K najbliższych sąsiadów - zadanie pierwsze

28 października 2017

Grzegorz Borkowski

1. WSTĘP

Zadanie zostało zaimplementowane w Pythonie 3.6.1
Do raportu został dołączony Jupyter Notebook z kodem źródłowym programu.

2. ZADANIE 1

2.1 Treść zadania

Celem zadania jest obserwacja jak zmiana wykorzystywanej przez klasyfikator k-NN metryki wpływa na kształt granicy decyzyjnej, a więc w efekcie na jego skuteczność.

Zadanie rozpoczynamy od przygotowania 2 dwuwymiarowych testowych zbiorów danych, składających się z obserwacji należących do przynajmniej 3 klas. Staramy się tak dobrać kształty zbiorów, by uzyskać ciekawe obserwacje (warto tu trochę poeksperymentować).

Następnie na obu z tych zbiorów trenujemy poniższe 4 warianty klasyfikatora k-NN:
* zwykły k-NN z k=1 i metryką Euklidesa
* zwykły k-NN z k=3 i metryką Euklidesa
* zwykły k-NN z k=1 i metryką Mahalanobisa
* zwykły k-NN z k=1 i metryką wyuczoną uprzednio z użyciem algorytmu Large Margin Nearest Neighbor (przyzwoicie opisanego choćby na poczciwej Wikipedii oraz oczywiście w źródłowej publikacji).

Dla każdej pary zbiór-klasyfikator zwizualizuj wygląd granicy decyzyjnej oraz oszacuj

2.2 Opis zbiorów danych

Wykorzystywane w tym zadaniu zbiory to dwa dostępne na stronie UCI Machine Learning Repository: Iris i Wine. Iris to zbiór danych opisany w publikacji [1]. Zbiór składa się z trzech klas, każda ma po 50 instancji. Każda klasa to jeden typ Irysa. Jedna klasa jest liniowa odseparowana od pozostałych dwóch. Zadaniem jest przewidzieć klasę irysa. Informacje w zbiorze danych to:
1. Długość sepal w cm
2. Szerokość sepal w cm
3. Długość petal w cm
4. Szerokość petal w cm
5. Klasa: Iris Setosa, Iris Versicolour, Iris Virginica

Wine to zbiór danych opisany w [2]. Ten zbiór danych to rezultat chemicznej analizy win dojrzewających z tego samego regionu we Włoszech, ale z winoroślny zebranych z trzech różnych upraw. Analiza wyznaczyła ilości trzynastu różnych chemicznych składowych w kazdym z trzech typów wina. Zadaniem jest przewidzieć klasę wina.
Atrybuty (jak w oryginale): 1) Alcohol 2) Malic acid 3) Ash 4) Alcalinity of ash 5) Magnesium 6) Total phenols 7) Flavanoids 8) Nonflavanoid phenols 9) Proanthocyanins 10)Color intensity 11)Hue 12)OD280/OD315 of diluted wines 13)Proline

2.3 Analiza zbioru Iris

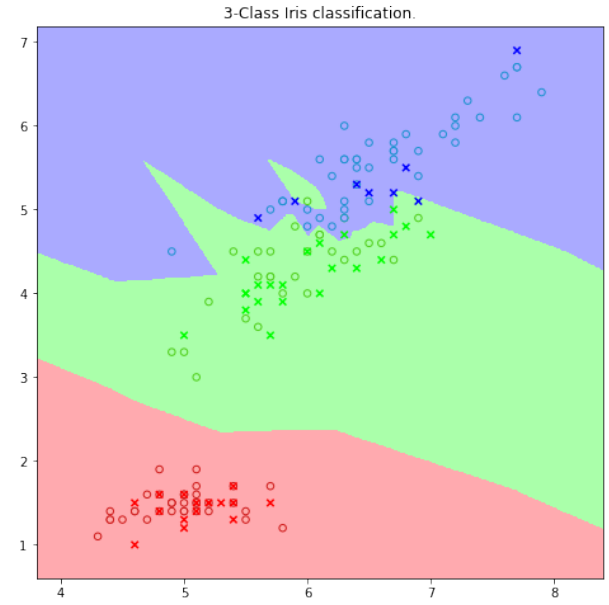
Zbiór wejściowy składa się ze 150 elementów, po 50 elementów na każdą klasę. Podzieliłem zbiór na dwa zbiory, treningowy i testowy (treningowy: 105, testowy: 45 elementów) w losowy sposób. Aby łatwo zwizualizować granicę decyzyjną wybieram tylko dwie cechy (długość length (cm), długość petal (cm)).

Na poniższych wykresach granic decyzyjnych, kółkami zaznaczone są elementy zbioru treningowe, krzyżykami elementy zbioru testowego. Kolor krzyżyka lub kółka odpowiada klasie przypisanej przez klasyfikator. (Czerwona - Iris setosa,

Large Margin Nearest Neighbors, k=1

klasa	precision	recall	f1score
Iris-setosa	1.00	1.00	1.00
Iris-versicolor	0.91	0.95	0.93
Iris-virginica	0.86	0.75	0.80
avg-total	0.93	0.93	0.93

Skuteczność klasyfikatora: 0.933

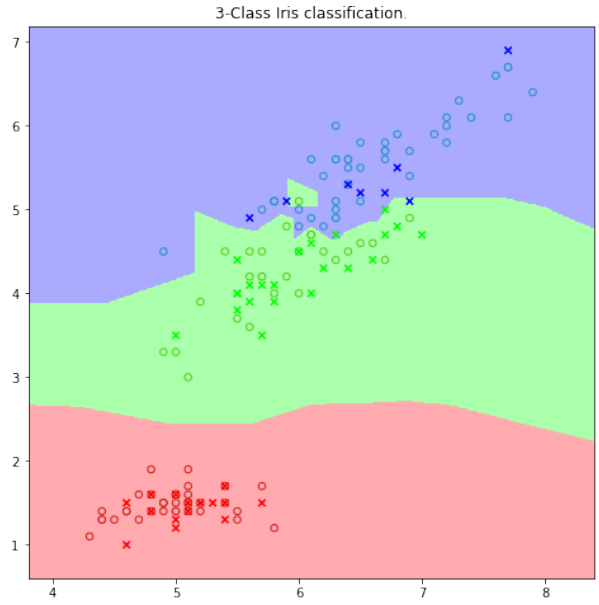


Rysunek 1. Granica decyzyjna dla Large Margin Nearest Neighbours

K-NN, k=1, Euklides

klasa	precision	recall	f1score
Iris-setosa	1.00	1.00	1.00
Iris-versicolor	0.91	0.95	0.93
Iris-virginica	0.86	0.75	0.80
avg-total	0.93	0.93	0.93

Skuteczność klasyfikatora: 0.93

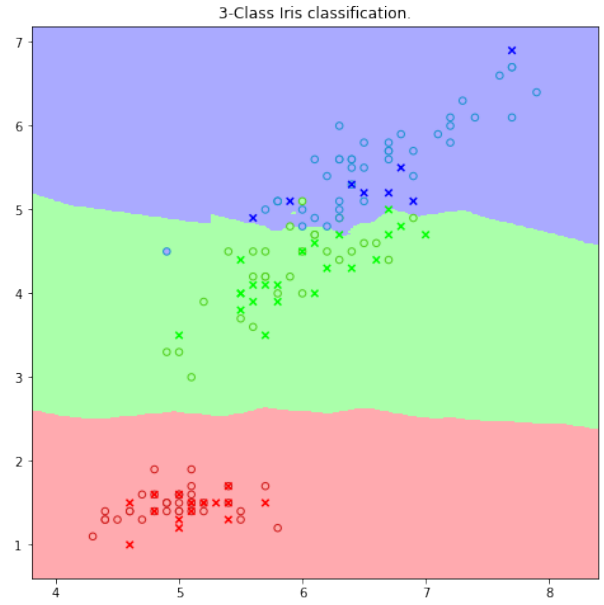


Rysunek 2. Granica decyzyjna dla K-NN z metryką Euklidesa dla k=1

K-NN, k=3, Euklides

klasa	precision	recall	f1score
Iris-setosa	1.00	1.00	1.00
Iris-versicolor	1.00	0.95	0.93
Iris-virginica	0.80	1.00	0.89
avg-total	0.96	0.96	0.96

Skuteczność klasyfikatora: 0.96

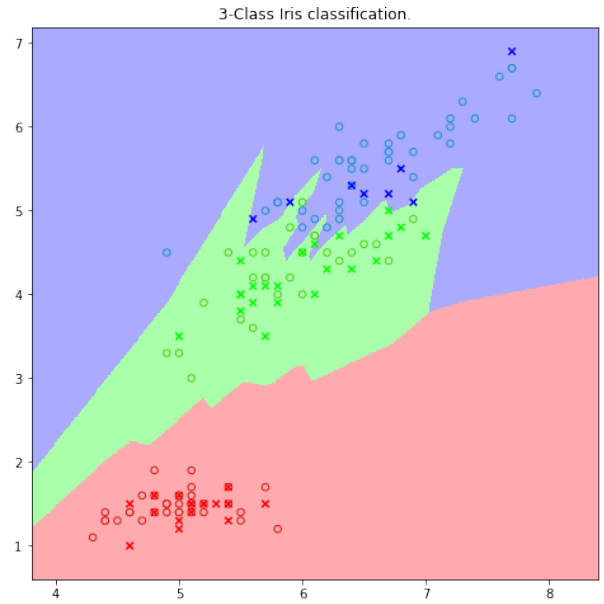


Rysunek 3. Granica decyzyjna dla K-NN z metryką Euklidesa dla k=3

kNN, k=1, Mahalonobis

klasa	precision	recall	f1score
Iris-setosa	1.00	1.00	1.00
Iris-versicolor	0.92	1.00	0.96
Iris-virginica	1.00	0.75	0.86
avg-total	0.96	0.96	0.95

Skuteczność klasyfikatora: 0.96



Rysunek 4. Granica decyzyjna dla K-NN z metryką Mahalanobisa dla k=1

2.4 Analiza zbioru Wine

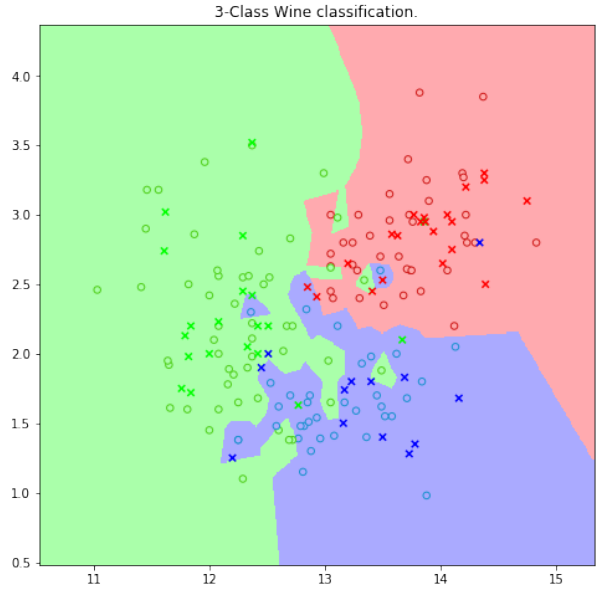
Zbiór wejściowy składa się ze 178 elementów. Klasa pierwsza :59, klasa druga: 71, klasa 3: 48 elementów. Podzieliłem zbiór na dwa zbiory, treningowy i testowy (treningowy: 124, testowy: 54 elementów) w losowy sposób. Aby łatwo zwizualizować granicę decyzyjną wybieram tylko dwie cechy (alcohol, total phenols).

Na poniższych wykresach granic decyzyjnych, kółkami zaznaczone są elementy zbioru treningowe, krzyżykami elementy zbioru testowego. Kolor krzyżyka lub kółka odpowiada klasie przypisanej przez klasyfikator. (Czerwona - klasa 1, niebieskie - klasa 2, zielona - klasa 3)

Large Margin Nearest Neighbours, k=1

klasa	precision	recall	f1score
Klasa 1	0.89	0.81	0.85
Klasa 2	0.84	0.80	0.82
Klasa 3	0.62	0.77	0.69
avg-total	0.81	0.80	0.80

Skuteczność klasyfikatora: 0.8148

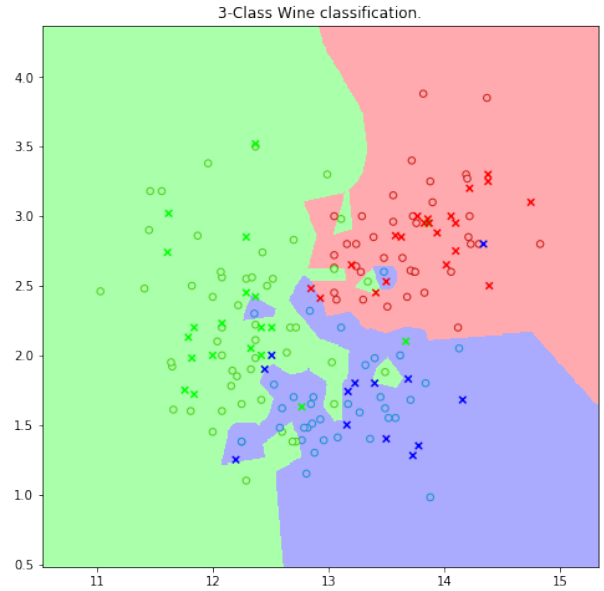


Rysunek 5. Granica decyzyjna dla LMNN dla k=1

k-NN z metryką Euklidesa, k=1

klasa	precision	recall	f1score
Klasa 1	0.90	0.86	0.88
Klasa 2	0.84	0.80	0.82
Klasa 3	0.67	0.77	0.71
avg-total	0.82	0.81	0.81

Skuteczność klasyfikatora: 0.81

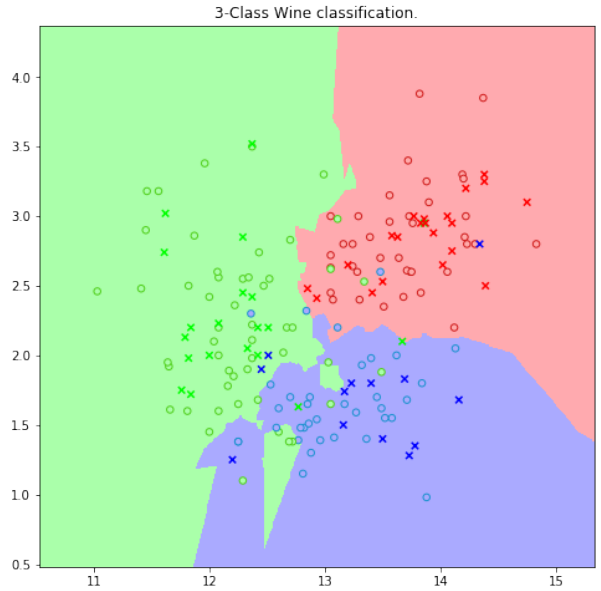


Rysunek 6. Granica decyzyjna dla k-NN dla k=1

k-NN z metryką Euklidesa, k=3

klasa	precision	recall	f1score
Klasa 1	0.91	1.00	0.95
Klasa 2	0.89	0.85	0.87
Klasa 3	0.83	0.77	0.80
avg-total	0.89	0.89	0.89

Skuteczność klasyfikatora: 0.88

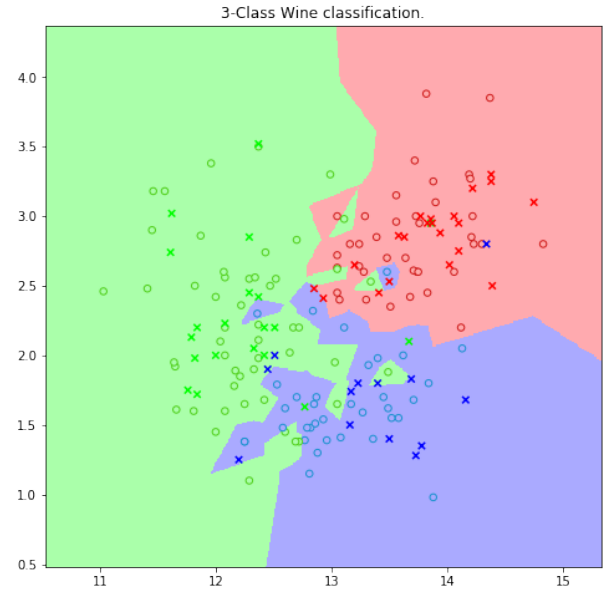


Rysunek 7. Granica decyzyjna dla kNN z metryką Euklidesa dla k=3

Mahalonobis, k=1

klasa	precision	recall	f1score
Klasa 1	0.90	0.90	0.90
Klasa 2	0.85	0.85	0.85
Klasa 3	0.69	0.69	0.69
avg-total	0.83	0.83	0.83

Skuteczność klasyfikatora: 0.83



Rysunek 8. Granica decyzyjna dla kNN z metryką Mahalonobisa dla k=1

2.5 Wnioski

(1) Wybór odpowiedniej metryki dla tego samego klasyfikatora ma ogromny wpływ na jego skuteczność. Wybór lepszej metryki dla danego zbioru zwiększył skuteczność klasyfikatora dla zbioru Irysy o 3 punkty procentowe(różnica między kNN=1 lub LMNN, a k-NN=3 lub k-NN=1 Mahalonibsem), a dla zbioru Wine aż o 7 punktów procentowych(LMNN k=1 lub kNN=1, a kNN=3).

(2) Nie ma dla klasyfikatora k najbliższych sąsiadów najlepszej metryki lub parametru k dla wszystkich możliwych problemów

(3) Wybór klasyfikatora ma ogromne znaczenie aby zminimalizować prawdopodobieństwo wystąpienia pewnych rodzajów błędów (dla zbioru Irys i klasy Iris-virginca i metryki Mahalonobisa precision wynosi 1.00, a recall 0.75, a dla k-NN k=3 z metryką Euklidesa, precision wynosi 0.80, a recall 1.00

(4) Eksperyment podkreśla wagę techniki cross-validation, aby lepiej dobrać metrykę, model i parametry modelu

(5) Bardziej skomplikowane metryki raczej będą miały bardziej skomplikowane granice decyzyjne - dla pewnych zbiorów danych niekoniecznie jest to dobra granica decyzyjna, może zachodzić overfitting (Mahalonobis - Irysy)

LITERATURA

[1] "Fisher, R.A. "The use of multiple measurements in taxo-nomic problems (1936) *Nature*, 135:7–9, 1956.

[2] "Forina, M. et al, PARVUS - An Extendible Package for Data Exploration, Classification and Correlation. Institute of Pharmaceutical and Food Analysis and Technologies"