sprawdź czy pierwsza liczba z `predict(iris.pca)` powstaje przez pomnożenie pierwszego wiersza macierzy `iris.preproc` z pierwsza kolumną macierzy `iris.pca[2]`.

d) Zostawiamy dwie najlepsze główne składowe. Jak można je zinterpretować? Np. jeśli PC1 jest skorelowane dodatnio z PetalLength i PetalWidth, a pozostałymi nie to można PC1 nazwać po prostu „Wielkość płatka”. Czasami korelacja jest ujemna. Zbadaj korelację między PC1,PC2 a wartościami pierwotnymi (iris.preproc$SepalLength i 3 innymi), by spróbować odkryć ich znaczenie (podobnie jak w zadaniu 1).

e) Zapisz `predict(iris.pca)` pod nową nazwą `iris.pca.data`, obetnij niepotrzebne kolumny, dodaj kolumnę z gatunkiem irysa. Będą zatem trzy kolumny w bazie danych.

Zadanie 4 *(1pkt)*
Naniesiemy teraz dane na wykres. Osie x i y będą odpowiadały głównym składowym PC1, PC2. Każdy kwiat będzie punktem na wykresie. Zrób tak, aby każdy rodzaj kwiatu (setosa, versicolor, virginica) miał inny kolor na wykresie. Dodaj legendę.

Poniżej graf jeszcze przed kolorowaniem punktów i dodaniem legendy.



Gdy skończysz tworzyć wykres, sprawdź czy każdy z gatunków ma swoją własną charakterystykę, skupia się w określonym obszarze.

Zadanie 5 *(1 pkt)*
A co jeśli teraz przyjdzie ktoś z nowymi irysem, pomierzy jego cztery parametry i spyta się Ciebie jaki to gatunek? Napisz funkcję:
`plotIris(SeplLength,SepalWidth,PetalLength,PetalWidth),`
która oznaczy inną, dodatkową (czarną) kropką irysa na przygotowanym wykresie (zadanie 4). Funkcja nie musi rozpoznawac jaki to gatunek irysa, wystarczy, że go oznaczy na wykresie. Rozpoznanie wykona użytkownik.

Funkcja musi:

- zlogarytmować cztery parametry na wejściu
- ustandaryzować wszystkie cztery parametry wzorem:
  (parametr – średnia_parametru_z_tabeli) / odchyl_stand_parametru_z_tabeli
  (nie należy korzystać tu z wbudowanych funkcji scale, bo taka funkcja obliczy złą średnią i odhcylenie)
- dane po takim preprocessingu należy zamienić na dwa parametry: PC1 i PC2 obliczając je na podstawie tabeli współczynników z `iris.pca[2]`
- Na końcu dodajemy punkt na wykresie.

Przetestuj funkcję dla poniższych danych:
plotIris(6.9, 3.2, 5.6, 2.2)
plotIris(6.0, 2.6, 4.4, 1.6)
plotIris(4.8, 3.6, 1.4, 0.2)
Czy jesteś w stanie oszacować jakie to irysy patrząc na wykres?