

# **Aprendizagem de Máquina**

## **Problema 2**

---

1.

**Solução**

# Passo a passo da solução



- 1- Importar os dados dos arquivos recebidos
  - amazon\_cells\_labelled.txt
  - imdb\_labelled.txt
  - Yelp\_labelled.txt

# Passo a passo da solução

## ■ 2- Tratar os dados

Talvez seja vantajoso fazer o seguinte...

- Tudo em letra minúscula/maiuscula?
- Remover pontuação?
- Remover caracteres especiais?
- Remover palavras que não nos auxiliam na classificação?

# Passo a passo da solução



- OBS: Tratamento dos dados, seleção de palavras (features) poderia ser feito utilizando módulos, bibliotecas de terceiros.
  - Por exemplo, plataformas de processamento natural de linguagens, como NLTK( Natural Language Toolkit).
  - Eu resolvi fazer manualmente...

# Passo a passo da solução

## ■ 3- Criar um dicionário de palavras

- Uma possível forma de resolver esse problema de análise de sentimentos é tratar cada sentença como um vetor.

**Hello word** -> [0,0,0,1,0,0,0,1]

- Porém são necessárias "funções de base" para que haja essa transformação de sentenças para vetores.

**Hello** -> [0,0,0,1,0,0,0,0] |||| **word** -> [0,0,0,0,0,0,0,1]

- Cada uma das palavras do texto será utilizada como "função de base", ou seja, elas serão utilizadas para compor a sentença completa.

**Hello word** = [0,0,0,1,0,0,0,0] + [0,0,0,0,0,0,0,1]

**Hello word** = [0,0,0,1,0,0,0,1]

# Passo a passo da solução

- Agora que as sentenças são vetores, podemos utilizar os métodos de classificação presentes no sklearn.

Entrada	Vetor Associado	Saída
abc bcd efg	[1, 1, 1, 0]	1
bcd efg hij	[0, 1, 1, 1]	0
abc hij	[0, 0, 0, 1]	1

**Bases: [ abc. bcd. efg, hij ]**

# Melhor resultado

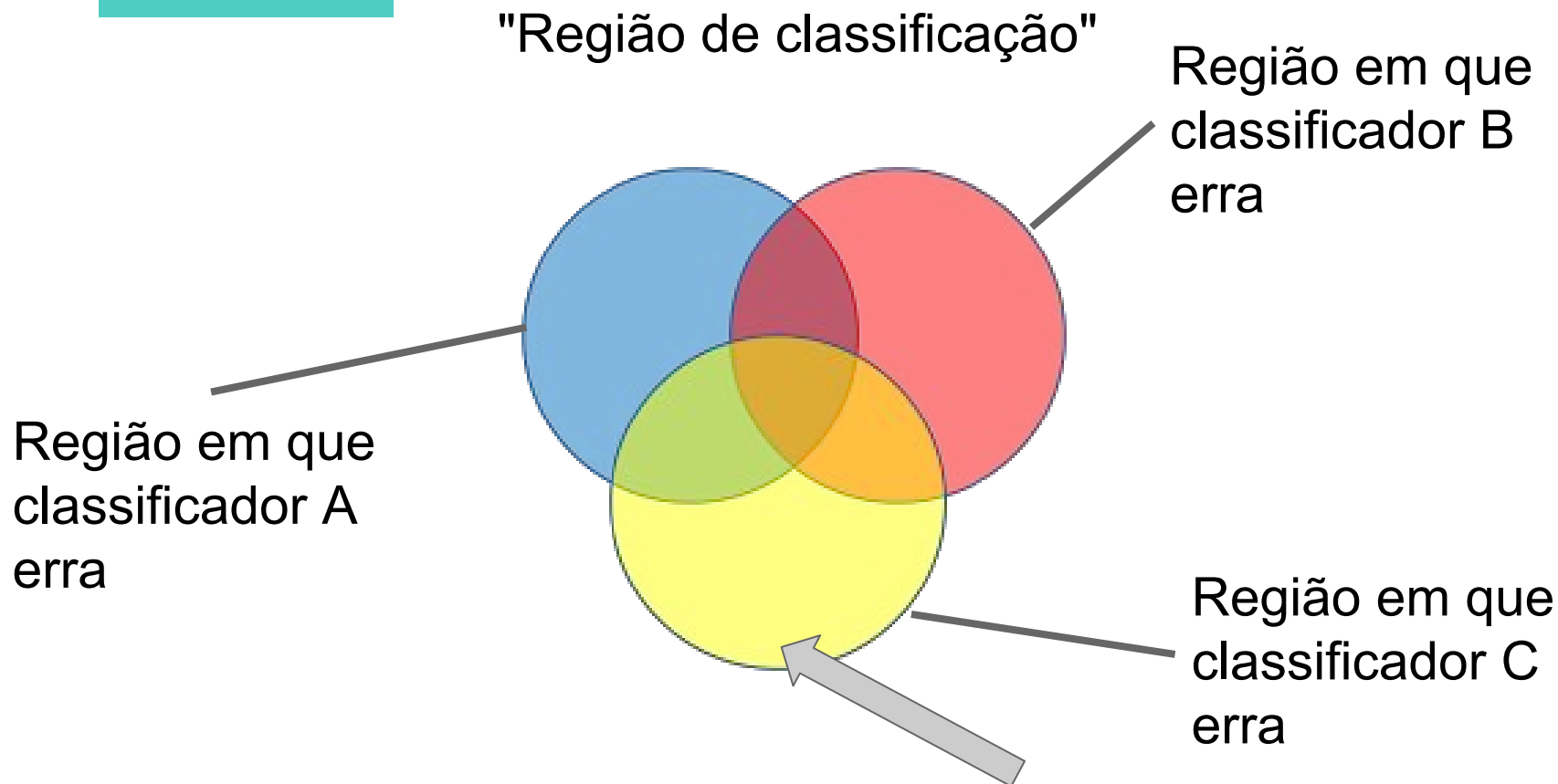


- Validação do modelo foi feita utilizando K-Fold, com  $k = 10$ .
- O melhor resultado obtido foi usando **Ensemble**.
  - Foi utilizado um esquema de votação em que os algoritmos utilizados e seus respectivos pesos foram:
    - Multinomial Naive Bayes - Peso 3
    - KNN - Peso 2
    - Logistic Regression - Peso 2

Precisão de aproximadamente **85.3%**



# Intuição Ensemble



Considerando pesos iguais: O que acontece se C erra mas A,B acertam? Classificamos corretamente.

# Ensemble no Projeto




Classificação é definida basicamente pelo Multinomial Naive Bayes (Peso 3), a não ser que:


Classificação do KNN (Peso 2) e Logistic Regression (Peso 2) sejam iguais E diferente do Multinomial Naive Bayes.

$(\text{Peso } 2) + (\text{Peso } 2) > (\text{Peso } 3)$ .

2.

**Comentários**

- 
- Multinomial Naive Bayes - Comumente usado para sentiment analysis.
    - Em comparação com outras variações do NB, ele é recomendado quando a quantidade de palavras repetidas na frase não importa.
  - Notou-se que algoritmos mais "simples", próximos de lineares generalizaram melhor.

- 
- Nesse projeto trabalhamos com dados de grande dimensão, sendo necessário, portanto, uma grande quantidade de dados.
  - O ideal seria inclusive ter mais dados, pois são relativamente poucas sentenças em comparação com o tamanho da base (quantidade de palavras do dicionário).
  - Começamos a perceber também demora para treinar alguns modelos.