



## Relatório - Trabalho Prático

Universidade de Évora

Aprendizagem Automática 2022/2023

Bernardo Vitorino (148463), Daniel Barreiros (148452) e Tomás Antunes (148511)

### I – Introdução

O relatório que se segue é alusivo ao trabalho prático da disciplina de aprendizagem automática, lecionado pelo docente Luís Rato e tem como objetivo explicar a abordagem definida por nós relativamente às estruturas de dados escolhidas para a implementação de alguns métodos.

Iremos também colocar algumas análises de desempenho referentes aos diversos tipos de dados e uma opinião crítica relativa à realização deste trabalho.

De uma forma geral, conseguimos calcular diversas probabilidades através de fórmulas estudadas na disciplina. Entre elas utilizamos o estimador suavizado e o teorema de Bayes.

### II – Desenvolvimento

Na realização deste trabalho começámos por criar uma classe chamada “NaiveBayesUevora” que implementa o classificador Naive Bayes. No seu conteúdo definimos o estimador suavizado, o método fit, predict, accuracy\_score e precision\_score.

Método fit: Calculamos as probabilidades à priori das classes, de seguida calculamos a probabilidade dos atributos dada a classe e por último calculamos a probabilidade dos atributos. Todos estes valores são guardados em dicionários.

Método predict: Vai prever a classe de um conjunto de atributos calculando a probabilidade de pertencerem, por exemplo, à classe A ou B, e de seguida escolhe a que tem maior probabilidade.

As classes resultantes são guardadas numa lista (results).

Método accuracy\_score: Calcula a exatidão de um classificador comparando os resultados previstos com os dados fornecidos. O resultado é devolvido através de um float.

Método precision\_score: Calcula a precisão de um classificador sendo este a média aritmética da precisão de cada classe. De igual forma, o resultado é devolvido através de um float.

### III – Considerações

Uma das dificuldades com a qual nos confrontámos foi o facto de existirem nos dados de teste valores que não existiam nos dados de treino. Através da criação de uma função (`check_attribute`), verificamos se a probabilidade de certo atributo já foi calculada e no caso de não ter sido é calculado o valor da mesma. Através desta correção, conseguimos eliminar o erro e assim prever a que classe corresponde o conjunto de dados.

Outra das dificuldades encontradas por nós incidiu sobre a realização da tarefa de split ao ficheiro `weather-nominal` (ficheiro que contém um conjunto de dados). O problema ocorreu visto que a função (`train_test_split` da biblioteca `sklearn`) não coloca de uma forma linear os índices correspondentes a cada linha, ou seja, a primeira linha poderia ter índice p.e 20 e a segunda índice 1.

A solução encontrada consistiu em aceder à lista de dados mediante a ordem da lista de chaves (utilizando o método `.keys()` dado que este retorna a lista das chaves de um dicionário). Deste modo foi proporcionada uma abstração da maneira como o método split tinha reorganizado os dados, podendo assim aceder a cada posição de dados mediante um ciclo.

### IV - Análises de Desempenho

#### Breast-Cancer1

- Para o primeiro conjunto de dados utilizando um  $\alpha=0$  obtivemos uma exatidão de cerca de 86% e uma precisão de cerca de 88%.
- Utilizando um  $\alpha=1$  obtivemos uma exatidão de cerca de 81% e uma precisão de cerca de 86%.
- Utilizando um  $\alpha=5$  obtivemos uma exatidão de cerca de 76% e uma precisão de cerca de 83%.

Os resultados mantiveram-se uniformes independentemente do valor do  $\alpha$ , apesar de que à medida que o  $\alpha$  aumenta a exatidão e a precisão diminuem.

#### Breast-Cancer2

- Para o segundo conjunto de dados utilizando um  $\alpha=0$  obtivemos uma exatidão de cerca de 86% e uma precisão de cerca de 88% .
- Utilizando um  $\alpha=1$  obtivemos uma exatidão de cerca de 81% e uma precisão de cerca de 86%.
- Utilizando um  $\alpha=5$  obtivemos uma exatidão de cerca de 76% e uma precisão de cerca de 83%.

Os resultados mantiveram-se uniformes independentemente do valor do alpha, apesar de que à medida que o alpha aumenta a exatidão e a precisão diminuem.

#### Weather-Nominal

- Para o terceiro conjunto de dados utilizando um  $\alpha=0$  obtivemos uma exatidão de cerca de 75% e uma precisão de cerca de 83%.
- Utilizando um  $\alpha=1$  obtivemos uma exatidão de cerca de 100% e uma precisão de cerca de 100%.
- Utilizando um  $\alpha=5$  obtivemos uma exatidão de cerca de 100% e uma precisão de cerca de 100%.

Com o aumento do alpha a exatidão e a precisão ficam ambas a 100%. Isto acontece pois com o aumento do alpha a probabilidade vai ao encontro da distribuição uniforme. Visto que este é um conjunto com poucos dados os resultados vão ser pobres do ponto de vista analítico.

A utilização do estimador suavizado ajuda a anular o problema de existir uma probabilidade igual a 0, no entanto com o aumento do alpha os resultados irão ficar menos exatos pois o modelo ficará menos complexo.

#### IV - Conclusão

Aprofundámos todos os temas deste trabalho de modo a que pudéssemos enriquece-lo e de maneira a que ficássemos a compreender melhor a matéria lecionada nesta disciplina.

Do nosso ponto de vista, conseguimos implementar o classificador de acordo com o enunciado e considerámos que os valores obtidos nas análises de desempenho representam a veracidade das previsões efetuadas pelo classificador (percentagem de previsões elevada).

No desenvolvimento esteve presente alguma dificuldade pois a linguagem de programação utilizada na disciplina não pertence ao nosso percurso académico, sendo necessário mais tempo para procurar informações/soluções não referentes diretamente com o conteúdo dos classificadores mas sim com a linguagem utilizada.