**Przedmiot**: Metody Eksploracji Danych

Sprawozdanie z laboratoriów

**Prowadzący**: dr inż. Romuald Hoffmann

**Osoby wykonujące:**

Mateusz Gajda

Kuba Kaczmarski

Radosław Relidzyński

**Grupa**: WCY20IJ1S1

Laboratoria 1

* Dane

Tabela 1. Liczba użytkowników portalu społecznościowego „Facebook”

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Kwartał | Liczba użytkowników w mln | | Kwartał | Liczba użytkowników w mln | | Kwartał | Liczba użytkowników w mln |
| Q3 '08 | 100 |  | Q1 '09 | 197 |  | Q1 '14 | 1276 |
|  |  |  | Q2 '09 | 242 |  | Q2 '14 | 1317 |
|  |  |  | Q3 '09 | 305 |  | Q3 '14 | 1350 |
|  |  |  | Q4 '09 | 360 |  | Q4 '14 | 1393 |
|  |  |  | Q1 '10 | 431 |  | Q1 '15 | 1441 |
|  |  |  | Q2 '10 | 482 |  | Q2 '15 | 1490 |
|  |  |  | Q3 '10 | 550 |  | Q3 '15 | 1545 |
|  |  |  | Q4 '10 | 608 |  | Q4 '15 | 1591 |
|  |  |  | Q1 '11 | 680 |  | Q1 '16 | 1654 |
|  |  |  | Q2 '11 | 739 |  | Q2 '16 | 1712 |
|  |  |  | Q3 '11 | 800 |  | Q3 '16 | 1788 |
|  |  |  | Q4 '11 | 845 |  | Q4 '16 | 1860 |
|  |  |  | Q1 '12 | 901 |  | Q1 '17 | 1936 |
|  |  |  | Q2 '12 | 955 |  | Q2 '17 | 2006 |
|  |  |  | Q3 '12 | 1007 |  | Q3 '17 | 2072 |
|  |  |  | Q4 '12 | 1056 |  | Q4 '17 | 2129 |
|  |  |  | Q1 '13 | 1110 |  |  |  |
|  |  |  | Q2 '13 | 1155 |  |  |  |
|  |  |  | Q3 '13 | 1189 |  |  |  |
|  |  |  | Q4 '13 | 1228 |  |  |  |

Tabela 2. Przychody, zysk i zatrudnienie przedsiębiorstwa „Facebook”

|  |  |  |  |
| --- | --- | --- | --- |
| Rok | Przychód w mln $ | Zysk w mln $ | Zatrudnienie |
| 2007 | 153 | -138 | 450 |
| 2008 | 272 | -56 | 850 |
| 2009 | 777 | 229 | 1218 |
| 2010 | 1974 | 606 | 2127 |
| 2011 | 3711 | 1000 | 3200 |
| 2012 | 5089 | 53 | 4619 |
| 2013 | 7872 | 1500 | 6337 |
| 2014 | 12466 | 2940 | 9199 |
| 2015 | 17928 | 3688 | 12691 |
| 2016 | 27638 | 10217 | 17048 |
| 2017 | 40653 | 15934 | 25105 |

* Pierwsza Analiza

Pytanie: **Jaka będzie liczba użytkowników w najbliższych latach ?**

Model: **Regresja Liniowa**

Na podstawie wybranej metody obliczamy parametry strukturalne, odchylenie standardowe i miary dopasowania.

Model 1: Estymacja KMNK, wykorzystane obserwacje 1-25

Zmienna zależna (Y): LiczbauAytkownikAwwmln

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | *Współczynnik* | *Błąd stand.* | *t-Studenta* | *wartość p* |  |
| const | 110,220 | 14,0044 | 7,870 | <0,0001 | \*\*\* |
| KwartaA | 53,9092 | 0,942037 | 57,23 | <0,0001 | \*\*\* |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Średn.aryt.zm.zależnej | 811,0400 |  | Odch.stand.zm.zależnej | 398,1520 |
| Suma kwadratów reszt | 26534,25 |  | Błąd standardowy reszt | 33,96562 |
| Wsp. determ. R-kwadrat | 0,993026 |  | Skorygowany R-kwadrat | 0,992723 |
| F(1, 23) | 3274,844 |  | Wartość p dla testu F | 2,62e-26 |
| Logarytm wiarygodności | −122,5649 |  | Kryt. inform. Akaike'a | 249,1298 |
| Kryt. bayes. Schwarza | 251,5676 |  | Kryt. Hannana-Quinna | 249,8059 |

Współczynnik determinacji, oznaczany jako , jest miarą, która służy do oceny jakości dopasowania modelu regresji do danych empirycznych.  jest liczbą z zakresu 0 do 1, gdzie 0 oznacza brak dopasowania modelu do danych, a 1 oznacza pełne dopasowanie modelu do danych. Definicja współczynnika determinacji jest następująca:

=

Gdzie:

- Wartość modelu zmiennej objaśnianej Y

– Średnia arytmetyczna z obserwacji

– wartość (zaobserwowanej) zmiennej objaśnianej Y (Obserwacja)

Obliczone przez nas parametry regresji liniowej wyliczane są na podstawie metody najmniejszych kwadratów. Metoda najmniejszych kwadratów przedstawiona jest następującym wzorem:

Gdzie:

Wartość zaobserwowanej zmiennej niezależnej (W tym przypadku to Kwartał)

Wartość (zaobserwowanej) zmiennej objaśnianej Y (Obserwacja, w tym przypadku liczba aktywnych użytkowników)

Parametry obliczane za pomocą minimalizacji funkcji (W tym przypadku wartości współczynnika const i KwartaA)

Cała funkcja jest liczona za pomocą macierzy(Przyspiesza to obliczenia). Wykorzystując zapis macierzowy wygląda ona następująco:

Obraz zawierający wykres

Opis wygenerowany automatycznie

Rysunek 1 - Wykres zależności liczby użytkowników od kwartałów z lat 2008 - 2014

Obraz zawierający wykres

Opis wygenerowany automatycznie

Rysunek 2 - Wykres zależności liczby użytkowników od kwartału z 95% przedziałem ufności dla lat 2015 – 2017

Wnioski:

Wybrany przez nas model regresji liniowej okazał się poprawny. Liczba użytkowników cały czas wzrasta w kolejnych kwartałach. Dodatkowo punkty na wykresie znajdują się w okolicy linii regresji oraz wyniki obserwacji z lat 2014 – 2017 mieszczą się w 95 % przedziale ufności.

* Druga Analiza

Pytanie: **Jaki będzie koszt w zależności od zatrudnienia ?**

Model: **Regresja Sześcienna**

Na podstawie wybranej metody obliczamy parametry strukturalne, odchylenie standardowe i miary dopasowania.

Model 2: Estymacja KMNK, wykorzystane obserwacje 1-11

Zmienna zależna (Y): Koszt

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | *Współczynnik* | *Błąd stand.* | *t-Studenta* | *wartość p* |  |
| const | −668,239 | 333,422 | −2,004 | 0,0851 | \* |
| Zatrudnienie | 1,13764 | 0,159187 | 7,147 | 0,0002 | \*\*\* |
| Zatrudnienie\_kw | 1,99769e-06 | 1,64652e-05 | 0,1213 | 0,9068 |  |
| Zatrudnienie\_3 | −2,84341e-010 | 4,37230e-010 | −0,6503 | 0,5362 |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Średn.aryt.zm.zależnej | 7505,455 |  | Odch.stand.zm.zależnej | 8149,045 |
| Suma kwadratów reszt | 1489134 |  | Błąd standardowy reszt | 461,2304 |
| Wsp. determ. R-kwadrat | 0,997758 |  | Skorygowany R-kwadrat | 0,996797 |
| F(3, 7) | 1038,201 |  | Wartość p dla testu F | 1,24e-09 |
| Logarytm wiarygodności | −80,59528 |  | Kryt. inform. Akaike'a | 169,1906 |
| Kryt. bayes. Schwarza | 170,7821 |  | Kryt. Hannana-Quinna | 168,1873 |

Współczynnik determinacji, oznaczany jako , ma identyczną definicję jak w przypadku pierwszej analizy i jest obliczany tak samo.

Obliczone przez nas parametry regresji sześciennej również są wyliczone na podstawie metody najmniejszych kwadratów. W tym wypadku jednak funkcja ta jest sześcienna i wygląda następująco:

Następnie funkcja ta jest sprowadzana do modelu liniowego (proces linearyzacji).

Gdzie:

Zatem równoważny model liniowy jest następujący:

W kolejnym punkcie uzyskany model jest liczony analogicznie do pierwszej analizy.

Obraz zawierający tekst, linia, diagram, zrzut ekranu

Opis wygenerowany automatycznie

Rysunek 3 - Wykres zależności kosztów od zatrudnienia z lat 2007 - 2017

Obraz zawierający tekst, diagram, linia, Wykres

Opis wygenerowany automatycznie

Rysunek 4 - Wykres zależności kosztów od zatrudnienia z 95% przedziałem ufności dla lat 2018 – 2022

Wnioski:

Dla podanych danych model regresji sześciennej wydaje się prawidłowy (punkty leżą blisko wyznaczonej linii). Dodatkowo przy predykcji 95% na lata 2018 - 2022 wartość kosztu znajduje się w granicach przedziału ufności dla funkcji zatrudnienia od kosztów. Niestety dla wartości w 2018 nie znajduje się w tym przedziale, a jest to spowodowane niezależnym i niespodziewanym czynnikiem zewnętrznym który wpłynął na zatrudnienie.

Laboratoria 2

* Zadanie 1

W wybranej losowo grupie studentów jednolitych studiów magisterskich z warszawskich uczelni badano ich stan cywilny w zależności od roku studiów. Wyniki obserwacji zebrano w tabeli 1. Przyjęto przez M oznaczać osoby będące w związku małżeńskim, natomiast przez W –osoby stanu wolnego.

**Dane:**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Rok studiów | 1 | 2 | 5 | 1 | 4 | 3 | 2 | 1 | 5 | 2 | 3 | 4 | 1 | 2 |
| Stan cywilny | W | W | M | W | M | W | M | W | M | W | M | M | W | W |
| Rok studiów | 5 | 4 | 3 | 1 | 4 | 5 | 2 | 5 | 3 | 4 | 3 | 2 | 5 | 1 |
| Stan cywilny | M | M | M | M | M | W | W | M | W | W | M | W | M | W |

**Wyznaczanie zależności stanu cywilnego badanych studentów – liniowy model prawdopodobieństwa:**

Obliczone przez nas parametry modelu liniowego wyliczane są na podstawie metody najmniejszych kwadratów. Metoda najmniejszych kwadratów przedstawiona jest następującym wzorem:

Gdzie:

Wartość zaobserwowanej zmiennej niezależnej

Wartość (zaobserwowanej) zmiennej objaśnianej Y

Parametry obliczane za pomocą minimalizacji funkcji

Cała funkcja jest liczona za pomocą macierzy(Przyspiesza to obliczenia). Wykorzystując zapis macierzowy wygląda ona następująco:

Cały powyższy opis dotyczy liniowego modelu prawdopodobieństwa.

Tabela liniowego modelu zależności prawdopodobieństwa stanu wolnego od roku:

|  |  |  |  |
| --- | --- | --- | --- |
| **Rok studiów** | **W** | **Na kierunku** | **Wartość empiryczna** |
| 1 | 5 | 6 | 0,833333333 |
| 2 | 5 | 6 | 0,833333333 |
| 3 | 2 | 5 | 0,4 |
| 4 | 1 | 5 | 0,2 |
| 5 | 1 | 6 | 0,166666667 |

Wartość empiryczna to prawdopodobieństwo bycia w stanie wolnym na danym roku studiów (W/liczba osób na kierunku).

Obliczamy parametry strukturalne, odchylenie standardowe i miary dopasowania:

Model 1: Estymacja KMNK, wykorzystane obserwacje 1-5

Zmienna zależna (Y): WartoscEmpirycznaW

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | *Współczynnik* | *Błąd stand.* | *t-Studenta* | *wartość p* |  |
| const | 0,943333 | 0,162856 | 5,792 | 0,0102 | \*\* |
| RokstudiAw | −0,163333 | 0,0491031 | −3,326 | 0,0448 | \*\* |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Średn.aryt.zm.zależnej | 0,453333 |  | Odch.stand.zm.zależnej | 0,291166 |
| Suma kwadratów reszt | 0,072333 |  | Błąd standardowy reszt | 0,155278 |
| Wsp. determ. R-kwadrat | 0,786697 |  | Skorygowany R-kwadrat | 0,715596 |
| F(1, 3) | 11,06452 |  | Wartość p dla testu F | 0,044842 |
| Logarytm wiarygodności | 3,495078 |  | Kryt. inform. Akaike'a | −2,990155 |
| Kryt. bayes. Schwarza | −3,771279 |  | Kryt. Hannana-Quinna | −5,086615 |

Obraz zawierający tekst, linia, zrzut ekranu, Wykres

Opis wygenerowany automatycznie

Rysunek 5 - Zależność prawdopodobieństwa w stanie wolnym od roku studiów

**Wyznaczanie zależności stanu cywilnego badanych studentów – model logitowy:**

W modelu logitowym wartość oczekiwana zmiennej objaśnianej może być interpretowana jako warunkowe prawdopodobieństwo realizacji danego zdarzenia przy ustalonych wartościach zmiennej objaśniającej lub zmiennych objaśniających. Dodatkowo stosuje się tak zwaną transformację logitową:

Gdzie:

p – prawdopodobieństwo wystąpienia zdarzenia

– Iloraz szans

Liniową postać modelu logitowego można przedstawić w postaci:

L =

Gdzie:

X – Macierz zmiennych objaśniających (W tym przypadku macierz Lat)

– współczynniki funkcji liniowej

Stosując przekształcenie odwrotne do transformacji możemy otrzymać wzór na prawdopodobieństwo p:

=

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Rok studiów** | **W** | **Na kierunku** |  | **1 -** |  | **Ln()** |
| 1 | 5 | 6 | 0,833333 | 0,166667 | 5 | 1,609438 |
| 2 | 5 | 6 | 0,833333 | 0,166667 | 5 | 1,609438 |
| 3 | 2 | 5 | 0,4 | 0,6 | 0,666667 | -0,40547 |
| 4 | 1 | 5 | 0,2 | 0,8 | 0,25 | -1,38629 |
| 5 | 1 | 6 | 0,166667 | 0,833333 | 0,2 | -1,60944 |

Gdzie:

– prawdopodobieństwo bycia w stanie wolnym

1 - – prawdopodobieństwo zdarzenia przeciwnego

Ln() – Logit

Obliczamy parametry strukturalne, odchylenie standardowe i miary dopasowania:

Model 2: Estymacja KMNK, wykorzystane obserwacje 1-5

Zmienna zależna (Y): Lt

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | *Współczynnik* | *Błąd stand.* | *t-Studenta* | *wartość p* |  |
| const | 2,06055 | 0,796102 | 2,588 | 0,0812 | \* |
| RokstudiAw | −0,760090 | 0,240034 | −3,167 | 0,0506 | \* |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Średn.aryt.zm.zależnej | −0,219722 |  | Odch.stand.zm.zależnej | 1,369841 |
| Suma kwadratów reszt | 1,728486 |  | Błąd standardowy reszt | 0,759053 |
| Wsp. determ. R-kwadrat | 0,769715 |  | Skorygowany R-kwadrat | 0,692953 |
| F(1, 3) | 10,02734 |  | Wartość p dla testu F | 0,050613 |
| Logarytm wiarygodności | −4,439213 |  | Kryt. inform. Akaike'a | 12,87843 |
| Kryt. bayes. Schwarza | 12,09730 |  | Kryt. Hannana-Quinna | 10,78197 |

Współczynniki modelu logitowego obliczane są z następującego równania macierzy:

Gdzie:

X – Macierz zmiennych objaśniających (W tym przypadku macierz Lat)

L – Macierz logitów/logitowa

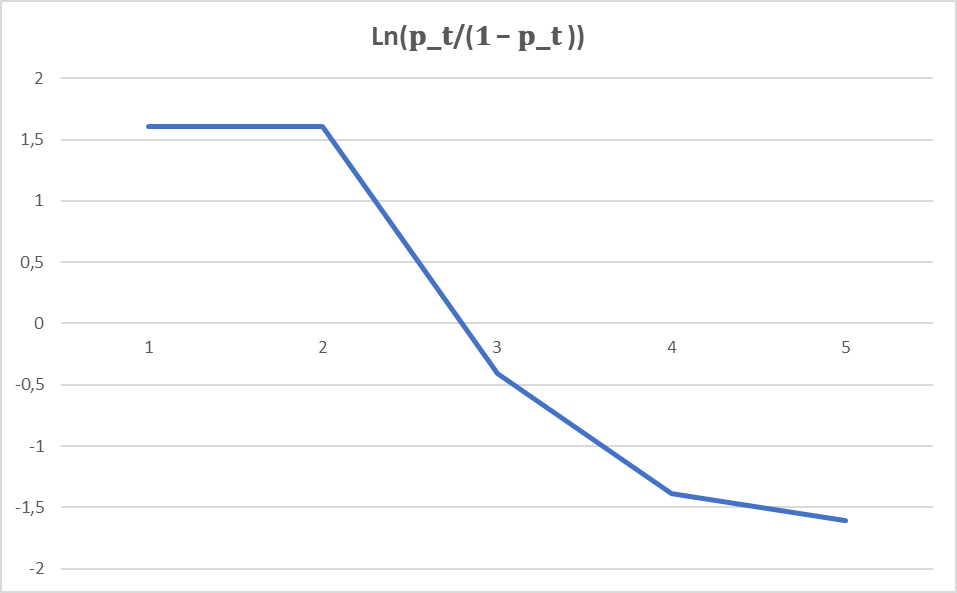
Korzystając z powyższego wzoru udało nam się obliczyć współczynniki .

2,06055

−0,760090

Następnie możemy oszacować prawdopodobieństwo p.

p =



Rysunek 6 - wykres zależności logitów od roku

Wnioski:

Zarówno wykres zależności prawdopodobieństwa w stanie wolnym od roku studiów jak i wykres zależności logitów od roku mają tendencje spadkową oraz przyjmują zbliżony kształt. Z obu wykresów wynika, że z biegiem lat maleje ilość osób w stanie wolnym.

* Zadanie 2

Przez cały okres eksploatacji pewnego systemu operacyjnego (OS) zbierano dane dotyczące liczby błędów krytycznych wykrytych w tym czasie w oprogramowaniu. Zebrane obserwacje w układzie miesięcznym przedstawionow tabeli 3.

Dane:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Nr miesiąca | Liczba błędów | Nr miesiąca | Liczba błędów | Nr miesiąca | Liczba błędów | Nr miesiąca | Liczba błędów | Nr miesiąca | Liczba błędów |
| 1 | 1 | 21 | 9 | 41 | 3 | 61 | 2 | 81 | 8 |
| 2 | 0 | 22 | 0 | 42 | 2 | 62 | 19 | 82 | 0 |
| 3 | 0 | 23 | 0 | 43 | 1 | 63 | 7 | 83 | 6 |
| 4 | 0 | 24 | 3 | 44 | 6 | 64 | 2 | 84 | 5 |
| 5 | 0 | 25 | 1 | 45 | 3 | 65 | 5 | 85 | 10 |
| 6 | 0 | 26 | 1 | 46 | 0 | 66 | 1 | 86 | 0 |
| 7 | 0 | 27 | 0 | 47 | 1 | 67 | 12 | 87 | 5 |
| 8 | 0 | 28 | 1 | 48 | 1 | 68 | 4 | 88 | 1 |
| 9 | 0 | 29 | 0 | 49 | 0 | 69 | 6 | 89 | 2 |
| 10 | 0 | 30 | 2 | 50 | 0 | 70 | 4 | 90 | 1 |
| 11 | 0 | 31 | 10 | 51 | 14 | 71 | 7 | 91 | 2 |
| 12 | 0 | 32 | 0 | 52 | 1 | 72 | 2 | 92 | 1 |
| 13 | 0 | 33 | 16 | 53 | 4 | 73 | 2 | 93 | 0 |
| 14 | 0 | 34 | 0 | 54 | 1 | 74 | 3 |  |  |
| 15 | 0 | 35 | 2 | 55 | 1 | 75 | 8 |  |  |
| 16 | 0 | 36 | 2 | 56 | 7 | 76 | 4 |  |  |
| 17 | 0 | 37 | 1 | 57 | 14 | 77 | 6 |  |  |
| 18 | 0 | 38 | 1 | 58 | 6 | 78 | 3 |  |  |
| 19 | 0 | 39 | 1 | 59 | 0 | 79 | 3 |  |  |
| 20 | 1 | 40 | 0 | 60 | 1 | 80 | 5 |  |  |

W pierwszym kroku należy wyznaczyć drugą wartość liczby błędów jako sumę wszystkich błędów, jakie zadziały się od początku do danego miesiąca.

Następnie należy wyznaczyć funkcję logistyczną na podstawie przykładowych wartości afla, beta oraz gamma (np. 1, 1, 1):

Dodatkowo należy wyznaczyć wartość błędu dla każdej obliczonej wartości funkcji logistycznej, jako wartość , a następnie policzyć sumę.

W następnym kroku w Excelu należy wykorzystać solver wstawiając jako funkcję celu obliczoną sumę błędów, jako pole komórek zmiennych miejsca na wartości zmiennych , a metodę rozwiązywania „Nieliniową GRG”. Dla poprawienia dokładności wyniku należy zmienić dokładność na wartość 0,0001.

Zadaniem solvera jest minimalizacja funkcji sumy błędów funkcji logistycznej, w wyniku czego otrzymujemy zoptymalizowane parametry funkcji logistycznej. W ramach swojego działania dokonuje kilka iteracji, aby jak najbardziej zbliżyć się do wartości minimalnej.

Obliczone wartości parametrów:

|  |  |
| --- | --- |
| alfa | 310,394 |
| beta | 91,33092 |
| gamma | 0,068679 |
| suma | 3345,74 |

Zbieżność funkcji logistycznej z danymi można zobaczyć na poniższym wykresie:

Wnioski:  
Widząc po wykresie, otrzymana funkcja logistyczna bardzo dobrze pokrywa się z danymi wejściowymi. Wskazuje to na poprawne dobranie wartości parametrów .

Analizując kształt funkcji można stwierdzić, że błędy krytyczne systemu operacyjnego wystąpiły w niewielkiej ilości na początku działania, a z każdym kolejnym miesiącem pojawiało się ich coraz więcej (funkcja szybciej zwiększała swoją wartość). Przez ostatnie zmierzone liczba błędów nieznacznie spowolniła.

Laboratoria 3

* Zadanie 1

Mieszkańcy pewnej małej kamienicy, mieszczącej się zaraz tuż przy plaży morskiej, zaczęli od paru dni z ciekawością obserwować treningi grupy zawodników siatkówki plażowej. Mieszkańcy zauważyli, że ta grupa graczy niekiedy dzieli się na dwa zespoły w celu rozegrania meczu. Niestety obserwatorzy nie znają zamiarów trenujących, ale odnotowali, że gracze grają w różnych warunkach pogodowych. Swoje obserwacje odnotowali w tabeli.

**Dane:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Nr obserwacji | Siła wiatru | Zachmurzenie | Odczuwalna temperatura | Zagrano mecz |
| 1 | Silny | Pochmurnie | Zimno | Nie |
| 2 | Silny | Pochmurnie | Ciepło | Nie |
| 3 | Brak | Słonecznie | Ciepło | Tak |
| 4 | Brak | Słonecznie | Gorąco | Nie |
| 5 | Słaby | Pochmurnie | Gorąco | Tak |
| 6 | Słaby | Słonecznie | Ciepło | Tak |
| 7 | Brak | Pochmurnie | Zimno | Nie |
| 8 | Silny | Słonecznie | Zimno | Tak |
| 9 | Brak | Pochmurnie | Gorąco | Tak |
| 10 | Silny | Pochmurnie | Ciepło | Tak |
| 11 | Silny | Słonecznie | Ciepło | ??? |

Dzisiaj jest ciepło i słonecznie, ale wieje silny wiatr. Wobec tego mieszkańcy kibicujący grze zastanawiają się czy gracze pojawią się na plaży, aby rozegrać swój mecz treningowy, czy tylko wykonać ćwiczenia.

Twierdzenie Bayes’a można przedstawić wzorem prawdopodobieństwa warunkowego a posteriori, że obserwacja o wartości zaobserwowanej X pochodzi z klasy C:

Gdzie:

prawdopodobieństwo bezwarunkowe wystąpienia klasy C

prawdopodobieństwo warunkowe że obserwacja o wartości X należy do klasy C

prawdopodobieństwo bezwarunkowe wystąpienia obserwacji X

Jako, że prawdopodobieństwo Pr{X} możemy traktować jako dane stałe to dla każdej klasy mamy proporcjonalność:

Oznacza to, że poszukiwaną klasą spośród , i = 1,2,….,m, jest ta klasa, dla której wartość jest największa.

W skrócie można napisać:

I jest to nazywane klasyfikacją maksymalnego a posteriori. Niestety wartość jest praktycznie niemożliwa do obliczenia dlatego należy naiwnie założyć, że:

Wobec tego naiwny klasyfikator Bayes’a wyraża się następującym wzorem:

Dodatkowo możemy wykorzystać metodę najbliższego sąsiedztwa. W tej metodzie wykorzystuje się miarę odległości między badanymi obiektami oraz liczbę k najbliższych sąsiadów. Miarę odległości można przedstawić następującymi wzorami:

Odległość euklidesowa:

Odległość miejska:

Gdzie:

x, y – to punkty w n-wymiarowej przestrzeni

Przedstawiony wzór jest wzorem na odległość euklidesową(można też wykorzystać inne wzory na odległość takie jak miejską czy Czebyszewa). W następnym kroku określamy parametr k, który określa liczbę obiektów ze zbioru {x} determinujących decyzję o przynależności do konkretnej klasy. Po wybraniu tego parametru wybieramy k najbliższych sąsiadów. Później porównujemy nowy obiekt y z k najbliższymi sąsiadami i wybieramy przynależność badanego obiektu do konkretnej klasy poprzez głosowanie proste. Głosowanie proste polega na przypisaniu jednego głosu jednemu obiektowi. W sytuacji remisu eksperymentator podejmuje decyzję.

**Rozwiązanie:**

* Z wykorzystaniem naiwnego klasyfikatora Bayes’a

Pr{C1} = Pr{d=tak} = 0,6

Pr{C2} = Pr{d=nie} = 0,4

Pr{x1=silny|d=tak} = 0,333333333

Pr{x1=silny|d=nie} = 0,5

Pr{x2=słonecznie|d=tak} = 0,5

Pr{x2=słonecznie|d=nie} = 0,25

Pr{x3=ciepło|d=tak} = 0,5

Pr{x3=ciepło|d=nie} = 0,25

= 0,05

= 0,0125

Szukana przez nas wartość jest równa „Tak”, ponieważ wartość iloczynu dla C1 jest większa niż dla C2

* Z wykorzystaniem metody najbliższego sąsiedztwa

Wyniki obliczeń można przedstawić w postaci tabeli

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Nr obserwacji | Siła wiatru | Zachmurzenie | Odczuwalna temperatura | Zagrano mecz |  |  |
| 1 | 2 | 1 | 0 | Nie | 1,41 | 2 |
| 2 | 2 | 1 | 1 | Nie | 1 | 1 |
| 3 | 0 | 0 | 1 | Tak | 2 | 2 |
| 4 | 0 | 0 | 2 | Nie | 2,24 | 3 |
| 5 | 1 | 1 | 2 | Tak | 1,73 | 3 |
| 6 | 1 | 0 | 1 | Tak | 1 | 1 |
| 7 | 0 | 1 | 0 | Nie | 2,45 | 4 |
| 8 | 2 | 0 | 0 | Tak | 1 | 1 |
| 9 | 0 | 1 | 2 | Tak | 2,45 | 4 |
| 10 | 2 | 1 | 1 | Tak | 1 | 1 |
| 11 | 2 | 0 | 1 | ??? |  |  |

Jako, że metodę najbliższego sąsiedztwa można użyć jedynie na wartościach liczbowych wartości w postaci słów zamienione zostały na liczby.

Dla odległości euklidesowej wyniki dla poszczególnych k wykładają następująco:

* k=(od 1 do 4): Tak
* k=5: Tak
* k=6: Tak
* k=7: Tak
* k=8: Tak
* k=(od 9 do 10): Tak

Widać że dla dowolnego k zawsze zagrano mecz.

Dla odległości miejskiej wyniki dla poszczególnych k wykładają następująco:

* k=(od 1 do 4): Tak
* k=(od 5 do 6): Tak
* k=(od 7 do 8): Tak
* k=(od 9 do 10): Tak

Wnioski:

Dla obu metod wynik wychodzi ten sam, co oznacza że dla podanych danych mecz zostanie rozegrany. Warto zauważyć, że zarówno dla odległości euklidesowej jak i miejskiej ostateczne wyniki są takie same.

* Zadanie 2

Załóżmy, że chcemy zbudować algorytm odfiltrowania spamu z otrzymanych wiadomości pocztą lub za pomocą czatu. W spamie występują treści wg. określonych słów kluczowych, których statystyka występowania została zawarta w tabeli wraz z dokonaną klasyfikacją wg. pewnego algorytmu.

**Dane:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Nr wiad. | Słowa kluczowe | | | | | Klasyfikacja: spam |
| Pieniądz | Darmowy | Bogaty | Nieprzyzwoicie | Tajny |
| 1 | Nie | Nie | Tak | Nie | Tak | Tak |
| 2 | Tak | Tak | Tak | Nie | Nie | Tak |
| 3 | Nie | Nie | Nie | Nie | Nie | Nie |
| 4 | Nie | Tak | Nie | Nie | Nie | Tak |
| 5 | Tak | Nie | Nie | Nie | Nie | Nie |
| 6 | Nie | Tak | Nie | Tak | Tak | Tak |
| 7 | Nie | Tak | Nie | Tak | Nie | Tak |
| 8 | Nie | Nie | Nie | Tak | Nie | Tak |
| 9 | Nie | Tak | Nie | Nie | Nie | Nie |
| 10 | Nie | Nie | Nie | Nie | Tak | Nie |
| 11 | Tak | Tak | Tak | Nie | Tak | Tak |
| 12 | Tak | Nie | Nie | Nie | Tak | Tak |
| 13 | Nie | Tak | Tak | Nie | Nie | Nie |
| 14 | Tak | Nie | Tak | Nie | Tak | ??? |

Analogicznie do zadania 1 wykorzystujemy naiwną klasyfikację Bayes’a, która dokładnie jest opisana wyżej.

Pr{C1} = Pr{d=tak} = 0,615

Pr{C2} = Pr{d=nie} = 0,385

Pr{x1=tak|d=tak} = 0,375

Pr{x1=tak|d=nie} = 0,2

Pr{x2=nie|d=tak} = 0, 375

Pr{x2=nie|d=nie} = 0,6

Pr{x3=tak|d=tak} = 0, 375

Pr{x3=tak|d=nie} = 0,2

Pr{x4=nie|d=tak} = 0,625

Pr{x4=nie|d=nie} = 1

Pr{x5=tak|d=tak} = 0,5

Pr{x5=tak|d=nie} = 0,2

= 0,0101

= 0,0018

Szukana przez nas wartość jest równa „Tak”, ponieważ wartość iloczynu dla C1 jest większa niż dla C2

Wnioski:

Dla podanych danych słów kluczowych według klasyfikacji bayesowskiej wiadomość nr 14 zostanie sklasyfikowana jako spam.