Metody analizy i wizualizacji dużych zbiorów danych
Laboratorium 2 – Metody MDS

Jan Gołda, Wojciech Basiura - WIET 2018/2019

Opis zagadnienia

Laboratorium 2 poświęcone było wizualizacji małych zbiorów danych przy użyciu różnych metod multidimensional scaling. Dotyczyło również metryk k-nearest neighbors i neighbours hit.

Zadaniem pierwszym była wizualizacja zbioru "lung cancer" metodami MDS, Isomap, oraz LLE i algorytmami poznanymi na pierwszych zajęciach – PCA i kernel PCA. Wyniki mieliśmy porównać oboma poznanymi metrykami.

Zadaniem drugim było odfiltrowanie części wymiarów ze spreparowanego zbioru sphere w którym tylko część wymiarów jest sprzężona z etykietami – oryginalny zbiór rozszerzono o losowo wygenerowane cechy.

Zadanie pierwsze

Isomap dla parametru num_of_neighbors = 2
LLE dla parametru num_of_neighbors = 9

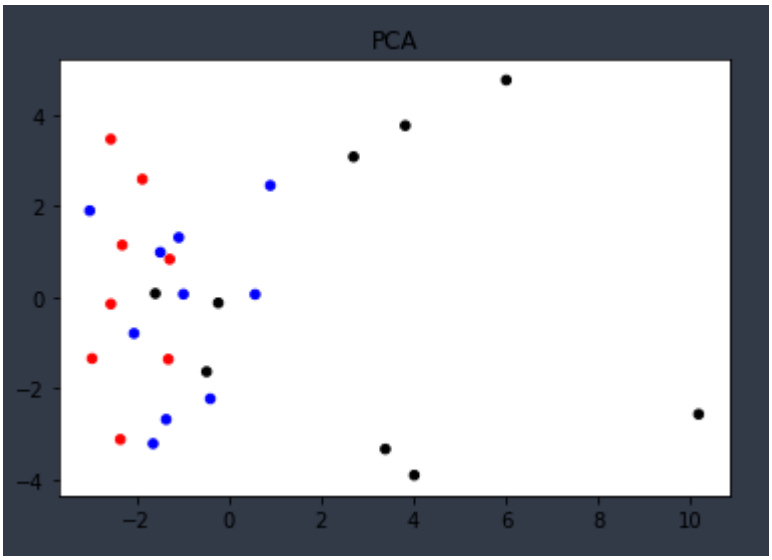
	PCA	Kernel PCA	MDS	Isomap	LLE
hit-to-all	59%	61%	63% (1)	60%	68%
Neighbors-hit (2)	3.04	2.93	3.26	2.74	2.70

(1) ma duże wahania – to jest maksimum jakie znaleźliśmy
(2) out of 5

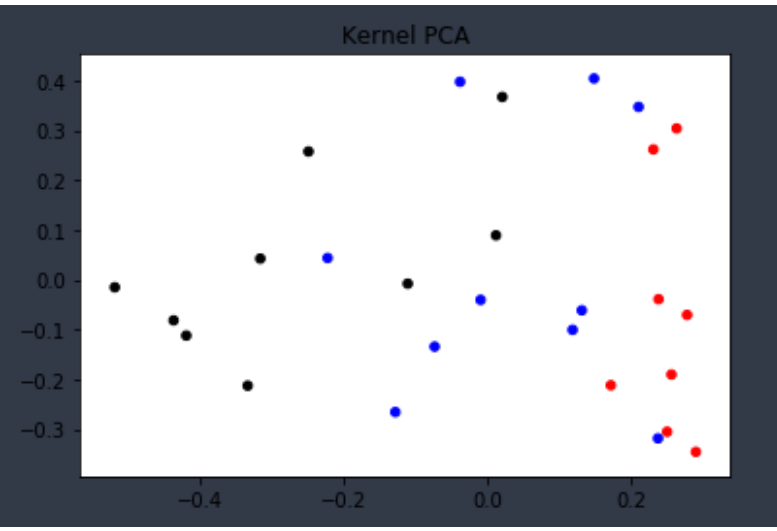
1. Zbiór „lung cancer”

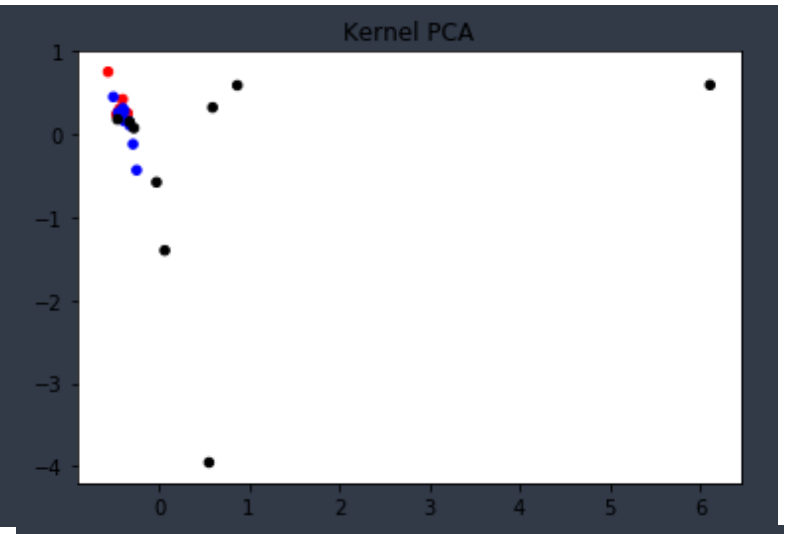
Zadany zbiór jest wyjątkowo mały – ma 32 wiersze i 57 kolumn. Zawierał również wartości nieokreślone (oznaczone przez znak „?”) które odfiltrowaliśmy. Pozostało 27 wierszy. Wartości pierwszej kolumny przyjmują wartości 1-3 i są etykietami trzech typów nowotworu płuc.

2. Metody PCA i KPCA

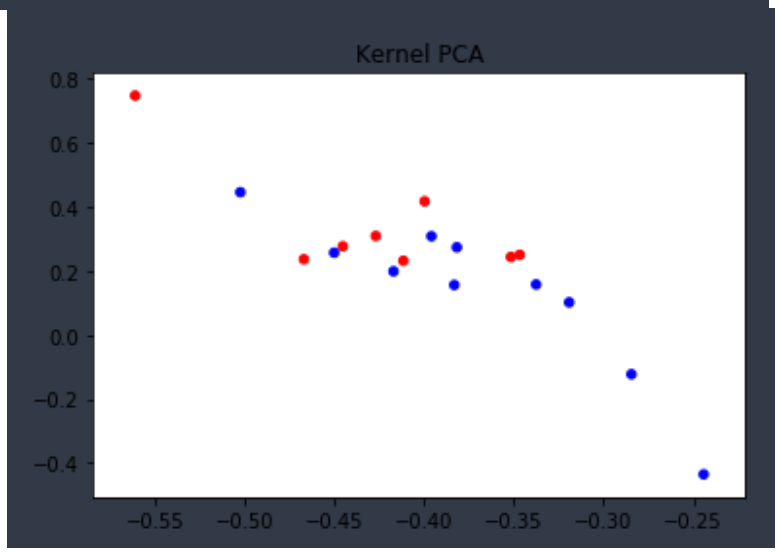


ze StandardScaler





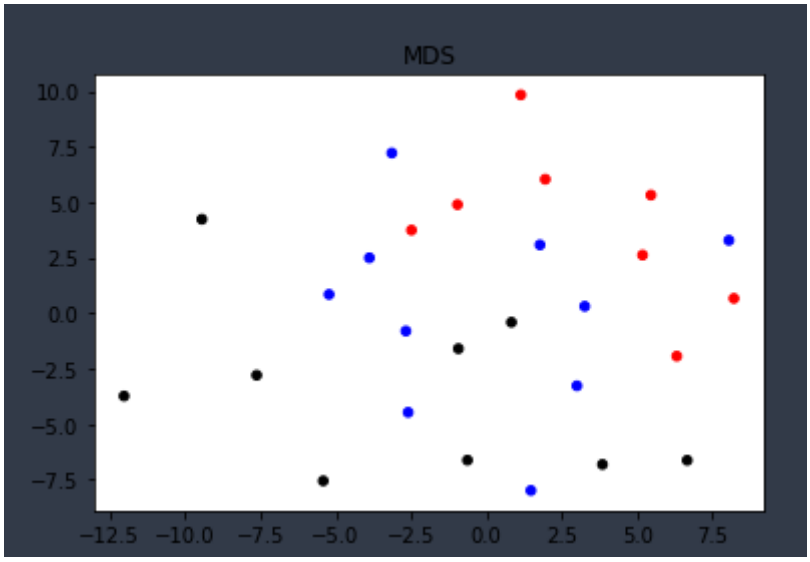
kernel = „rbf”, bez skalowania



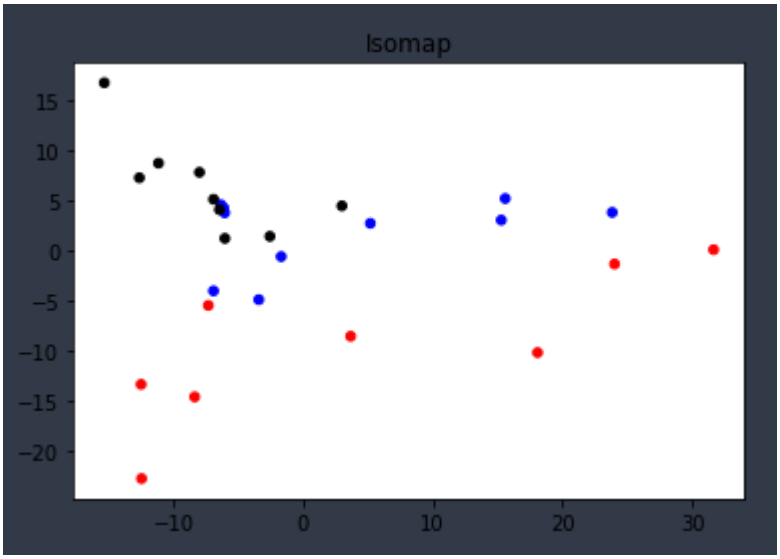
kernel = „poly”, StandardScaler

Część czarnych punktów udało się wydzielić. Grupy czerwonych i niebieskich punktów są bardzo blisko i zazwyczaj się przeplatają.

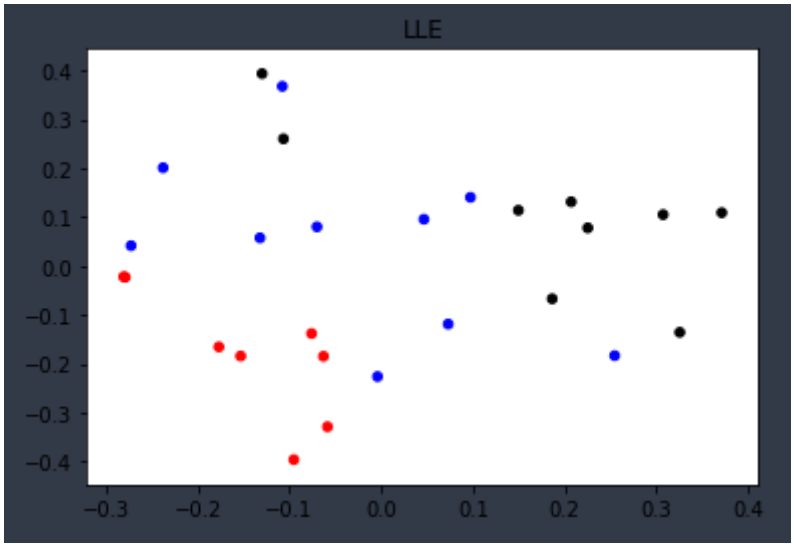
3. Metody mds, Isomap, LLE



eps=0.023, StandardScaler



num_of_neighbors=2



num_of_neighbors=9

Zadanie drugie

1. Zbiór „sphere ”

Zbiór sphere.csv ma wymiary (9976, 61). Pierwsza kolumna określa przynależność do grupy – wartości „one” i „two”. Ze względu na czas obliczeń wybieramy losowe 2000 wierszy.

2. Metoda

- Wykonujemy LLE na wybranych wierszach.
- Wyliczamy wartość metryki hit-to-all.
- Usuwamy jedną z kolumn i powtarzamy poprzednie kroki.
- Jeżeli zmiana wartości metryki jest duża oznacza to, że kolumna była istotna przy rozdzielaniu zbiorów. W przeciwnym wypadku można podejrzewać, że jest sztucznie dodana.
- Kwestią kluczową jest dobór parametru określającego kiedy kolumna jest istotna, a kiedy nie.
- W naszym przypadku, przyjęcie 2.5% idealnie wyznaczyło 10 kolumn które prawdopodobnie są „prawdziwymi” danymi.

3. Potwierdzenie

Jako potwierdzenie policzyliśmy wariancję w kolumnach. 10 najniższych wyników to kolumny wybrane przez naszą metodę. W pierwszej grupie wariancja waha się od 24.7 do 27.4. Po 10

najniższych wynikach następuje skok z 27.4 na 70.0. Najwyższa wartość to 79.6. Wyraźnie widać podział na dwie grupy znaleziony przez naszą metodę.