**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 2.

Obraz zawierający tekst, numer, Czcionka, Równolegle

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”).

Zadanie 2.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Nr wiad. | Słowa kluczowe | | | | | d Klasyfikacja: spam |
| x1 | x2 | x3 | x4 | x5 |
| pieniądz | darmowy | bogaty | nieprzyzwoicie | tajny |
| 1 | 0 | 0 | 1 | 0 | 1 | 1 |
| 2 | 1 | 1 | 1 | 0 | 0 | 1 |
| 3 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 0 | 1 | 0 | 0 | 0 | 1 |
| 5 | 1 | 0 | 0 | 0 | 0 | 0 |
| 6 | 0 | 1 | 0 | 1 | 1 | 1 |
| 7 | 0 | 1 | 0 | 1 | 0 | 1 |
| 8 | 0 | 0 | 0 | 1 | 0 | 1 |
| 9 | 0 | 1 | 0 | 0 | 0 | 0 |
| 10 | 0 | 0 | 0 | 0 | 1 | 0 |
| 11 | 1 | 1 | 1 | 0 | 1 | 1 |
| 12 | 1 | 0 | 0 | 0 | 1 | 1 |
| 13 | 0 | 1 | 1 | 0 | 0 | 0 |
| 14 | 1 | 0 | 1 | 0 | 1 | ?????? |

Dla słów kluczowych "tak" i "nie", można przyporządkować wartości binarne:

"tak" = 1,

"nie" = 0.

1. Opierając się na klasyfikacji bayesowskiej proszę określić status wiadomości/posta nr 14 na podstawie występowania lub niewystępowania słów kluczowych.

Naiwny klasyfikator Bayesa przyporządkował nowy przypadek do klasy 𝐶2 = {𝑑 = 1} spam.

Podsumowanie:

Wynik otrzymany z naiwnej klasyfikacji bayesowskiej wskazuje, że nowy przypadek został przyporządkowany do klasy (spamu). Do zalet klasyfikacji bayesowskiej należy prosta implementacja i szybkość działania. Do wad - złożenie, równość wagi słów (tzn. każde słowo kluczowe traktowane jest jako równie ważne), brak uwzględnienia kontekstu.

W związku z tym, choć naiwna klasyfikacja bayesowska może być stosunkowo prosta i szybka, jej skuteczność w wykrywaniu spamu może być ograniczona, zwłaszcza wobec bardziej zaawansowanych metod i bardziej złożonych danych.

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