A blue and black logo

Description automatically generatedAprendizagem 2024/2025

**Homework II – Group 034**

(ist199417, ist1106481)

**I. Pen-and-paper**

**1)**

Distância de Hamming: número de features diferentes entre e .

Temos então a tabela de distâncias

j

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| i | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 1 | 0 | 2 | 1 | 0 | 1 | 1 | 1 | 2 |
| 2 | 2 | 0 | 1 | 2 | 1 | 1 | 1 | 0 |
| 3 | 1 | 1 | 0 | 1 | 2 | 2 | 0 | 1 |
| 4 | 0 | 2 | 1 | 0 | 1 | 1 | 1 | 2 |
| 5 | 1 | 1 | 2 | 1 | 0 | 0 | 2 | 1 |
| 6 | 1 | 1 | 2 | 1 | 0 | 0 | 2 | 1 |
| 7 | 1 | 1 | 0 | 1 | 2 | 2 | 0 | 1 |
| 8 | 2 | 0 | 1 | 2 | 1 | 1 | 1 | 0 |

- Para

A k-vizinhança é .

- Para - Para

Logo, deve ser classificado como N.

- Para

A k-vizinhança é .

- Para - Para

Logo, deve ser classificado como N.

- Para

A k-vizinhança é .

- Para - Para

Logo, deve ser classificado como P.

- Para

A k-vizinhança é .

- Para - Para

Logo, deve ser classificado como N.

- Para

A k-vizinhança é .

- Para - Para

Logo, deve ser classificado como P.

- Para

A k-vizinhança é .

- Para - Para

Logo, deve ser classificado como P.

- Para

A k-vizinhança é .

- Para - Para

Logo, deve ser classificado como P.

- Para

A k-vizinhança é .

- Para - Para

Logo, deve ser classificado como N.

Condensando os resultados obtidos, temos:

A white rectangular box with black text

Description automatically generated

|  |  |  |
| --- | --- | --- |
|  | Pred | True |
| 1 | N | P |
| 2 | N | P |
| 3 | P | P |
| 4 | N | P |
| 5 | P | N |
| 6 | P | N |
| 7 | P | N |
| 8 | N | N |

**2)**

Propomos a seguinte métrica:

Seja .

d e

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| j  i | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 1 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 1 |
| 2 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 0 |
| 3 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 1 |
| 4 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 1 |
| 5 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 0 |
| 6 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 0 |
| 7 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 1 |
| 8 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 0 |

- Para

A k-vizinhança é .

- Para - Para

Logo, deve ser classificado como P.

- Para

A k-vizinhança é .

- Para - Para

Logo, deve ser classificado como N.

- Para

A k-vizinhança é .

- Para - Para

Logo, deve ser classificado como P.

- Para

A k-vizinhança é .

- Para - Para

Logo, deve ser classificado como P.

- Para

A k-vizinhança é .

- Para - Para

Logo, deve ser classificado como N.

- Para

A k-vizinhança é .

- Para - Para

Logo, deve ser classificado como N.

- Para

A k-vizinhança é .

- Para - Para

Logo, deve ser classificado como P.

- Para

A k-vizinhança é .

- Para - Para

Logo, deve ser classificado como N.

Condensando os resultados obtidos, temos:

A white squares with black text

Description automatically generated

|  |  |  |
| --- | --- | --- |
|  | Pred | True |
| 1 | P | P |
| 2 | N | P |
| 3 | P | P |
| 4 | P | P |
| 5 | N | N |
| 6 | N | N |
| 7 | P | N |
| 8 | N | N |

Portanto, a medida F1 aumentou em 3 vezes.

**3)**

Para aprender um Bayesian classifier, precisamos de:

- prior probabilities de cada classe (menos uma);

- os parâmetros associados com .

Comecemos por calcular as prior probabilities.

Agora precisamos dos parâmetros associados a .

Como e são dependentes, temos de calcular a sua probabilidade conjunta.

A white grid with black numbers and a square in the middle

Description automatically generatedA table with numbers and a triangle

Description automatically generated with medium confidence Para Para

Como {} e {} são conjuntos de variáveis independentes,

() = ()

Logo , precisamos saber . Como é normal , precisamos de calcular a média e a variância para saber a função de densidade de probabilidade.

Para

A função de densidade de probabilidade está bem definida.

Para

A função de densidade de probabilidade está bem definida.

(Usámos variância corrigida por se tratar de uma amostra)

**4)**

Para

Para

Para

Logo, deverá ser classificado como P.

Para

Para

Para

Logo, deverá ser classificado como N.

Para

Para

Para

Logo, deverá ser classificado como N.

**5)**

- Vocabulário = {“Amazing”, “run”, “I”, “like”, “it”, “too”, “tired”, “Bad”}

- = |Vocabulário| = 8

-

-

Para

Para

Logo,deverá ser classificado como P.

**II. Programming and critical analysis**

**1)**

A diagram of a graph

Description automatically generated with medium confidence**a.**

kNN Accuracy: ;

GaussianNB Accuracy:;

kNN Mean Accuracy: ;

kNN Std Accuracy: ;

GaussianNB Mean Accuracy: ;

GaussianNB Std Accuracy: .

Como vemos pelo gráfico e pelo desvio padrão obtidos, o GaussianNB é o mais estável relativamente a performance. Esta observação está de acordo com o esperado, porque o GaussianNB é um modelo probabilístico (paramétrico), que assume independência entre as variáveis, ajudando na generalização do modelo. O kNN é baseado na amostra obtida (não paramétrico), sujeito a maior variabilidade, estando mais dependente dos dados de treino (pode dar overfit a cada fold).

**A diagram of a graph

Description automatically generated with medium confidenceb.**

kNN Accuracy (Min-Max Scaled):

GaussianNB Accuracy (Min-Max Scaled):

kNN Mean Accuracy (Min-Max Scaled):

kNN Std Accuracy (Min-Max Scaled):

GaussianNB Mean Accuracy (Min-Max Scaled):

GaussianNB Std Accuracy (Min-Max Scaled):

Como vemos pelo gráfico e resultados em cima apresentados, o modelo kNN melhorou bastante ao fazer o pré-processamento. Esta observação está de acordo com o esperado, uma vez que o kNN é um modelo baseado em distâncias (entre duas instâncias). Como cada variável à partida tem uma escala diferente, variáveis com escalas maiores têm maior impacto na distância. Ao fazer o Min-Max scaling todas as variáveis têm a mesma escala 0-1, tendo assim o mesmo impacto na distância.

Os resultados do modelo GaussianNB permaneceram iguais, porque este modelo não é baseado em distâncias, assim é irrelevante a escala de cada variável.

**c.**

kNN < GaussianNB? pval = 0.2537311948784664

Optámos por não usar os dados sem estarem escalados, com o Min-Max, porque o modelo GaussianNB era claramente superior (sem escalamento) e acreditamos que o mais justo seria comparar os dois modelos na sua melhor forma. Portanto, com o escalamento Min-Max, não conseguimos concluir que o modelo kNN é estatísticamente superior ao modelo GaussianNB, de acordo com a accuracy, uma vez que o pval > 0.05 (nível de confiança = 0.95).

Teríamos de diminuir o nível de confiança para 0.74 para rejeitar a hipótese nula (kNN < GaussianNB), ou seja, afirmar que o modelo kNN é estatísticamente superior ao modelo GaussianNB.

**2)**

**a.**

**A graph of a number of numbers

Description automatically generated with medium confidence**

**b.**

Observando o gráfico (dados não escalados), vemos que aumentar o número de vizinhos leva a um aumento da accuracy do classificador kNN do grupo de teste, para uniform weights e distance weights. Isto era de esperar, porque não nos estamos a limitar a analisar a vizinhança próxima, mas sim a um maior número de vizinhos, o que melhora o poder de generalização do modelo.

**3)**

**A screenshot of a computer

Description automatically generated**Heat map da correlação de Pearson entre variáveis:

**A screenshot of a computer

Description automatically generated**

Um dos problemas do GaussianNB, para este dataset, é assumir independência entre variáveis. Como vemos no heat map acima, há vários pares de variáveis, cuja correlação é significativa. Por exemplo, , o que sugere um nível significativo de correlação (caso fossem independentes a correlação seria zero). Por isso, estamos a ignorar relações importantes entre variáveis, diminuindo a accuracy do modelo.

Gráficos da distribuição das variáveis age e sex:

A graph with blue lines and a line

Description automatically generatedA graph of a distribution of age

Description automatically generated

Outro problema é assumir distribuição Gaussiana de todas as variáveis. Como vemos acima, a variável age parece ter uma distribuição similar à Gaussiana, no entanto, em variáveis categóricas, como o sex, a distribuição normal não se adequa. A distribuição Gaussiana é mais adequada para variáveis contínuas, neste caso temos dados mistos (categóricos e contínuos).