Indukcyjna Analiza Danych
Laboratorium 4
Algorytm:
K-najbliższy sąsiadów
K-nearest neighbours

Piotr Błoński

14 maja 2020

Prowadzący: Dr inż. Paweł Myszkowski

1 Cel ćwiczenia

Zapoznanie się z metodą klasyfikacji k-najbliższych sąsiadów (knn, k-nearest neighbours) na przy samodzielnej implementacji przy użyciu języka Python.

2 Plan badań

a. Implementacja algorytmu Knn

b. Implementacja 3 sposobów głosowania :

   (a) większościowe równoprawne

   (b) ważone odległością

   (c) ważone różnicą odległości najbliższego i najdalszego sąsiada danej klasy.

c. Implementacja 2 miar długość : Euklidesowa i Manhatan.

d. Użycie kroswalidacji stratyfikowanej.

e. Użycie standaryzacji

3 Knn - wstęp teoretyczny

Jest to jeden z podstawowych algorytmów nie parametrycznych używany w klasyfikacji i regresji. Należy do kategorii uczenia 'instance-based'. Jego działanie jest dość proste :

a. Przed przekazaniem danych należy je znormalizować / standaryzować

b. Dla zadanej nowej obserwacji obliczamy odległość do znanych już obserwacji za pomocą zadanej funkcji obliczania dystansu. Np. Odległość euklidesowa, Manhattan i inne.

c. Następnie wybieramy K najbliżej leżących obserwacji - dalej nazywamy je 'sąsiadami'.

d. Korzystając z którejś metody głosowania wybieramy do jakieś klasy będzie przynależeć Nowa obserwacja.

Jeśli chodzi o K czyli liczbę sąsiadów to powinniśmy celować w liczbę parzystą aby zmniejszyć ryzyko wystąpienia 'remisów' podczas głosowania. Podczas wystąpienia remisu jedną z metod radzenia sobie z nimi jest użycie wagi (np. dystans do najbliższego) lub zmienienie K. Generalnie im większe K tym knn będzie bardziej odporny na szumy i outlierów kosztem złożoności obliczeniowej. Musimy także pamiętać aby K nie było zbyt duże bo może zdarzyć się sytuacja że K będzie większe od ilości obserwacji danej klasy przez co obserwacja może nie zostać sklasyfikowana poprawnie.

3.1 Lazy learning

Tego terminu używa się do algorytmów uczenia maszynowego które nie tworzą swoistego 'modelu' czyli wyuczonych zestawów parametrów. Podczas 'lazy learnigu' algorytm tylko 'zapamiętuje' dane które są mu podane, nie dokonuje aproksymacji żadnej funkcji. Dopiero w podczas używania modelu model wykonuje większość obliczeń, co skutkuje tym że są o wiele wolniejsze od zwykłych modeli podczas używania, jednakże proces nauki to w większości po prostu podanie danych z odroczeniem obliczeń na później.

3.2 Koszt

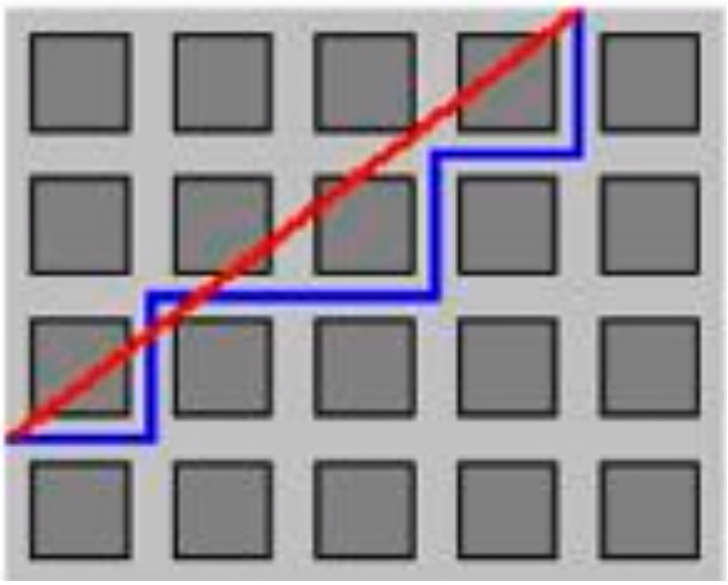
Niestety, ponieważ Knn musi przy każdej predykcji obliczyć odległość każdej obserwacji od nowej obserwacji, jest dość kosztowny obliczeniowo. Im więcej danych tym dłużej zajmuje wykonanie tej operacji, ale także powinna zwiększyć się jego skuteczność.

4 Funkcje obliczania odległości

Note: Jako że dane mamy numeryczne posłużymy się Euclidian i Manhattan distance. Gdybyśmy mieli dane kategoryczne moglibyśmy zastosować np Hamming'a. Obydwie te funkcje należą do kategorii Minkowski distance (p-norm). Gdzie dla Euclidian p=2 a dla manhatann p=1.

$$d\left(p,q\right) = \sqrt[p]{\sum_{i=1}^{n}\left(q_i - p_i\right)^p} \tag{1}$$

Niestety funkcje te są dość wrażliwe na outlierów.



Rysunek 1: Porównanie dystansów Manhatann(niebieski) i Euclidian(czerwony)

Miara dystansu powinna spełniać warunki:

a. Powinna być dodatnia

b. Równa 0 tylko i wyłącznie wtedy gdy x=y

c. Symetryczna

d. $d(x,z) <= d(x,y) + d(y,z)$ Traingle inequality

$d_1(\mathbf{p},\mathbf{q}) = \|\mathbf{p}-\mathbf{q}\|_1 = \sum_{i=1}^{n}|p_i - q_i|$

5 Standaryzacja

Jest to proces przed uczeniem maszynowym w którym 'wyrównujemy dane' tak aby ich rozkład miał średnią wartość = 0 i odchylenie standardowe równe 1.
$z = \frac{x-\mu}{\sigma}$ gdzie z to nowa standaryzowana wartość, x to zmienna niestandaryzowana, μ to średnia z populacji a σ to odchylenie standardowe populacji.

6 Knn - pytania pomocnicze

a. Co jest modelem (klasyfikacji) w algorytmie k-nn? - modelem jest algorytm który po dostaniu danych umie wykonać obliczenia które odpowiedzą na klasyfikacje/regresje.

b. Jak miary odległości wpływają wpływają na skuteczność modelu? -

c. Dlaczego zwykle nie stosuje się parzystych k? - aby uniknąć ryzyka remisów w głosowaniu

7 Wykonane badania

W ramach zajęć laboratoryjnych zostały zmierzone metryki: precision, recall, fscore dla kombinacji 2 różnych funkcji dystansów, 3 metod głosowania, K=[3,5,$\sqrt{N}$],Foldów kroswalidacji = [3,5,10]. Gdzie N jest ilością obserwacji w zbiorze.
Zbiory użyte w badaniach to: Wine, Glass i Seed. A także Iris dla testów czy algorytm w ogóle działa prawidłowo.
Użyto także Standaryzacji z biblioteki Sklearn. Podczas kroswalidacji z każdego foldu zbierane są metryki i następnie uśredniane dla każdej kombinacji.

8 Wnioski

Pragnę zaznaczyć że w załączniku do sprawozdania w postaci kodu i wersji pdf znajdują się także wyniki wyższych foldów które nie zostały wzięte pod uwagę podczas pisania sprawozdania ze względu na bardzo długi czas ewaluacji i zagorzeniu czytelności sprawozdania.

	Set	precision	recall	fscore
0	Iris	0.943915	0.939951	0.940029
[] 0	Glass	0.648614	0.539372	0.558123
0	Wine	0.318155	0.260635	0.256092
0	seeds	0.907945	0.900161	0.899600

Tabela 1: Głosowanie większościowe Euclidean distance Głosowanie Kroswalidacja 3 Fold

	Set	precision	recall	fscore
0	Iris	0.954784	0.953023	0.953340
[] 0	Glass	0.599691	0.533169	0.546775
0	Wine	0.304031	0.237863	0.238708
0	seeds	0.914209	0.904388	0.904959

Tabela 2: Głosowanie większościowe Manhatan distance Kroswalidacja 3 Fold K = 3

	Set	precision	recall	fscore
0	Iris	0.943915	0.939951	0.940029
0	Glass	0.648614	0.539372	0.558123
0	Wine	0.318155	0.260635	0.256092
0	seeds	0.907945	0.900161	0.899600

Tabela 3: Głosowanie większościowe Euclidean distance Kroswalidacja 3 Fold K = 3

	Set	precision	recall	fscore
0	Iris	0.947476	0.946487	0.946618
[] 0	Glass	0.596217	0.577580	0.561213
0	Wine	0.250178	0.249978	0.246672
0	seeds	0.899740	0.881441	0.882888

Tabela 4: Głosowanie ważone dystansem Euclidean distance Kroswalidacja 3 Fold K = 3

	Set	precision	recall	fscore
0	Iris	0.941815	0.939951	0.940009
[] 0	Glass	0.614848	0.626664	0.598928
0	Wine	0.237315	0.240005	0.236117
0	seeds	0.870634	0.852858	0.854478

Tabela 5: Głosowanie ważone dystansem Manhattan distance Kroswalidacja 3 Fold K = 3

	Set	precision	recall	fscore
0	Iris	0.943915	0.939951	0.940029
[] 0	Glass	0.648614	0.539372	0.558123
0	Wine	0.380879	0.257533	0.255441
0	seeds	0.907945	0.900161	0.899600

Tabela 6: Głosowanie ważone różnicą dystansów Euclidean distance Kroswalidacja 3 Fold K = 3

	Set	precision	recall	fscore
0	Iris	0.954784	0.953023	0.953340
[] 0	Glass	0.599691	0.533169	0.546775
0	Wine	0.358523	0.230684	0.231764
0	seeds	0.914209	0.904388	0.904959

Tabela 7: Głosowanie ważone różnicą dystansów Manhattan distance Kroswalidacja 3 Fold K = 3

	Set	precision	recall	fscore
0	Iris	0.961057	0.959967	0.960136
[] 0	Glass	0.700044	0.509188	0.507590
0	Wine	0.493459	0.255739	0.257422
0	seeds	0.931249	0.923913	0.923804

Tabela 8: Głosowanie większościowe Euclidean distance Kroswalidacja 3 Fold K = 5

	Set	precision	recall	fscore
0	Iris	0.943396	0.939542	0.939846
[] 0	Glass	0.562438	0.478703	0.471956
0	Wine	0.535107	0.243418	0.242237
0	seeds	0.909118	0.899960	0.900356

Tabela 9: Głosowanie większościowe Manhatan distance Kroswalidacja 3 Fold K = 5

	Set	precision	recall	fscore
0	Iris	0.909469	0.906454	0.907116
[] 0	Glass	0.615712	0.557465	0.537253
0	Wine	0.242934	0.252855	0.241872
0	seeds	0.887757	0.871981	0.874037

Tabela 10: Głosowanie większościowe Euclidean distance Kroswalidacja 3 Fold K = 5

	Set	precision	recall	fscore
0	Iris	0.933996	0.933007	0.933138
[] 0	Glass	0.622636	0.617118	0.593237
0	Wine	0.238920	0.244781	0.237962
0	seeds	0.877752	0.857488	0.859146

Tabela 11: Głosowanie ważone dystansem Euclidean distance Kroswalidacja 3 Fold K = 5

	Set	precision	recall	fscore
0	Iris	0.949343	0.946078	0.946242
[] 0	Glass	0.579951	0.482233	0.488873
0	Wine	0.433616	0.246722	0.249129
0	seeds	0.917647	0.909420	0.909347

Tabela 12: Głosowanie ważone dystansem Manhattan distance Kroswalidacja 3 Fold K = 5

	Set	precision	recall	fscore
0	Iris	0.949343	0.946078	0.946242
[] 0	Glass	0.579951	0.482233	0.488873
0	Wine	0.433616	0.246722	0.249129
0	seeds	0.917647	0.909420	0.909347

Tabela 13: Głosowanie ważone różnicą dystansów Euclidean distance Kroswalidacja 3 Fold K = 5

	Set	precision	recall	fscore
0	Iris	0.949261	0.946078	0.946518
[] 0	Glass	0.505927	0.446342	0.437612
0	Wine	0.383139	0.245301	0.247751
0	seeds	0.921946	0.914050	0.913776

Tabela 14: Głosowanie ważone różnicą dystansów Manhattan distance Kroswalidacja 3 Fold K = 5

9 Omówienie

W pokazanych w sprawozdaniu modeli najlepiej sprawował się :

a. Iris - 0.960 - Głosowanie większościowe Euclidean distance Kroswalidacja 3 Fold K= 5

b. Glass - 0.598 - Głosowanie ważone dystansem Manhattan distance Kroswalidacja 3 Fold K = 3

c. Wine - 0.256 - Głosowanie większościowe Euclidean distance Kroswalidacja 3 Fold K = 3

d. seeds - 0.923 - Głosowanie większościowe Euclidean distance Kroswalidacja 3 Fold K= 5

9.1 Wyniki z poprzednich laboratorium

Set	Fscore	
Iris	0.946	
Glass	0.736	Wyniki dla naiwnego bayesa z kroswaldacją 5 Fold
Wine	0.83	
Seed	0.77	

W porównaniu z Naiwnym Bayesem z pierwszych laboratorium można zauważyć że Knn gorzej radzi się z wielowymiarowymi danymi. Wyniki rzędu 0.256 dla Wine jest naprawdę kiepski.