

Relatório da implementação do Naïve Bayes

Aprendizagem Automática Jailson Varela nº40699

28 de julho de 2025



1 Introdução

Nesta submissão, a melhoria do trabalho se trata da implementação do algoritmo de aprendizagem Naïve Bayes e os seus métodos que lidam com os dados (compostos por atributos numéricos e classes nominais) para que, deste modo, sejam criados modelos de acordo com os diferentes valores do hiperparâmetro suave ou `var_smoothing`.

2 Metodologia

2.1 Objetivo

O objetivo para esta melhoria é criar 2 modelos Naive Bayes com os diferentes valores do hiperparâmetro suave=1e-9, 1e-5 para cada um dos dados em estudo iris.csv, rice.csv e entrega_antecipada.csv.

2.2 Descrição do algoritmo

No projeto foi implementado o tratamento dos ficheiros recebidos, os quais foram iris.csv, rice.csv, wdbc.csv (conjunto de dados descartado da avaliação) e entrega_antecipada.csv que passou por consultar e analisar os dados e também por separá-los por classes e atributos.

Com o tratamento de dados feito, implementei o algoritmo Naive Bayes(NB) como descrito, com os valores por omissão/default (suave=1e-9) . Guardamos os dados do NB em:

- self.X_train com formato numpy.array;
- self.y_train com formato numpy.array;

contendo os valores de atributos de treino, classes de treino, como o nome do atributo indica.

2.2.1 Métodos Implementados

Também foram implementados os métodos principais:

- `__init__(self, var_smoothing=1e-9)`- Inicializa o algoritmo e atribui a `var_smoothing = 1e-9` por default.
- `fit(self, X, y)`- Treina o modelo Naive Bayes (NB).
- `predict(self, example)`- Faz a predição do modelo Naive Bayes (NB) para os dados de teste.
- `score(self, X, y)`- Calcula a exatidão do modelo.

Para além dos métodos acima, implementei alguns métodos auxiliares:

- `count_class(x, cond)`- Conta o número de vezes que um elemento aparece em uma lista.
- `gaussian_probability(x_val, feature_values, var_smoothing)`- Calcula a probabilidade de um valor dado uma distribuição gaussiana.

3 Método de avaliação de desempenho

Para avaliar o modelo, foi optada a divisão treino-teste e assim analisar consoante cada par de desempenho nos conjuntos de treino e de teste.

3.1 Análises dos modelos

As análises dos modelos relatam, do mesmo que o apresentado na submissão em época normal, os desempenhos:

3.2 Modelos Naïve Bayes (NB)

3.2.1 *Iris*

Modelo NB: suave=1e-9:

- • Desempenho sobre conjunto de treino : 0.955
- • Desempenho sobre conjunto de teste : 0.974

Modelo NB: suave=1e-5:

- • Desempenho sobre conjunto de treino : 0.955
- • Desempenho sobre conjunto de teste : 0.974

GNB	suave=1e-9	suave=1e-5
conj. treino	0.955	0.974
conj. teste	0.955	0.974

Tabela 1: Desempenho do Conjunto de dados iris

3.2.2 Análise do Conjunto de dados iris

Nota-se que os modelos com...

- suave=1e-9 e suave=1e-5 são idênticos em termos de desempenho nos conjuntos de treino e de teste. Apresentam ser modelos generalistas, ou seja, com bons desempenhos no conjunto de teste, a exatidão e os modelos não estão demasiado ajustados aos dados de treino, ou seja, não sofrem nem de sub-ajustamento nem de sobre-ajustamento, que é o pretendido para poder generalizar.

3.2.3 *Rice*

Modelo NB: suave=1e-9:

- • Desempenho sobre conjunto de treino : 0.915
- • Desempenho sobre conjunto de teste : 0.925

Modelo NB: suave=1e-5:

- • Desempenho sobre conjunto de treino : 0.915
- • Desempenho sobre conjunto de teste : 0.924

GNB	suave=1e-9	suave=1e-5
conj. treino	0.915	0.915
conj. teste	0.925	0.924

Tabela 2: Desempenho do Conjunto de dados rice

3.2.4 *Análise do Conjunto de dados rice*

Nota-se que os modelos com...

- suave=1e-9 e suave=1e-5 são iguais em termos de desempenho no conjunto de treino. Apresentam ser modelos muito generalistas, pois não está demasiado próximo de 1. Tem maiores desempenhos no conjunto de teste, a exatidão, e os modelos não estão suficientemente ajustados aos dados de treino, ou seja, não sofrem nem de sub-ajustamento nem de sobre-ajustamento, que é o ideal para poder generalizar.

3.2.5 *Entrega Antecipada*

Modelo NB: suave=1e-9:

- • Desempenho sobre conjunto de treino : 0.977
- • Desempenho sobre conjunto de teste : 1.0

Modelo NB: suave=1e-5:

- • Desempenho sobre conjunto de treino : 0.977
- • Desempenho sobre conjunto de teste : 1.0

GNB	suave=1e-9	suave=1e-5
conj. treino	0.977	0.977
conj. teste	1.0	1.0

Tabela 3: Desempenho do Conjunto de dados dados entrega_antecipada

3.2.6 Análise do Conjunto de dados entrega_antecipada

Nota-se que os modelos com...

- suave=1e-9 e suave=1e-5 são idênticos em termos de desempenhos nos conjuntos de treino e de teste. Apresentam ser modelos generalistas, ou seja, com ótimos desempenhos no conjunto de teste, a exatidão, e os modelos não estão demasiado ajustados aos dados de treino, ou seja, não sofrem nem de sub-ajustamento nem de sobre-ajustamento, que é o pretendido para bem generalizar.

4 Conclusão

Noto que o conjunto de dados que criei não apresenta grande variação de dados, como melhoria poderia ter mais atributos e mais exemplos para classificar o nosso modelo e assim poder apresentar uma maior variância e dimensão dos dados.