

Classificador Ingênuo de Bayes

Arthur Abrahão Santos Barbosa
Universidade Federal de Pernambuco
Centro de Informática
Pernambuco, Brasil
aasb2@cin.ufpe.br

Filipe Samuel da Silva
Universidade Federal de Pernambuco
Centro de Informática
Pernambuco, Brasil
fss8@cin.ufpe.br

Nigel Mendes de Lima
Universidade Federal de Pernambuco
Centro de Informática
Pernambuco, Brasil
nml@cin.ufpe.br

I. OBJETIVOS

A. Objetivo Geral

Através da análise de algumas informações referentes a um indivíduo, usando um classificador ingênuo de Bayes, prever se o mesmo irá se inscrever em um depósito a prazo.

B. Objetivos Específicos

- Compreender a implementação do classificador ingênuo de Bayes
- Demonstrar a Importância do Aprendizado de máquina e suas aplicações
- Investigar o uso do Aprendizado de máquina no marketing bancário

II. JUSTIFICATIVA

Este projeto foi escolhido com base na maneira organizada e completa que o conjunto de dados foi disponibilizado e por sua afinidade em aplicar-se os conceitos existentes, o banco de dados pode ser encontrando do site Machine Learning Repository, com o nome de "Bank Marketing Data Set" [1].

Sua função é prover dados sobre a possibilidade de um cliente aderir ou não o serviço prestado pela agência com base em testes com múltiplas entradas de dados e com duas saídas possíveis, sim ou não. Seu público alvo são principalmente bancos, qualquer área de estudo sobre comportamento social e estudos sobre aprendizagem de máquina.

III. BASE DE DADOS

Os dados são referentes a campanhas de marketing direto, por meio de telefonemas, muitas vezes repetindo o contato com um mesmo cliente, de uma instituição bancária portuguesa. O objetivo de sua classificação é prever de antemão se um cliente irá aderir ou não um depósito a prazo (identificado como a variável y). O banco de dados completo está distribuído em quatro conjuntos sendo eles:

- bank-additional-full.csv com todos os exemplos (41188) e 20 entradas, ordenadas por data (de maio de 2008 a novembro de 2010), muito próximo aos dados analisados em [Moro et al., 2014]
- bank-additional.csv com 10% dos exemplos (4119), selecionados aleatoriamente de 1) e 20 entradas.
- bank-full.csv com todos os exemplos e 17 entradas, ordenadas por data (versão mais antiga deste conjunto de dados com menos entradas).

- bank.csv com 10% dos exemplos e 17 entradas, selecionadas aleatoriamente a partir de 3 (versão mais antiga deste conjunto de dados com menos entradas).

Para o projeto será usado o item 3 (bank-full.csv) contendo 16 variáveis de entrada e uma de saída sendo elas (Nota: os exemplos de entrada abaixo estão todos em inglês, pois é assim que se encontra no banco de dados):

A. 16 Variáveis de entrada:

- age: (numérico).
- job: tipo de trabalho (categórico: 'admin.', 'blue-collar', 'entrepreneur', 'housemaid', 'management', 'retired', 'self-employed', 'services', 'student', 'technician', 'unemployed', 'unknown').
- marital: estado civil (categórico: 'divorced', 'married', 'single', 'unknown'; nota: 'divorced' significa divorciado ou viúvo).
- education: (categórico: 'basic.4y', 'basic.6y', 'basic.9y', 'high.school', 'illiterate', 'professional.course.4y', 'professional.course.6y', 'professional.course.9y', 'university.degree').
- default: possui crédito inadimplente? (categórico: 'no', 'yes', 'unknown').
- balance (numérico).
- housing: possui crédito de habitação? (categórico: 'no', 'yes', 'unknown').
- loan: possui crédito pessoal? (categórico: 'no', 'yes', 'unknown').
- contact: tipo de comunicação do contato (categórico: 'cellular', 'telephone').
- day: dia do último contato (numérico).
- duration: duração do último contato em segundos (numérico, nota importante, este atributo pode afetar muito a saída, por exemplo se a duração for '0' então y = 'no', todavia a duração é desconhecida até que a chamada tenha terminado, nesse caso y é conhecido, sendo assim a entrada só será incluída se realmente for necessária).
- month: último mês do contato (categórico: 'jan', 'feb', 'mar', ..., 'nov', 'dec').
- campaign: número de contatos realizados durante a campanha para este contato (numérico, inclui o último contato feito).
- pdays: número de dias que se passaram desde o último contato de uma campanha anterior para este cliente (numérico; 999 significa que este cliente não foi contatado antes).

- **previous**: número de contatos realizados para este cliente antes dessa campanha (numérico).
- **poutcome**: resultado da campanha de marketing anterior (categórico: 'failure', 'nonexistent', 'success').

B. Uma Variável de saída:

- **y** : o cliente assinou o depósito a prazo (binário: 'yes', 'no').

IV. ANÁLISE EXPLORATÓRIA DOS DADOS

A. Descrição Estatística dos dados

- O campo **count**, representa a quantidade de instancias que contém aquele atributo.
- O campo **unique** se refere a quantas categorias existem daquele atributo, caso ele seja do tipo texto descritivo (não numérico).
- O campo **top** se refere a categoria mais encontrada daquele atributo.
- O campo **freq** é o número de vezes que a categoria top daquele atributo foi encontrada.
- O campo **mean** informa a média dos valores daquele atributo, caso sejam numéricos.
- O campo **std** informa o desvio padrão dos dados daquele atributo.
- O campo **min** mostra o menor valor numérico que aquele atributo possui nas amostras, considerando todas as instâncias.
- O campo **25%** se refere ao primeiro quartil das amostras daquele atributo.
- O campo **50%** se refere ao segundo quartil das amostras daquele atributo.
- O campo **75%** se refere ao terceiro quartil das amostras daquele atributo.
- O campo **max** informa o maior valor que aquele atributo possui entre todas as amostras.

B. Dados Disponíveis ou ausentes

- Os atributos **age**, **day**, **duration**, **campaign**, **pdays**, **previous**, possuem informação do tipo numérico e são valores quantitativos discretos.
- O atributo **balance**, possui informação do tipo numérico e são valores quantitativos contínuos.
- Os atributos **job**, **marital**, **education**, **default**, **housing**, **loan**, **contact**, **poutcome**, possuem informação do tipo texto descritivo(categorias), seus valores são qualitativos nominais.
- Os atributo **month**, possui informação do tipo texto descritivo(categorias), seus valores são qualitativos ordinais.
- Não existem dados faltando(ausentes) em nenhum campo de nenhuma instância.

C. Gráficos

- (1) Primeiro temos o Histograma (Fig.1) das idades em relação ao total de amostras.
 - Com destaque para a maior parte dos valores estarem próximos ao intervalo entre 30 e 40 anos.

- Temos a média como 40,936210, e o desvio padrão de 10,618762.

- (2) Segundo temos o número de pessoas que obtiveram empréstimo de casa em relação a sua grau de educação escolar (Fig.2).

- Destaque para o grau secundário que tem a maior parte dos que pegaram empréstimo.

- O grau secundário(ensino médio) é o que mais aparece nas amostras.

- (3) Temos o grau da educação escolar em relação ao Saldo médio anual (Fig.3)

- O destaque fica para os valores Outliers do saldo, principalmente pro grau tertiary(Faculdade).

Fig. 1. Exemplo

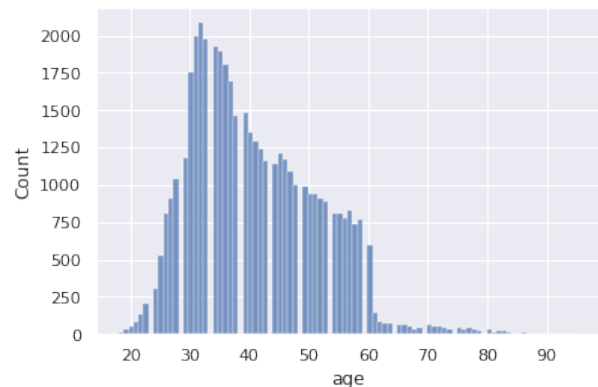


Fig. 2. Exemplo2

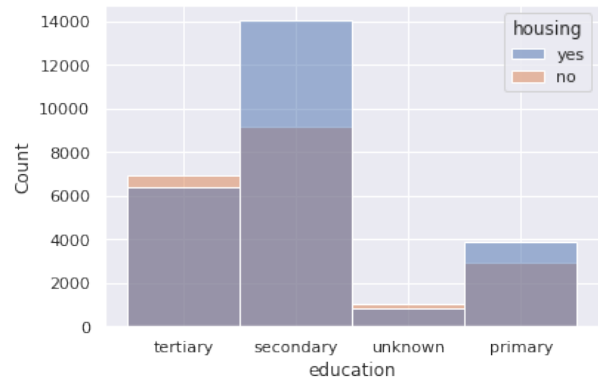
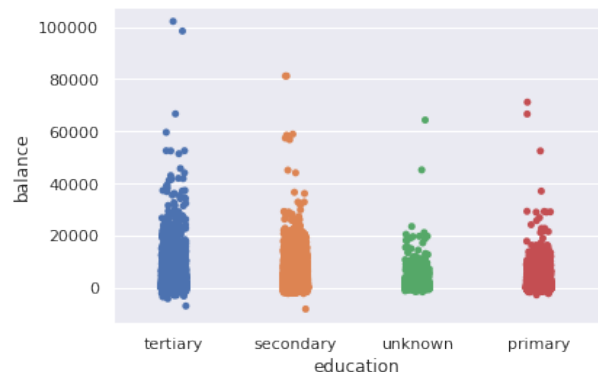


Fig. 3. Exemplo4



- (4) Temos o número de empréstimos de casa em relação ao tipo de contato que foi registrado (Fig. 4).
- O destaque está no tipo de contato celular, que possui um número de empréstimo 'sim' muito próximo ao 'não', apesar de 25130 de todas as amostras serem com valor *sim*.
- (5) Temos o boxplot (Fig.5) de empréstimo pessoal em relação a idade.
- Temos os valores do primeiro quartil, do terceiro quartil e da mediana de cada boxplot, para as respostas *sim* e *não*.
 - Podemos ver que a mediana de ambos os boxplots, têm uma valores muito próximos.
 - Outro destaque vai para os outliers das amostras com respostas *sim*.

Fig. 4. Exemplo4

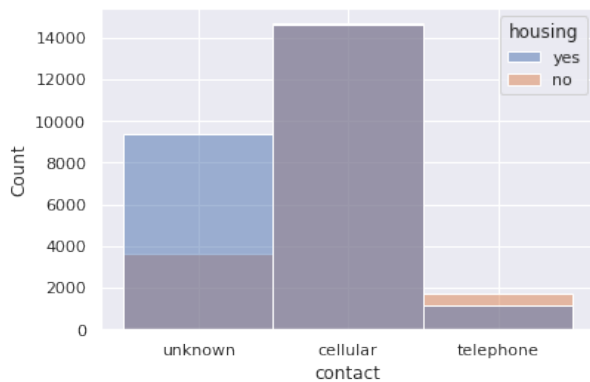
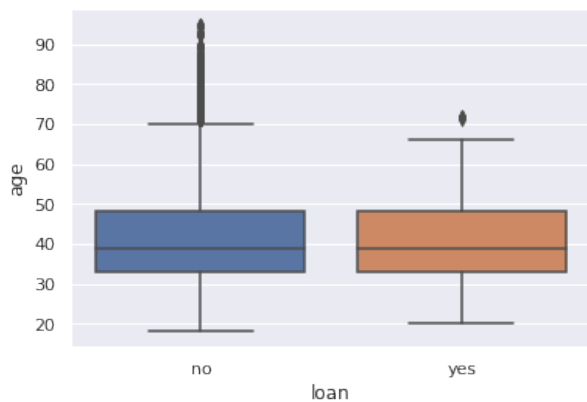


Fig. 5. Exemplo5



- (6) Temos o strippplot (Fig.6) de empréstimo pessoal em relação a idade.
- Com destaque para as idades que foram os outliers, em relação ao Não Empréstimo, temos os mais velhos.
 - Já do outro lado, os que sim, tiveram empréstimo pessoal estavam em um intervalo de idade menor, e mais jovem.
- (7) Temos o strippplot (Fig.7) dos meses em relação a campanha que foi realizada.
- Com destaque para os meses do meio do ano que contemplam a maior parte das amostras.

- Enquanto que no final do ano, foram registrados as menores parcelas das amostras.

Fig. 6. Exemplo6

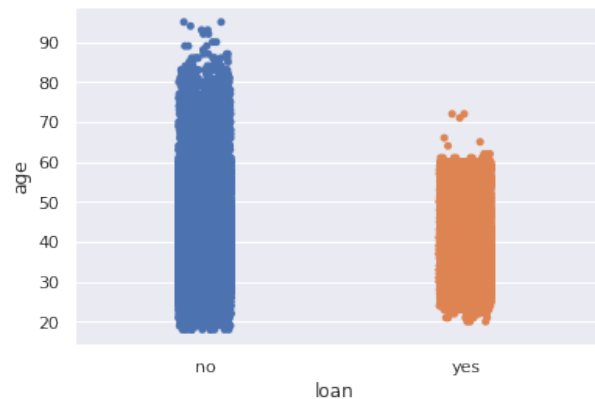
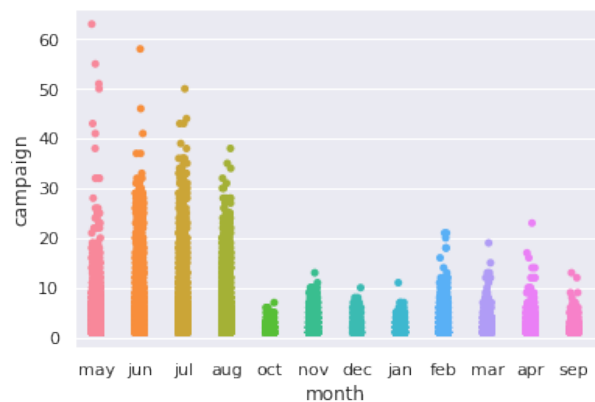


Fig. 7. Exemplo7



D. Outliers

Nesta seção temos a quantidade de pessoas que aceitaram ou recusaram assinar um investimento, em cada classe ou categoria de cada atributo para os atributos do tipo texto descritivo, e em relação aos valores numéricos de um atributo, para um atributo do tipo numérico.

V. CLASSIFICADOR INGÊNUO DE BAYES

Baseado no Teorema de Bayes, nome em homenagem ao matemático e pastor presbiteriano inglês Thomas Bayes, que formulou uma função probabilística com o ideal de provar a existência de Deus, Naive Bayes é um algoritmo de classificação probabilística muito utilizado para aprendizado de máquina (Machine Learning). O algoritmo possui a habilidade de categorizar textos baseado na frequência em que as palavras são dispostas, o exemplo mais comum são os filtros de e-mail que podem utilizar o Naive Bayes para identificar se uma mensagem é um spam apenas lendo a disposição das palavras utilizadas. O nome Naive, do português ingênuo, vem do fato que o algoritmo desconsidera totalmente a correlação entre as variáveis, tratando cada uma de maneira independente.

A. Definição Formal do Teorema de Bayes

O teorema é um corolário da lei da probabilidade total e é descrito da seguinte maneira, sejam A e B dois eventos e P(A) e P(B) as probabilidades de A e B, respectivamente, sendo P(B) diferente de 0, então o Teorema de Bayes nos diz que,

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}, \quad (1)$$

De maneira análoga, com P(A) diferente de 0,

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}. \quad (2)$$

B. Tipos de Classificadores Ingênuos de Bayes

A biblioteca Scikit learn apresenta diversos tipos de Classificadores. As diferenças principais entre eles são principalmente em relação as suposições feitas em relação a distribuição de probabilidade $P(x_i|y)$. No projeto foram-se aplicados dois tipos de Classificadores, o Gaussiano e o Categórico:

1) *Bayes Ingênuo Gaussiano*: O classificador de bayes gaussiano supõe que as variáveis seguem uma distribuição normal, a verossimilhança das variáveis são supostas como gaussianas:

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} e^{-\frac{(x_i-\mu_y)^2}{2\sigma_y^2}} \quad (3)$$

Os parametros σ_y e μ_y são estimados usando máxima verossimilhança.

2) *Bayes Ingênuo Categórico*: O Bayes Ingênuo Categórico implementa o classificador de bayes ingênuo para distribuição categórica de dados.

A probabilidade de x_i ser da categoria t dado a classe é c é estimado como:

$$P(x_i = t|y = x; \alpha) = \frac{N_{tic} + \alpha}{N_c + \alpha n_i} \quad (4)$$

onde:

- N_{tic} é o número de vezes que a categoria t aparece na amostra e pertence a classe c
- N_c é o número de amostras que pertence a classe c
- α é um parâmetro de calibração

C. Sobre o Projeto

Para montar o classificador foi necessário passar pelas seguintes etapas:

- 1) Carregar o dataframe através da biblioteca Pandas
- 2) Separação dos valores de x e de y
- 3) Conversão dos valores da Base de Dados para números inteiros através do módulo preprocessing da biblioteca sklearn.model_selection
- 4) Separação dos Dados para treino e para teste através do módulo train_test_split da biblioteca sklearn.model_selection

- 5) Treino do Classificador Ingênuo de Bayes através dos módulos GaussianNB e CategoricalNB da biblioteca sklearn.naive_bayes. Foram criados dois classificadores: Um Gaussiano que assume que todas as variáveis de treino são normais (o que não é verdade), e um categórico que assume que as variáveis de treino seguem uma distribuição categórica (Mais próximo da realidade).

VI. EXPERIMENTOS

A. Experimentos Iniciais

Após treinar ambos os classificadores (Gaussiano e Categórico), usando 20 por cento dos dados para teste. Foi verificado alguns valores referente ao teste:

- Precision: Precision é a razão

$$\frac{t_p}{t_p + f_p} \quad (5)$$

onde:

- t_p é o número de verdadeiros positivos
- f_p é o número de falsos positivos.

Precision é intuitivamente a habilidade do classificador não marcar como positivo uma amostra que é negativa. O melhor valor de Precision é 1 e o pior é zero. [2]

- Accuracy:
- Recall-Score: O Recall Score é a razão:

$$\frac{t_p}{t_p + f_n} \quad (6)$$

onde:

- t_p é o número de verdadeiros positivos
- f_n é o número de falsos negativos

O Recall Score é intuitivamente a habilidade do classificador de encontrar todas as amostras positivas. O melhor valor do Recall Score é 1 e o pior valor é 0. [3]

- F1-Score: O F1 Score pode ser interpretado como a média ponderada da precisão e recall. O melhor valor que o F1 score pode alcançar é 1, o pior é 0. A contribuição relativa da precisão e recall para o F1 score são iguais. A fórmula para o F1 score é:

$$F1 = \frac{2 \cdot (precision \cdot recall)}{precision + recall} \quad (7)$$

[4]

- Confusion Matrix

	Categórico	Gaussiano
Precision	0.89	0.84
Accuracy	0.89	0.84
Recall Score	0.89	0.84
F1-Score	0.89	0.84

TABLE I
RELATÓRIO DE CLASSIFICAÇÃO POR LABEL DO CLASSIFICADOR
CATEGÓRICO.

	0	1
Precision	0.93	0.53
Recall Score	0.95	0.43
F1-Score	0.94	0.47

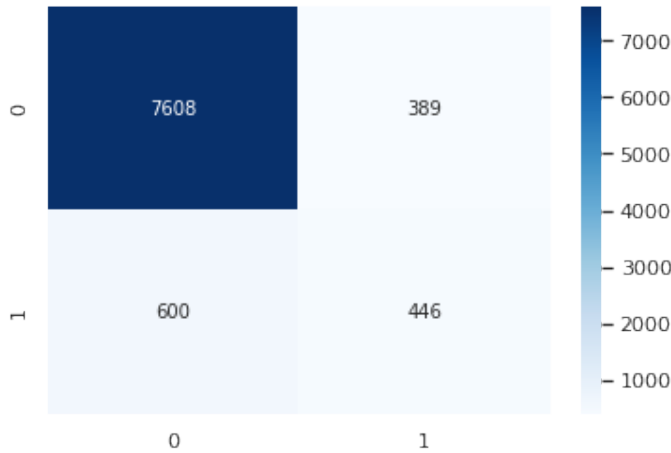


Fig. 8. Confusion Matrix do Classificador Categórico

TABLE II
RELATÓRIO DE CLASSIFICAÇÃO POR LABEL DO CLASSIFICADOR
GAUSSIANO

	0	1
Precision	0.89	0.49
Recall Score	1.00	0.03
F1-Score	0.94	0.06

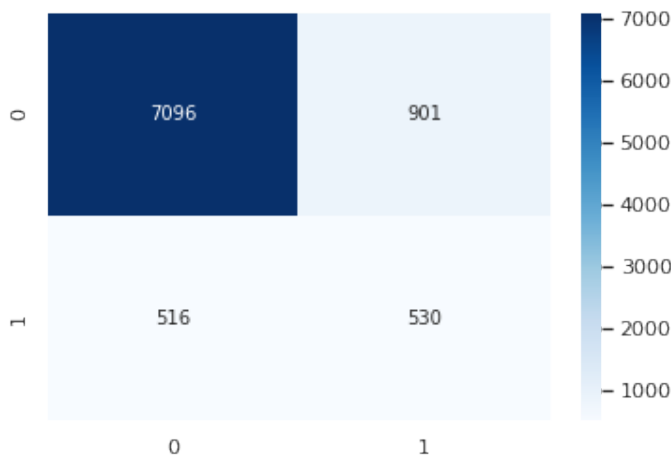


Fig. 9. Confusion Matrix do Classificador Gaussiano

VII. ANÁLISE DOS RESULTADOS

VIII. CONCLUSÕES E DISCUSSÕES

REFERENCES

- [1] Bank marketing data set. [Online]. Available: <https://archive.ics.uci.edu/ml/datasets/Bank+Marketing>
- [2] "sklearn.metrics.precision_score." [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.precision_score.html
- [3] "sklearn.metrics.recall_score." [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.recall_score.html
- [4] sklearn.metrics.f1_score. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html
- [5] S. K. Reddy, "Categorical naive bayes classifier implementation in python," Oct. 2020. [Online]. Available: <https://www.inblog.in/Categorical-Naive-Bayes-Classifier-implementation-in-Python-dAVqLWkf7E>
- [6] Muthu, "Understanding the classification report through sklearn," Jul. 2018. [Online]. Available: <https://muthu.co/understanding-the-classification-report-in-sklearn/>
- [7] "1.9. naive bayes." [Online]. Available: https://scikit-learn.org/stable/modules/naive_bayes.html