Informatyka, studia zaoczne, inż I st. semestr V

Sztuczna inteligencja i systemy ekspertowe 2017/2018

Prowadzący: mgr inż. Paweł Tarasiuk sobota, 14:45

Data oddania: _____ Ocena: _____

Natalia Mateuszuk 203940
Adrian Grzelak 200242

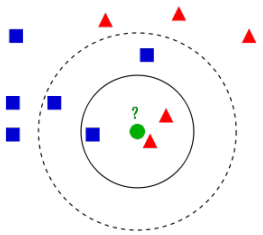
Zadanie 2
Klasyfikator oparty o metodę k najbliższych sąsiadów (k-NN).

1. Wstęp

Celem zadania jest zbadanie klasyfikatora opartego o metodę k najbliższych sąsiadów korzystając z dwóch zbiorów danych [**?**]*iris* oraz [**?**]*wine*. W tym celu metodą 3-krotnej kroswalidacji zostanie podjęta próba zaklasyfikowania każdego z elementów zbioru. Porównane ze sobą zostaną wyniki uzyskane przy pomocy różnych metryk, różnego doboru wartości k, po uprzednim znormalizowaniu zbioru danych jak i uprzednim jego ustandaryzowaniu. Wyniki uzyskane w sprawozdaniu zostały przy pomocy programu `knn.exe`, który specjalnie do tego celu został stworzony.

1.1. Klasyfikator k-NN

Klasyfikator oparty o metodę k najbliższych sąsiadów, jest to klasyfikator, który w celu stwierdzenia przynależności obserwacji do klasy, wyszukuje w zbiorze treningowym k najbliższych dla tej obserwacji sąsiadów (Na podstawie najmniejszych odległości wg. zdefiniowanej metryki). Następnie przyporządkowuje ją do klasy, która najliczniej występowała wśród tych sąsiadów. W przypadku gdy pewne klasy występują równolicznie, wybierana jest ta która ma swój element najbliżej.



Rysunek [**?**] W przypadku k=3, zielona kropka zostanie zakwalifikowana do czerwonych trójkątów. W przypadku k=5 natomiast do niebieskich kwadratów.

1.2. Metryki

W programie zostało zaimplementowane 5 metryk. Jednakże do naszej analizy wykorzystamy jedynie 3 (euklidesowa,manhattan,czebyszewa).

1.2.1. Metryka Euklidesowa

Odległość w $\mathbb{R}^n$ dla metryki euklidesowej zdefiniowana jest następująco:

$$dist_e(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Jest to najbardziej intuicyjna metryka.

1.2.2. Metryka Euklidesowa do kwadratu

Odległość w $\mathbb{R}^n$ dla metryki euklidesowej do kwadratu zdefiniowana jest następująco:

$$dist_{e^2}(x,y) = \sum_{i=1}^{n}(x_i - y_i)^2$$

Metryka która daje takie same wyniki przy klasyfikacji co metryka euklidesowa, jednakże ze względu na to że nie trzeba wykonywać pierwiastkowania to czas obliczeń ulega skróceniu.

1.2.3. Metryka Manhattan

Odległość w $\mathbb{R}^n$ dla metryki manhattan zdefiniowana jest następująco:

$$dist_m(x,y) = \sum_{i=1}^{n}|x_i - y_i|$$

1.2.4. Metryka Czebyszewa

Odległość w $\mathbb{R}^n$ dla metryki Czebyszewa zdefiniowana jest następująco:

$$dist_cz(x,y) = \max_{i=1..n}|x_i - y_i|$$

1.2.5. Metryka Minkowskiego

Odległość w $\mathbb{R}^n$ dla metryki Minkowskiego do potęgi p zdefiniowana jest następująco:

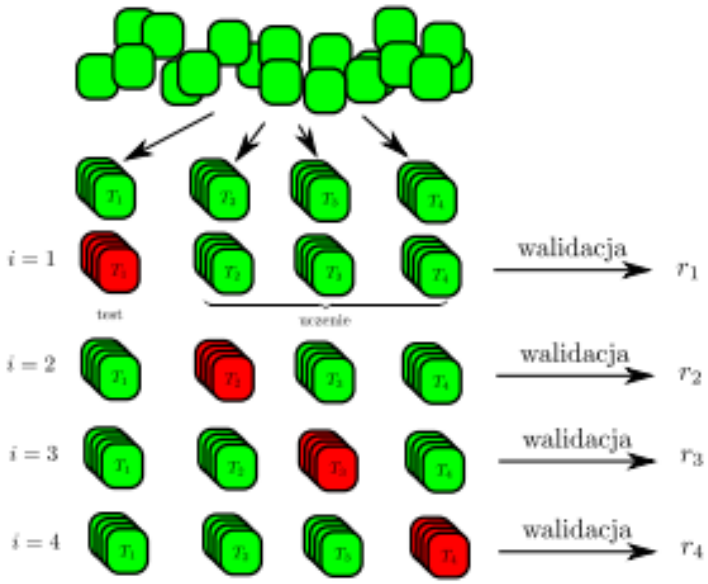
$$dist_{m^p}(x,y) = \left(\sum_{i=1}^{n}(x_i - y_i)^p\right)^{\frac{1}{p}}$$

Można zauważyć że metryka ta jest uogólnieniem dla poprzednich metryk. dla $p = 1$ jest równoważna metryce Manhattan, dla $p = 2$ metryce euklidesowej, dla $p = 4$ metryce euklidesowej do kwadratu i dla $p \to \infty$ metryce Czebyszewa.

1.3. Kroswalidacja

Sprawdzian krzyżowy to metoda statystyczna, polegająca na podziale próby statystycznej na podzbiory, a następnie przeprowadzaniu wszelkich analiz na niektórych z nich (zbiór uczący), podczas gdy pozostałe służą do potwierdzenia wiarygodności jej wyników (zbiór testowy, zbiór walidacyjny). Bez jej zastosowania nie można być pewnym czy model będzie dobrze działał dla danych, które nie były wykorzystywane do jego konstruowania (zob. overfitting).

K-krotna walidacja W tej metodzie, oryginalna próba jest dzielona na K podzbiorów. Następnie kolejno każdy z nich bierze się jako zbiór testowy, a pozostałe razem jako zbiór uczący i wykonuje analizę. Analiza jest więc wykonywana K razy. K rezultatów jest następnie uśrednianych (lub łączonych w inny sposób) w celu uzyskania jednego wyniku.



Rysunek [**?**] Schemat ideowy kroswalidacji

1.4. Błędy pierwszego i drugiego rodzaju

Błąd pierwszego rodzaju (błąd pierwszego typu, alfa-błąd, ang. false positive) - błąd polegający na odrzuceniu hipotezy zerowej, która w rzeczywistości nie jest fałszywa. Oszacowanie prawdopodobieństwa popełnienia błędu pierwszego rodzaju oznaczamy symbolem α (mała grecka litera alfa) i nazywamy poziomem istotności testu. Termin false positive jest często używany w odniesieniu do oprogramowania antywirusowego, które omyłkowo klasyfikuje zdrowy plik jako zainfekowany. Błąd drugiego rodzaju (błąd drugiego typu, błąd przyjęcia, beta-błąd, ang. false negative) - błąd polegający na nieodrzuceniu hipotezy zerowej, która jest w rzeczywistości fałszywa. Oszacowanie

prawdopodobieństwa popełnienia błędu drugiego rodzaju oznaczamy symbolem β (mała grecka litera beta), a jego dopełnienie do jedności nazywane jest mocą testu. W odniesieniu do oprogramowania antywirusowego, błąd drugiego rodzaju polega na niewykryciu szkodliwego kodu. Innymi słowy, zainfekowany plik uznawany jest za bezpieczny. W przypadku diagnoz medycznych, błąd drugiego rodzaju oznacza brak rozpoznania choroby u pacjenta. Tzn. pomimo występowania schorzenia, pacjent nie zostaje rozpoznany względem danego schorzenia.

Prawdopodobieństwo popełnienia błędu drugiego rodzaju wiąże się z tzw. mocą testu statystycznego i oznaczony jest mała grecką literą beta. Moc testu to 1 minus Beta. Błąd I i II rodzaju są ze sobą ściśle powiązane: im częściej odrzucana będzie hipoteza zerowa tym większe ryzyko popełnienia błędu I rodzaju, ale mniejsze ryzyko popełnienia błędu II rodzaju. Z kolei jeśli będziemy bardziej restrykcyjni w poszukiwaniu dowodów na to, że hipoteza zerowa jest nieprawdziwa (nie będziemy jej tak łatwo odrzucać) to oczywiście spada prawdopodobieństwo popełnienia błędu I rodzaju, ale rośnie prawdopodobieństwo popełnienia II rodzaju czyli tzw. moc testu spada.

		HIPOTEZA UZNANA PRZEZ NAS ZA PRAWDZIWĄ	
		H0	**H1**
RZECZYWIŚCIE PRAWDZIWA HIPOTEZA	**H0**	*POPRAWNE PRZYJĘCIE H0*	*BŁĄD I RODZAJU*
	H1	*BŁĄD II RODZAJU*	*POPRAWNE ODRZUCENIE H0*

Rysunek [**?**] Tabela Błędów I i II rodzaju

2. Wyniki

2.1. Tabela wyników

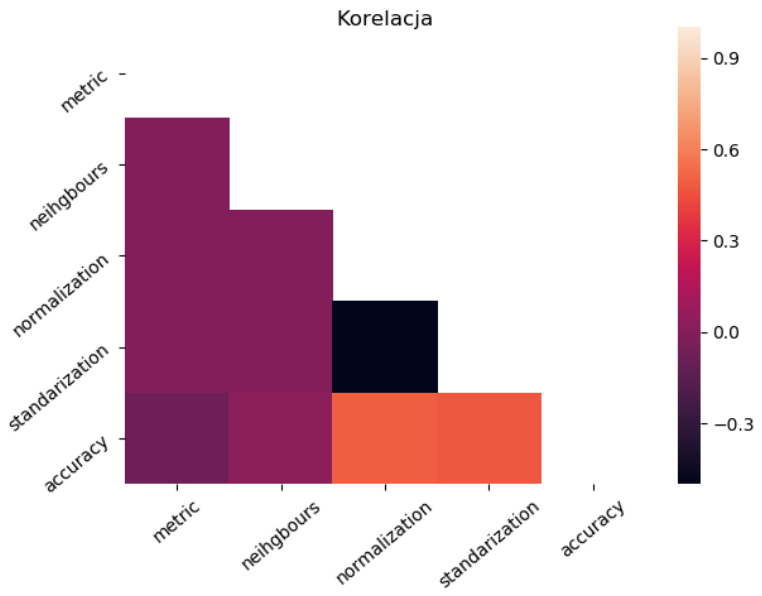
Zbiór Danych	Metryka	k	Normalizacja	Standaryzacja	Trafność
iris	Euclidean	1	no	no	96.00%
iris	Euclidean	3	no	no	96.67%
iris	Euclidean	5	no	no	95.33%
iris	Euclidean	10	no	no	96.00%
iris	Euclidean	1	yes	no	96.00%
iris	Euclidean	3	yes	no	95.33%
iris	Euclidean	5	yes	no	96.00%
iris	Euclidean	10	yes	no	96.67%
iris	Euclidean	1	no	yes	94.00%
iris	Euclidean	3	no	yes	94.67%
iris	Euclidean	5	no	yes	95.33%
iris	Euclidean	10	no	yes	96.67%

iris	Manhattan	1	no	no	96.00%
iris	Manhattan	3	no	no	96.00%
iris	Manhattan	5	no	no	94.67%
iris	Manhattan	10	no	no	95.33%
iris	Manhattan	1	yes	no	94.67%
iris	Manhattan	3	yes	no	94.67%
iris	Manhattan	5	yes	no	94.67%
iris	Manhattan	10	yes	no	95.33%
iris	Manhattan	1	no	yes	93.33%
iris	Manhattan	3	no	yes	95.33%
iris	Manhattan	5	no	yes	94.00%
iris	Manhattan	10	no	yes	95.33%
iris	Chebyshev	1	no	no	96.67%
iris	Chebyshev	3	no	no	96.00%
iris	Chebyshev	5	no	no	98.00%
iris	Chebyshev	10	no	no	98.00%
iris	Chebyshev	1	yes	no	95.33%
iris	Chebyshev	3	yes	no	96.00%
iris	Chebyshev	5	yes	no	96.00%
iris	Chebyshev	10	yes	no	96.67%
iris	Chebyshev	1	no	yes	94.00%
iris	Chebyshev	3	no	yes	94.67%
iris	Chebyshev	5	no	yes	96.00%
iris	Chebyshev	10	no	yes	95.33%
wine	Euclidean	1	no	no	74.16%
wine	Euclidean	3	no	no	68.54%
wine	Euclidean	5	no	no	73.60%
wine	Euclidean	10	no	no	73.60%
wine	Euclidean	1	yes	no	95.51%
wine	Euclidean	3	yes	no	96.63%
wine	Euclidean	5	yes	no	97.19%
wine	Euclidean	10	yes	no	96.63%
wine	Euclidean	1	no	yes	95.51%
wine	Euclidean	3	no	yes	95.51%
wine	Euclidean	5	no	yes	96.63%
wine	Euclidean	10	no	yes	97.19%
wine	Manhattan	1	no	no	81.46%
wine	Manhattan	3	no	no	76.97%
wine	Manhattan	5	no	no	78.65%
wine	Manhattan	10	no	no	79.21%
wine	Manhattan	1	yes	no	97.19%
wine	Manhattan	3	yes	no	97.19%
wine	Manhattan	5	yes	no	98.31%
wine	Manhattan	10	yes	no	98.31%
wine	Manhattan	1	no	yes	97.75%

wine	Manhattan	3	no	yes	97.19%
wine	Manhattan	5	no	yes	97.75%
wine	Manhattan	10	no	yes	97.75%
wine	Chebyshev	1	no	no	69.10%
wine	Chebyshev	3	no	no	68.54%
wine	Chebyshev	5	no	no	73.03%
wine	Chebyshev	10	no	no	71.35%
wine	Chebyshev	1	yes	no	93.26%
wine	Chebyshev	3	yes	no	96.07%
wine	Chebyshev	5	yes	no	93.82%
wine	Chebyshev	10	yes	no	94.94%
wine	Chebyshev	1	no	yes	94.94%
wine	Chebyshev	3	no	yes	93.82%
wine	Chebyshev	5	no	yes	93.82%
wine	Chebyshev	10	no	yes	94.38%

Tabela. Wyniki dla eksperymentów

dla zbioru iris najlepsze wyniki trafności zostały osiągnięte przy stosowaniu metryki Chebysheva, bez stosowania standaryzacji i normalizacji. Dla zbioru wine najlepsze wyniki trafności zostały osiągnięte przy stosowaniu metryki Manhattan, z zastosowaniem normalizacji. Najlepsze wyniki osiągano dla k równego 5 lub 10. Na poniższym obrazku możemy zobaczyć stopień skorelowania pomiędzy poszczególnymi parametrami.



Rysunek Skorelowanie pomiędzy wynikami a parametrami testu dla zbioru danych wine

Od razu rzuca się w oczy że zastosowanie normalizacji ma największy pozytywny wpływ na uzyskaną skuteczność (Silnie dodatnio skorelowana). Na drugim miejscu zaraz za nią plasuje się standaryzacja. Zastosowane liczba sąsiadów k oraz metryka mają również wpływ jednakże nie tak znacz-

ny. Warto zatem przy klasyfikowaniu metodą kNN zastosować normalizację wprowadzanych danych. Warto również pokusić się o sprawdzenie optymalnej liczby sąsiadów. Zbyt mała może przekłamywać niektóre wyniki, podobnie zbyt duża. Sprawdzenie różnych metryk również jest wskazane - na przykładzie naszych zbiorów danych widać, że dla irysów najlepsze wyniki mamy w metryce Czebyszewa, tymczasem dla win w metryce Manhattan.

2.2. Przykładowy test dla iris

```
==================INFO ABOUT TEST================
File: iris.data ,Attributes: 4 ,Instances: 150
Classes: Iris-setosa, Iris-versicolor, Iris-virginica
Method: 3-fold cross-validation
Metric: Euclidean
Nearest neighbours checked: 1
Standarized=no ,Normalization=no
Computing time: 17ms
==================TEST RESULTS==================
Error Matrix:
  50    0    0   | Iris-setosa
   0   47    3   | Iris-versicolor
   0    3   47   | Iris-virginica
Statistics:
 False Positive Ratio (I):
   Iris-setosa - 0.00% (0)
   Iris-versicolor - 6.00% (3)
   Iris-virginica - 6.00% (3)
 False Negative Ratio (II):
   Iris-setosa - 0.00% (0)
   Iris-versicolor - 6.00% (3)
   Iris-virginica - 6.00% (3)
 Accuracy: 96.00%
```

pierwsza grupa testowa ze zbioru irysy podlega bardzo dobrej klasyfikacji, grupa 2 i 3 kwiatów prawdopodobnie jest podobna ponieważ następują pewne pomyłki w klasyfikacji.

2.3. Przykladowy test dla wine

```
==================INFO ABOUT TEST================
File: wine.data ,Attributes: 13 ,Instances: 178
Classes: 1, 2, 3
Method: 3-fold cross-validation
Metric: Euclidean
Nearest neighbours checked: 5
Standarized=no ,Normalization=no
Computing time: 56ms
==================TEST RESULTS==================
Error Matrix:
```

```
  56    0    3   | 1
   4   49   18   | 2
   3   19   26   | 3
Statistics:
 False Positive Ratio (I):
   1 - 11.11% (7)
   2 - 27.94% (19)
   3 - 44.68% (21)
 False Negative Ratio (II):
   1 - 5.08% (3)
   2 - 30.99% (22)
   3 - 45.83% (22)
 Accuracy: 73.60%
```

pierwsza grupa testowa ze zbioru win podlega bardzo dobrej klasyfikacji, grupa 2 i 3 win prawdopodobnie jest podobna ponieważ następują pewne pomyłki w klasyfikacji. W grupie 2 i 3 występuje pewne podobieństwo do grupy 1.

Literatura

[1] Stefan Aeberhard. `http://archive.ics.uci.edu/ml/datasets/Wine`. [Online; Ostatni dostęp 20.01.2017].

[2] R.A. Fisher. `http://archive.ics.uci.edu/ml/datasets/Iris`. [Online; Ostatni dostęp 20.01.2017].

[3] Jimmy Wales. `https://pl.wikipedia.org/`. [Online; Ostatni dostęp 20.01.2017].