Porównanie wizualizacji PCA i LDA

Bazyli Reps, Łukasz Knigawka

16 kwietnia 2020

Spis treści

1 Opis problemu

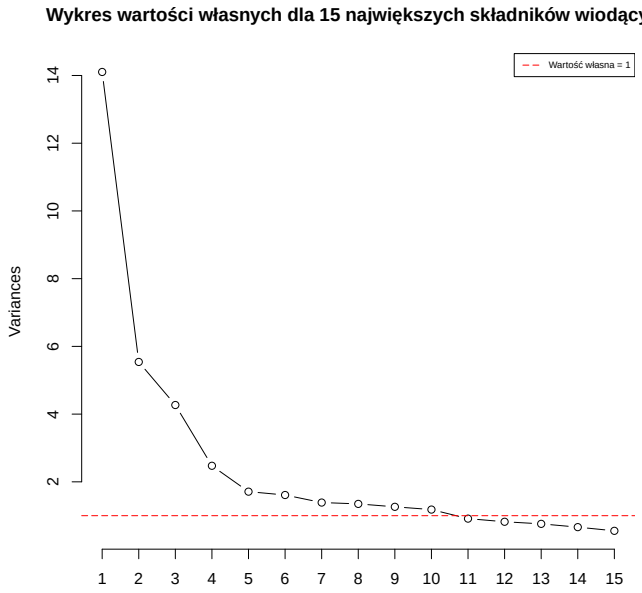
Celem niniejszej pracy jest porównianie kompresji wymiarów danych za pomocą metod PCA oraz LDA.
Zamierzamy przedstawić przede wszystkim analizę redukcji wymiarowości dla obu metod, a następnie porównać je pod względem czasu wykonania redukcji a także uzyskanego stopnia kompresji.
Skrypty potrzebne do wykonania analiz napisane zostały w języku R.

2 Opis danych

Analizy przeprowadzone zostały na danych pobranych ze strony UCI Machine Learning Repository: `http://archive.ics.uci.edu/ml/datasets/SPECTF+Heart`
Przedstawiają one badania wyniki badań serca dla poszczególnych ROI (Region Of Interest) w zależności od stanu badanej osoby (w spoczynku lub podczs wysiłku).
Dane zawierają 44 kolumny zawierające dane dla 22 ROI w obu stanach oraz jedną kolumnę opisującą która przyjmuje dwie wartości: 'normal' oraz 'abnormal'. Zestaw zawiera 265 rekordów, spośród których 55 obserwacji posiada etykietę 'abnormal', a 210 'normal'.

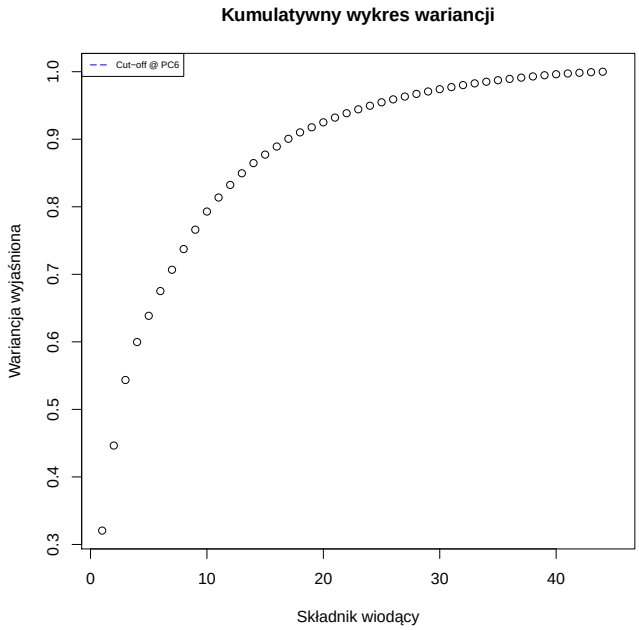
3 Analiza PCA

Dane zostały zredukowane do dwóch wymiarów. Na początku zostały wyznaczone wartości własne.



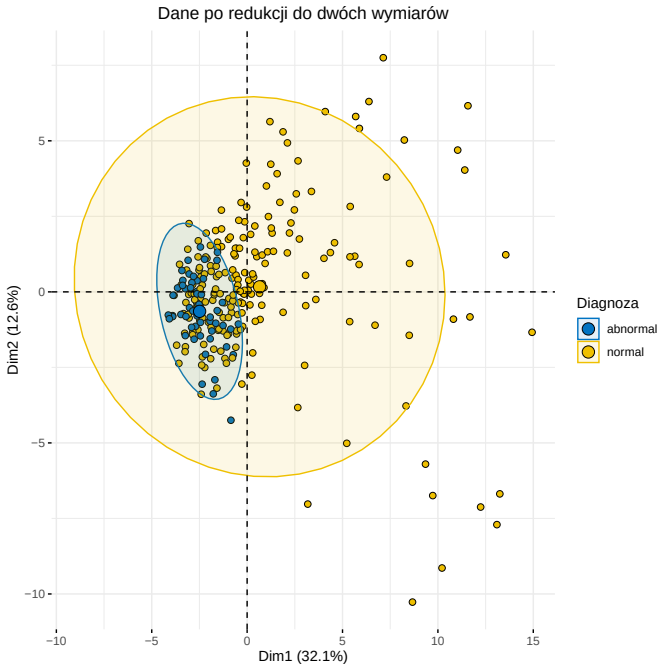
Rysunek 1: Istnieje dziesięć wartości własnych większych od 1. Pozostałe można swobodnie usunąć.

Następnie zbadana została wariancja wyjaśniona w zależności od ilości wziętych składników wiodących.



Rysunek 2: Największe dwa składniki wiodące wyjaśniają około 45% wariancji danych

W dalszej kolejności zwizualizowane zostały dane po redukcji do dwóch wymiarów. Na wykresie zaznaczone zostały klasy obiektów.

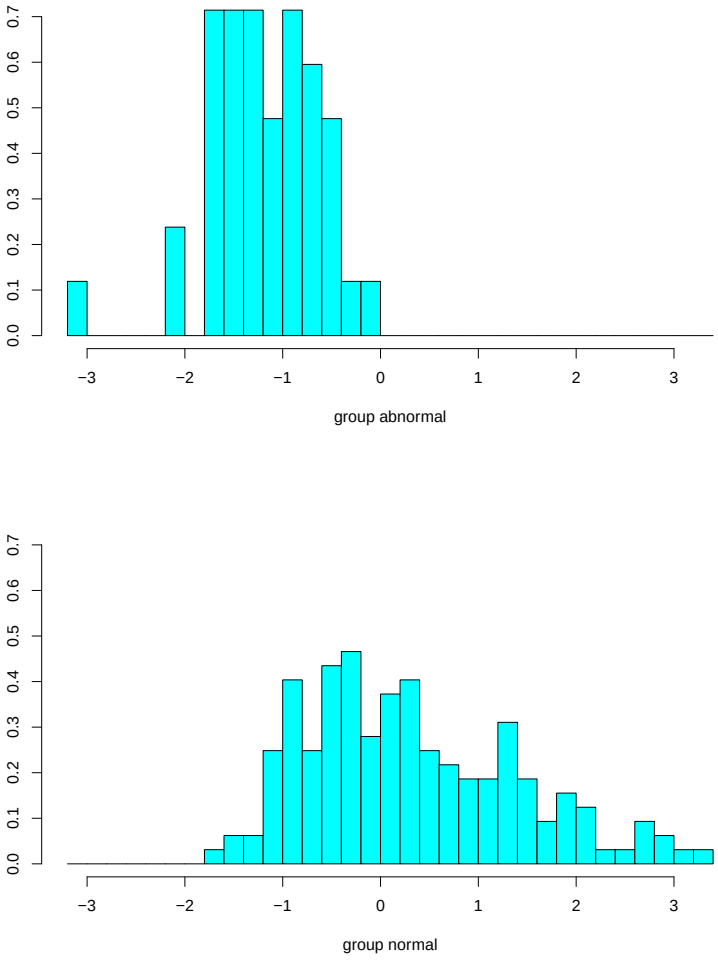


Rysunek 3: Wyraźnie widoczne jest zgrupowanie rekordów z diagnozowanych jako 'abnormal'. Niestety charakter danych, a co za tym idzie mniejsza od 60% wyjaśniona wariancja powodują, że pomimo redukcji obie klasy nie są wyraźnie od siebie odseparowane.

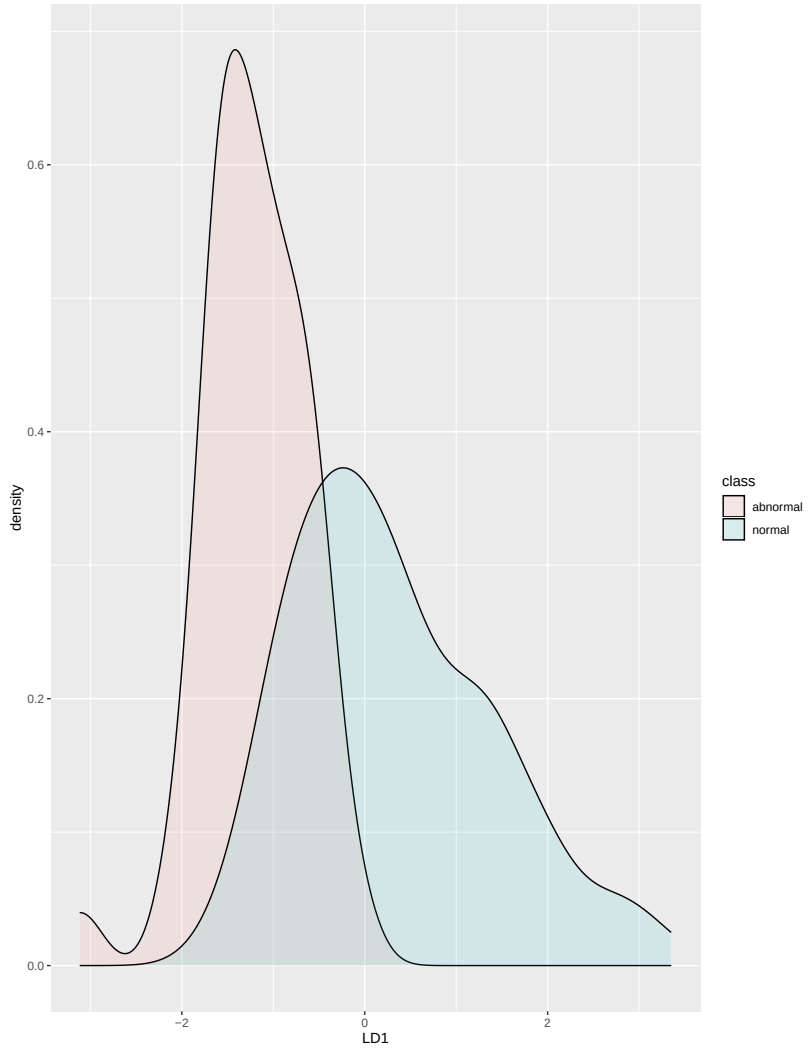
Wywołanie funkcji *prcomp()* trwało 1.170636 sekundy.

4 Analiza LDA

Dyskryminację liniową można wykorzystać w zadaniu redukcji wymiarowości, przy jej pomocy można zredukować liczbę wymiarów do N-1, gdzie N to liczba klas. W naszym przypadku liczba klas wynosi 2, stąd liczbę wymiarów można zmniejszyć do zaledwie jednego. LDA można także użyć jako klasyfikatora.
Aby przeprowadzić dyskryminację liniową dane zostały podzielone na zbiór treningowy oraz testowy (w klasycznych proporcjach odpowiednio 75% oraz 25% obserwacji). Dane zostały przeskalowane. Po nauczeniu klasyfikatora zbadaliśmy funkcje dyskryminacji dla poszczególnych klas.



Rysunek 4: Histogramy funkcji dyskryminacji dla poszczególnych klas



Rysunek 5: Funkcje gęstości wektora zmiennych dyskryminujących dla obu klas przedstawione na jednym wykresie. Widzimy, że klasy nie są jednoznacznie separowalne.

Po zbudowaniu klasyfikatora przeszliśmy do jego oceny. Na bazie obserwacji niewykorzystanych do budowy klasyfikatora zbadaliśmy zgodność klasyfikatora z rzeczywistymi danymi.

Listing 1: Macierz pomyłek dla danych treningowych

```
           Actual
predicted  abnormal normal
  abnormal       18      3
  normal         24    158
```

Dla danych treningowych osiągnięto precyzję 0.8669951.

Listing 2: Macierz pomyłek dla danych testowych

```
           Actual
predicted  abnormal normal
  abnormal        5      3
  normal          8     46
```

Dla danych testowych osiągnięto precyzję 0.8225806.
Wywołanie funkcji *lda()* z biblioteki *MASS* trwało 1.57568 sekundy.

5 Podsumowanie

- Oba algorytmy skutecznie redukują liczbę zmiennych opisujących zjawisko.
- Metoda LDA wymaga mniejszej liczby składowych głównych niż PCA.
- W przeciwieństwie do metody PCA, w przypadku LDA stosuje się wstępny podział danych ze względu na przynależność do klasy.