SAR 2020/2021 Laboratorium 10

22-23.12.2020

10.1 PCA

Dane w pliku *Miasta.txt* zawierają wartości trzech atrybutów dla 46 miast na świecie:

- Work - ważona średnia liczba godzin pracy dla 12 zawodów,
- Price - indeks kosztów utrzymania na podstawie cen 112 towarów i usług (wartość indeksu dla Zurichu równa się 100),
- Salary - indeks płacy za godzinę w 12 zawodach po odjęciu podatku (wartość indeksu dla Zurichu równa się 100).

a) Dokonać standaryzacji zmiennych.

b) Dla pary zmiennych: Work i Price wyznaczyć kierunki wzdłuż których występuje największa zmienność. Przedstawić wykres rozproszenia dla pary zmiennych Work i Price a następnie wrysować linie wzdłuż których występuje największa zmienność.

c) Wykonać analizę składowych głównych dla wszystkich zmiennych (standaryzowanych).

d) Jaki jest procent wariancji tłumaczony przez poszczególne składowe? Czy możemy dokonać redukcji wymiaru danych?

e) Obliczyć kierunki główne oraz składowe główne.

f) Znajdź miasto o największej wartości pierwszej składowej głównej. O czym świadczy duża wartość pierwszej składowej głównej?

10.2 PCR i PLSR

Wczytać dane z pliku *prostate.data*. Dane zawierają informacje dotyczące raka prostaty u 97 mężczyzn. W zbiorze znajduje się zmienna *train* przyjmująca 2 wartości: TRUE (dla obserwacji ze zbioru treningowego) oraz FALSE (dla obserwacji ze zbioru testowego).

a) Na podstawie próby treningowej dopasować model liniowy opisujący zależność zmiennej *lpsa* od pozostałych zmiennych (z wyjątkiem zmiennej *train*). Dokonać predykcji na zbiorze testowym, obliczyć RMSE.

b) Na podstawie próby treningowej dopasować model opisujący zależność zmiennej *lpsa* od pozostałych zmiennych (z wyjątkiem zmiennej *train*) metodą PCR. Dokonać wyboru zmiennych metodą kroswalidacji. Dokonać predykcji na zbiorze testowym, obliczyć RMSE.

c) Na podstawie próby treningowej dopasować model opisujący zależność zmiennej *lpsa* od pozostałych zmiennych (z wyjątkiem zmiennej *train*) metodą PLSR. Dokonać wyboru zmiennych metodą kroswalidacji. Dokonać predykcji na zbiorze testowym, obliczyć RMSE.