

# Preparando terreno no mundo da **Machine Learning**

Um passo de cada vez!

# AGENDA

---

- Classificadores Bayesianos
  - ◆ Análise Discriminante Gaussiano
  - ◆ Naive Bayes
- Métricas
  - ◆ Acurácia
  - ◆ Matriz de Confusão
  - ◆ Precision/Recall
  - ◆ F1-Score
- Analisador de Fake News
  - ◆ NLP
  - ◆ Criação de Modelos
  - ◆ Avaliação dos Modelos

# Classificadores Bayesianos

“Dado a probabilidade que choveu esta noite e dado a probabilidade que o sol não estava tão quente essa manhã, então o chão ainda está molhado?”

# Classificador Bayesiano

- Dado um conjunto de características (atributos)  $x$  de um padrão, a qual classe o padrão pertence

# Métricas

Como eu sei que um modelo é melhor que outro?

Nós iremos conhecer 4 métricas hoje!

# ACCURACY

$$Acc = \frac{1}{n} \sum 1(\hat{y}_i = y_i)$$

Predicted  $y$

True  $y$

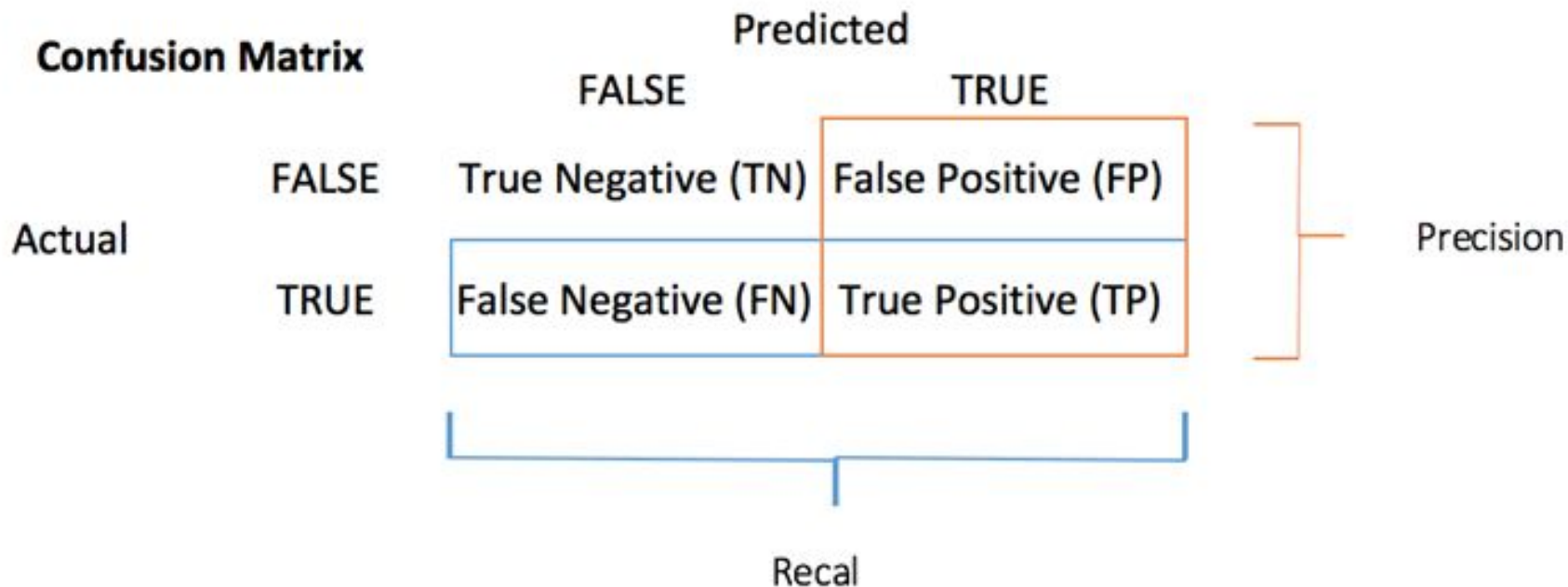
Indicator function

number of observations

A common metric in classification. Fails when we have highly imbalanced classes. In those cases F1 is more appropriate.

ChrisAlbon

# Matriz de Confusão





# PRECISION

Precision is the ability a classifier to not label a true negative observation as positive.

True Positive

---

True Positive + False Positive

ChrisAlbon

# RECALL

"Recall is about the real positives"

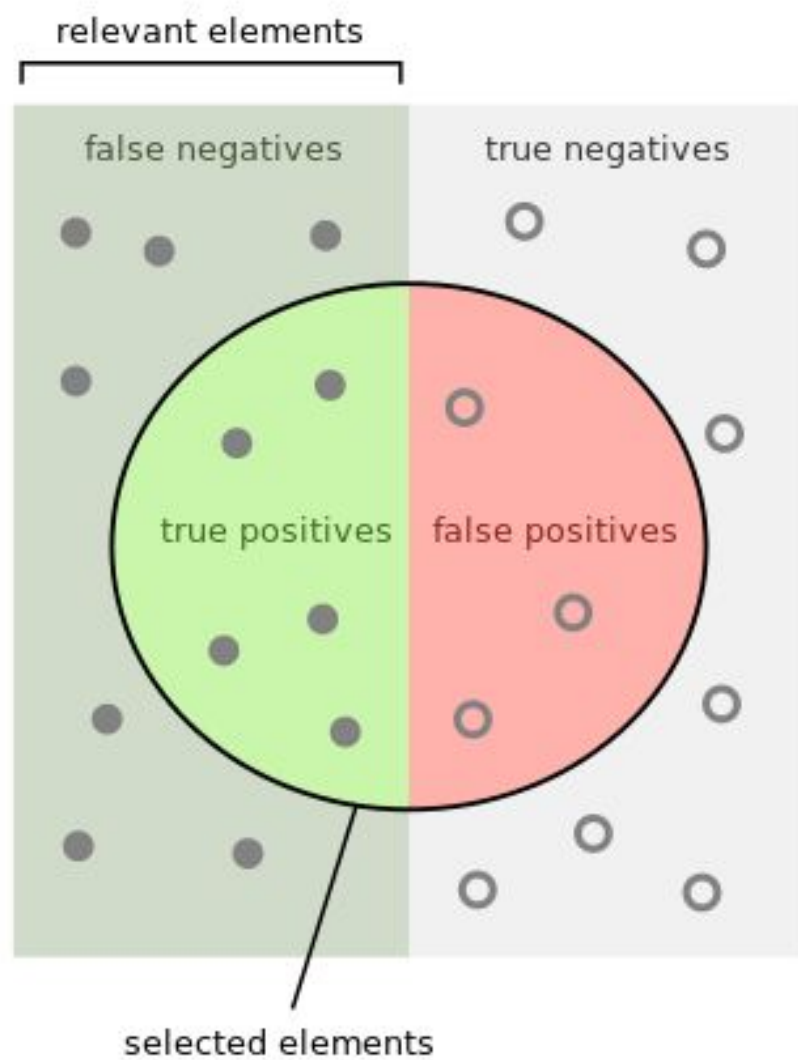
True Positives

---

True Positives + False Negatives

Recall is the ability of the classifier to find positive examples. If we wanted to be certain to find all positive examples, we could maximize recall.

Chris Albon



How many selected items are relevant?

$$\text{Precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

How many relevant items are selected?

$$\text{Recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

# F1 SCORE

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

F1 score is the harmonic mean of precision and recall. Values range from 0 (bad) to 1 (good).

Chris Albon

$$\text{Recall} = \frac{TP}{TP + FN} \quad \text{Actual}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Accuracy} = \frac{TP + TN}{\text{Total}}$$

Predicted

	Pos	Neg
Pos	TP	FP
Neg	FN	TN

# Construindo um Analisador de Fake News

Agora te peguei, Bolsonaro! hahahah

# Por que?

- Cenário político-social mundial tem se mostrado fortemente impactado pela disseminação de Fake News. (*Bolsonaro só fala merda*)
- Notoriedade com o crescente uso de redes sociais.
- **Objetivo de elaborar um modelo que possa classificar se determinada notícia é verdadeira ou falsa.**

**Que tarefa é essa?**



**Quais os algoritmos estudados  
poderíamos utilizar?**

# Tarefa

- Problema de **Classificação Binária**.
- Usaremos:
  - ◆ Regressão Logística
  - ◆ K-NN
  - ◆ Árvore de Decisão
  - ◆ Análise Discriminante Gaussiano

# Dado

- Conjunto de dados utilizado foi extraído do Kaggle, fruto de um web crawler de diferentes fontes americanas, composto por **4009 registros de notícias**.
  - ◆ 2137 notícias falsas e 1872 notícias verdadeiras,
- Os atributos originais desse conjunto de dados consistem:
  - ◆ *URL*
  - ◆ *Headline*
  - ◆ *Body*
  - ◆ *Label*
- Atributos utilizados
  - ◆ *Label*
  - ◆ *Headline*

# Pré-processamento dos dados

- Lower Case
- Remoção de pontuação
- Remoção de StopWords
- Remoção de palavra mais e menos frequentes
- Stemmização/Lemmatização
- Bag of Words
- Redução da Dimensionalidade
  - ◆ Fisher Score
  - ◆ PCA

# Métricas

- Precision-Recall
- F1-Score
- Matriz de Confusão
- Acurácia

**HANDS-ON!**