**METODY EKSPLORACJI DANYCH**

Laboratorium. Klasyfikacja na podstawie klasyfikatora bayesowskiego

i najbliższego sąsiedztwa

Prowadzący: Wykonali:

dr inż. Romuald Hoffmann pchor. Michał ADAMCZEWSKI

pchor. Mikołaj ADAMSKI

pchor. Przemysław SUJECKI

Zadanie 1.

Obraz zawierający tekst, zrzut ekranu, Czcionka, numer

Opis wygenerowany automatycznie

Naiwny klasyfikator Bayesa

Do budowy modeli klasyfikacji wykorzystuje się **naiwny klasyfikator Bayesa**.

Prawdopodobieństwo warunkowe (a posteriori), że obserwacja o wartości zaobserwowanej X pochodzi z klasy C wynosi:

gdzie:

- Pr{C} jest prawdopodobieństwem bezwarunkowym (a priori) wystąpienia klasy C,

- Pr{X|C} jest prawdopodobieństwem warunkowym, że obserwacja o wartości X należy do klasy C,

- Pr{X} jest prawdopodobieństwem bezwarunkowym wystąpienia obserwacji X.

W klasyfikacji bayesowskiej, zgodnie z regułą Bayesa, obiekt X przypisujemy do klasy Ci, dla której prawdopodobieństwo warunkowe Pr{Ci|X}, i = 1,2,…,m, jest największe, tzn. wyznaczamy wszystkie prawdopodobieństwa i wybieramy to największe.

Do wyznaczania prawdopodobieństwa Pr{Ci|X} jest wykorzystywane twierdzenie Bayesa:

Prawdopodobieństwo Pr{𝐶i} możemy estymować względną częstością występowania klasy w zbiorze testowym 𝑍 następująco:

Jeżeli założymy, że wszystkie klasy Ci mają to samo prawdopodobieństwo wystąpienia, to:

Wzór Bayesa przy założeniu „niezależności klas”:

Prawdopodobieństwo jest praktycznie niemożliwe do wyliczenia, stąd przyjmujemy naiwne założenie niezależności atrybutów w zbiorze danych:

Naszym zadaniem więc będzie:

Prawdopodobieństwa warunkowe wyliczymy w następujący sposób:

Algorytm k-NN

Do budowy modeli klasyfikacji wykorzystuje się również **algorytm k-NN** najbliższych sąsiadów. Algorytm ten określa miarę odległości między badanymi (analizowanymi) obiektami oraz określeniu liczby k najbliższych sąsiadów, wg. przyjętej miary odległości. W zadaniu laboratoryjnym wykorzystamy dwie miary odległości:

1. euklidesową, której bazą jest norma L2,
2. miejską (tzw. Manhattan) – norma L1.

Odległość euklidesowa:

Odległość miejska:

Ogólny algorytm postępowania wygląda następująco:

1. Obliczamy odległości badanego nowego obiektu y od pozostałych obiektów zbioru x,
2. Określamy parametr k, który określa liczbę obiektów ze zbioru x determinujących decyzję o przynależności do konkretnej klasy,
3. Wybieramy k-najbliższych sąsiadów w sensie przyjętej odległości. W przypadku identycznych odległości wliczamy wszystkie punkty o tej samej odległości do k-NN,
4. Porównujemy nowy obiekt y z k-najbliższymi sąsiadami i wybieramy przynależność badanego obiektu do konkretnej klasy poprzez głosowanie proste lub ważone.

Głosowanie proste – przypisanie jednego głosu jednemu obiektowi.

Głosowanie ważone – wpływ poszczególnych obiektów jest tym większy im mniejsza jest odległość od obiektu badanego.

Należy wybrać odpowiedni parametr k, aby wartość nie była zbyt mała oraz zbyt wysoka. W pierwszym przypadku klasyfikator będzie zbyt podatny na szumy w danych. W drugim przypadku czas działania algorytmu zostanie wydłużony oraz przeoczone mogą zostać ważne związki („wygładzenie predykcji”).

Czym różnią się dane odległości:

Euklidesowa: Mierzy najkrótszą odległość między dwoma punktami.

Miejska (Manhattan): Mierzy sumę bezwzględnych różnic między współrzędnymi.

Czebyszewa: Mierzy największą bezwzględną różnicę między współrzędnymi.

Zadanie 1.

Naiwny klasyfikator bayesowki

Tabele częstotliwości, w której pokazujemy liczbę zagranych meczy na podstawie warunków pogodowych:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Siła wiatru | zagrano | nie zagrano | Razem | zagranych | nie zagranych |
| silny | 2 | 2 | 10 | 6 | 4 |
| słaby | 2 | 0 |  |  |  |
| brak | 2 | 2 |  |  |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Zachmurzenie | zagrano | nie zagrano | Razem | zagranych | nie zagranych |
| pochmurnie | 3 | 3 | 10 | 6 | 4 |
| słonecznie | 3 | 1 |  |  |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Odczuwalna temp | zagrano | nie zagrano | Razem | zagranych | nie zagranych |
| zimno | 1 | 2 | 10 | 6 | 4 |
| ciepło | 3 | 1 |  |  |  |
| gorąco | 2 | 1 |  |  |  |

Prawdopodobieństwo a priori P(A) – prawdopodobieństwo samego zdarzenia A, przed uwzględnieniem jakichkolwiek informacji, w naszym przypadku siły wiatru, zachmurzenia i odczuwalnej tempretaury:

P(zagrano) = 0,6

P(nie zagrano) = 0,4



Prawdopodobieństwa warunkowe dla każdej cechy, odwrotność a posteriori inaczej nazywane jako wiarygodność zdarzenia P(B|A), np. silny wiatr pojawiał się podczas zagranego meczu:

Wiatr:

P(silny|zagrano) = 0,33333 P(silny|nie zagrano) = 0,5

Zachmurzenie:

P(słonecznie|zagrano) = 0,5 P(słonecznie|nie zagrano) = 0,25

Odczuwalna temperatura:

P(ciepło|zagrano) = 0,5 P(ciepło|nie zagrano) = 0,25

Prawdopodobieństwo a posteriori:

P(zagrano|silny,słonecznie,ciepło) =

P(nie zagrano|silny,słonecznie,ciepło) = = 0,0125

Wyliczenie prawdopodobieństwa P(silny,słonecznie,ciepło) na 2 sposoby:

1. Suma zagranych i nie zagranych meczy w podanych warunkach pogodowych (silny, słonecznie, ciepło) dzielona przez liczbę meczy (10);
2. Suma prawdopodobieństw P(zagrano|silny,słonecznie,ciepło) oraz P(nie zagrano|silny,słonecznie,ciepło);

ad. a) P(silny,słonecznie,ciepło) = 0,064

ad. b) P(silny,słonczenie,ciepło) = 0,0625

Końcowe prawdopodobieństwo:

P(zagrano|silny,słonecznie,ciepło) = = 0,8

P(nie zagrano|silny,słonecznie,ciepło) = = 0,2

Wniosek: wyniki wskazują na to, że z dużym prawdopodobieństwem mecz się odbędzie.

Metoda k-NN

|  |  |
| --- | --- |
| Wiatr |  |
| silny | 2 |
| słaby | 1 |
| brak | 0 |

|  |  |
| --- | --- |
| Zachmurzenie |  |
| słonecznie | 2 |
| pochmurnie | 1 |

|  |  |
| --- | --- |
| Temperatura |  |
| gorąco | 2 |
| ciepło | 1 |
| zimno | 0 |

Odległość Euklidesowa:

Dla każdej obserwacji liczymy odległość euklidesową dla nowego przypadku:

|  |  |
| --- | --- |
|  | nowy |
| ciepło | 1 |
| słonecznie | 2 |
| silny | 2 |

Najbliższe są przypadki dla q = 2,3,6,8,10. W każdym oprócz 6 nie zagrano meczu, więc naturalnie klasyfikujemy, że mecz zostanie rozegrany.

Odległość miejska:

Najbliższe są przypadki dla q = 1,2,3,6,8,10. Zdecydowana większość zagrała mecz, więc klasyfikujemy, że mecz się odbędzie.

Bibliografia:

[0] dr inż. Romuald Hoffmann, prof. WAT, Notatki dla studentów, Warszawa 2023.

[1] Larose D. T., Metody i modele eksploracji danych, Wydawnictwo Naukowe PWN, Warszawa 2012.

[2] <https://scikit-learn.org/stable/>

[3] Fred Nwanganga, Mike Chapple, Praktyczne uczenie maszynowe w języku R, Wiley, Warszawa, 2022.

[4] Laurence Moroney, Sztuczna inteligencja i uczenie maszynowe dla programistów, Helion O’Reilly, 2021.

[5] <https://pl.wikipedia.org/wiki/Wikipedia:Strona_g%C5%82%C3%B3wna>

[6] Aurelie Geron, Uczenie maszynowe z użyciem Scikit-Learn i TensorFlow, Helion O’Reilly, 2018