**Pokok Bahasan VI**

**Naïve Bayes**

**Kode Pokok Bahasan**: TIK.RPL03.005.00.01

**Deskripsi Pokok Bahasan**:

Membahas bagaimana penerapan Algoritma Naïve Bayes pada dataset titanic.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| No | Elemen Kompetensi | Indikator Kinerja | Jml Jam | Hal |
| 1. | Menampilkan peluang dari kasus yang diberikan. | Mampu melakukan analisis terhadap peluang atas kejadian yang ditentukan. | 1 | 12 |
| 2. | Melakukan perhitungan manual menggunakan excel. | Mampu melakukan perhitungan peluang secara manual menggunakan excel pada kondisi yang ditentukan | 2 | 15 |
| 3. | Menggunakan fungsi naiveBayes dalam memprediksi data | Mengimplementasikan fungsi naive bayes pada prediksi data |  |  |

**TUGAS PENDAHULUAN**

Hal yang harus dilakukan dan acuan yang harus dibaca sebelum praktikum :

1. Menginstal R pada PC masing-masing praktikan.

2. Menginstal R Studio pada PC masing-masing praktikan.

**DAFTAR PERTANYAAN**

1. Apa itu algoritma Naïve Bayes?

2. Apa kegunaan Naïve Bayes?

3. Sebutkan tahapan dari proses algoritma Naïve Bayes!

**TEORI SINGKAT**

Algoritma Naive Bayes merupakan sebuah metoda klasifikasi menggunakan metode probabilitas dan statistik yg dikemukakan oleh ilmuwan Inggris Thomas Bayes. Algoritma Naive Bayes memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya sehingga dikenal sebagai Teorema Bayes. Ciri utama dr Naïve Bayes Classifier ini adalah asumsi yg sangat kuat (naïf) akan independensi dari masing-masing kondisi / kejadian.

Naive Bayes Classifier bekerja sangat baik dibanding dengan model classifier lainnya. Hal ini dibuktikan pada jurnal Xhemali, Daniela, Chris J. Hinde, and Roger G. Stone. “Naive Bayes vs. decision trees vs. neural networks in the classification of training web pages.” (2009), mengatakan bahwa “Naïve Bayes Classifier memiliki tingkat akurasi yg lebih baik dibanding model classifier lainnya”.

Keuntungan penggunan adalah bahwa metoda ini hanya membutuhkan jumlah data pelatihan (training data) yang kecil untuk menentukan estimasi parameter yg diperlukan dalam proses pengklasifikasian. Karena yg diasumsikan sebagai variabel independent, maka hanya varians dari suatu variabel dalam sebuah kelas yang dibutuhkan untuk menentukan klasifikasi, bukan keseluruhan dari matriks kovarians.

**LAB SETUP**

Hal yang harus disiapkan dan dilakukan oleh praktikan untuk menjalankan praktikum modul ini.

1. Menginstall library yang dibutuhkan untuk mengerjakan modul.

2. Menjalankan R Studio.

**ELEMEN KOMPETENSI I**

**Deskripsi:**

Menampilkan peluang dari kasus yang diberikan.

**Kompetensi Dasar**:

Mampu melakukan analisis terhadap peluang atas kejadian yang ditentukan.

**Latihan**

**Penjelasan Singkat :**

Pada latihan ini anda akan diminta untuk menampilkan summarize dari data menggunakan library yang disediakan oleh R.

**Langkah-Langkah Praktikum:**

1.Instal dan panggil package berikut jika belum terinstall

install.packages("tidyverse")

library("tidyverse") #for data wrangling tools

install.packages("titanic")

library("titanic")

2. Inisialisasi titanic\_train ke dalam variabel tdf.

tdf <- titanic\_train #training set of Titanic data

head(tdf)

3. Lakukan perintah berikut :

Compute the probability that a randomly selected passenger on the Titanic was female given that the passenger was at least 35 years old.

tdf %>%

summarize(prob = sum(Age >= 35 & Sex == "female", na.rm = TRUE)/sum(Age >= 35, na.rm = TRUE))

Output :

|  |
| --- |
|  |

**ELEMEN KOMPETENSI II**

**Deskripsi:**

Melakukan perhitungan manual menggunakan excel.

**Kompetensi Dasar**:

Mampu melakukan perhitungan peluang secara manual menggunakan excel pada kondisi yang ditentukan

**Latihan 1.2.1**

**Penjelasan Singkat :**

Pada latihan ini anda akan diminta untuk melakukan perhitungan peluang secara manual menggunakan excel.

**Langkah-Langkah Praktikum:**

Lakukan perhitungan kasus di atas dengan menggunakan Excel. Pisahkan age dengan batas 35 tahun.

Output :

|  |
| --- |
|  |

**ELEMEN KOMPETENSI III**

**Deskripsi:**

Menggunakan fungsi naïve bayes dalam memprediksi data.

**Kompetensi Dasar**:

Mengimplementasikan fungsi naive bayes pada prediksi data

**Latihan 1.3.1**

**Penjelasan Singkat :**

Pada latihan ini anda akan diminta untuk mengimplementasikan naïve bayes pada kasus yang diberikan.

**Langkah-Langkah Praktikum:**

Gunakan [titanic.csv](http://web.stanford.edu/class/archive/cs/cs109/cs109.1166/stuff/titanic.csv) yang berisi data 887 penumpang Titanic passengers. Kolom data menggambarkan survived (*S*), age (*A*), passenger-class (*C*), sex (*G*) and the fare paid (*X*).

Hitung peluang bersyarat (conditional probability) di bawah ini

*P*(*S*= true | *G*=female)

*P*(*S*= true | *G*=male)

*P*(*S*= true | *C*=1)

*P*(*S*= true | *C*=2)

*P*(*S*= true | *C*=3)

*P*(*S*= true | *G*=female,*C*=1) =

*P*(*S*= true | *G*=female,*C*=2) =  
*P*(*S*= true | *G*=female,*C*=3) =  
*P*(*S*= true | *G*=male,*C*=1) =  
*P*(*S*= true | *G*=male,*C*=2) =  
*P*(*S*= true | *G*=male,*C*=3) =

Gunakan R dan excel :

|  |
| --- |
|  |

Jalankan perintah R di bawah ini :

|  |
| --- |
| # https://www.kaggle.com/brirush/naive-bayes-for-titanic  library(e1071)  train <- read.csv("F:/Kuliah Data Mining gasal 1819/Kaggle/Titanic/train.csv")  test <- read.csv("F:/Kuliah Data Mining gasal 1819/Kaggle/Titanic/test.csv")  BayesTitanicModel<-naiveBayes(as.factor(Survived)~., train)  BayesPrediction<-predict(BayesTitanicModel, test)  summary(BayesPrediction)  output<-data.frame(test$PassengerId, BayesPrediction)  str(output)  colnames(output)<-cbind("PassengerId","Survived")  write.csv(output, file = 'Rushton\_Solution.csv', row.names = F) |

Output :

|  |
| --- |
|  |

Berikan penjelasan terhadap output di atas

|  |
| --- |
|  |

Tugas : Kasus “playing golf”

Data excelNaive

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| id | Outlook | Temp | Humadity | Wndy | PlayGolf |
| 1 | Rainy | Hot | High | FALSE | No |
| 2 | Rainy | Hot | High | TRUE | No |
| 3 | Overcast | Hot | High | FALSE | Yes |
| 4 | Sunny | Mild | High | FALSE | Yes |
| 5 | Sunny | Cool | Normal | FALSE | Yes |
| 6 | Sunny | Cool | Normal | TRUE | No |
| 7 | Overcast | Cool | Normal | TRUE | Yes |
| 8 | Rainy | Mild | High | FALSE | No |
| 9 | Rainy | Cool | Normal | FALSE | Yes |
| 10 | Sunny | Mild | Normal | FALSE | Yes |
| 11 | Rainy | Mild | Normal | TRUE | Yes |
| 12 | Overcast | Mild | High | TRUE | Yes |
| 13 | Overcast | Hot | Normal | FALSE | Yes |
| 14 | Sunny | Mild | High | TRUE | No |

Data excelNaiveTest

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| id | Outlook | Temp | Humadity | Wndy |
| 16 | Rainy | Mild | Normal | TRUE |

Script R :

|  |
| --- |
| > excelNaive <- read.delim('clipboard')  > excelNaiveTest <- read.delim('clipboard')  > excelNaiveModel <- naiveBayes(as.factor(PlayGolf)~., excelNaive)  > excelNaivePredic <- predict(excelNaiveModel, excelNaiveTest)  > summary(excelNaivePredic)  > excelNaiveOutput <- data.frame(excelNaiveTest$id, excelNaivePredic)  > str(excelNaiveOutput)  > colnames(excelNaiveOutput) <- cbind('id', 'PlayGolf')  > write.csv(excelNaiveOutput, file = "rainy.csv", row.names = F) |

Output :

|  |
| --- |
|  |

Hitungan manual data test :

|  |
| --- |
|  |

Sumber :

<http://web.stanford.edu/class/archive/cs/cs109/cs109.1166/problem12.html>

<http://rstudio-pubs-static.s3.amazonaws.com/6595_b57093a21dfc46a4b3338cfee29ec61e.html>

<https://community.rstudio.com/t/conditional-probability-with-dplyr/5117>

<https://www.kaggle.com/brirush/naive-bayes-for-titanic>

**CEK LIST**

|  |  |  |  |
| --- | --- | --- | --- |
| Elemen Kompetensi | No Latihan | Penyelesaian | |
| Selesai | Tidak selesai |
| 1 | 1.1.1 |  |  |
| 2 | 1.2.1 |  |  |
| 3 | 1.3.1 |  |  |

**FORM UMPAN BALIK**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Elemen Kompetensi** | **Tingkat Kesulitan** | | | **Tingkat Ketertarikan** | | | **Waktu Penyelesaian dalam menit** |
| Memahami data pre-processing. |  |  |  |  |  |  |  |
|  |  | Sangat Mudah |  |  | Tidak Tertarik |  |
|  |  |  |  |  |  |  |
|  |  | Mudah |  |  | Cukup Tertarik |  |
|  |  |  |  |  |  |  |
|  |  | Biasa |  |  | Tertarik |  |
|  |  |  |  |  |  |  |
|  |  | Sulit |  |  | Sangat Tertarik |  |
|  |  |  |  |  |  |  |
|  |  | Sangat Sulit |  |  |  |  |
|  |  |  |  |  |  |  |
| Mengimplementasikan pre-processing data. |  |  |  |  |  |  |  |
|  |  | Sangat Mudah |  |  | Tidak Tertarik |  |
|  |  |  |  |  |  |  |
|  |  | Mudah |  |  | Cukup Tertarik |  |
|  |  |  |  |  |  |  |
|  |  | Biasa |  |  | Tertarik |  |
|  |  |  |  |  |  |  |
|  |  | Sulit |  |  | Sangat Tertarik |  |
|  |  |  |  |  |  |  |
|  |  | Sangat Sulit |  |  |  |  |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Elemen Kompetensi** | **Tingkat Kesulitan** | | | **Tingkat Ketertarikan** | | | **Waktu Penyelesaian dalam menit** |
| Menggunakan fungsi naiveBayes dalam memprediksi data |  |  |  |  |  |  |  |
|  |  | Sangat Mudah |  |  | Tidak Tertarik |  |
|  |  |  |  |  |  |  |
|  |  | Mudah |  |  | Cukup Tertarik |  |
|  |  |  |  |  |  |  |
|  |  | Biasa |  |  | Tertarik |  |
|  |  |  |  |  |  |  |
|  |  | Sulit |  |  | Sangat Tertarik |  |
|  |  |  |  |  |  |  |
|  |  | Sangat Sulit |  |  |  |  |
|  |  |  |  |  |  |  |