

# Introdução ao Aprendizado de Máquinas (Machine Learning)

Douglas Rodrigues

Universidade Federal Fluminense



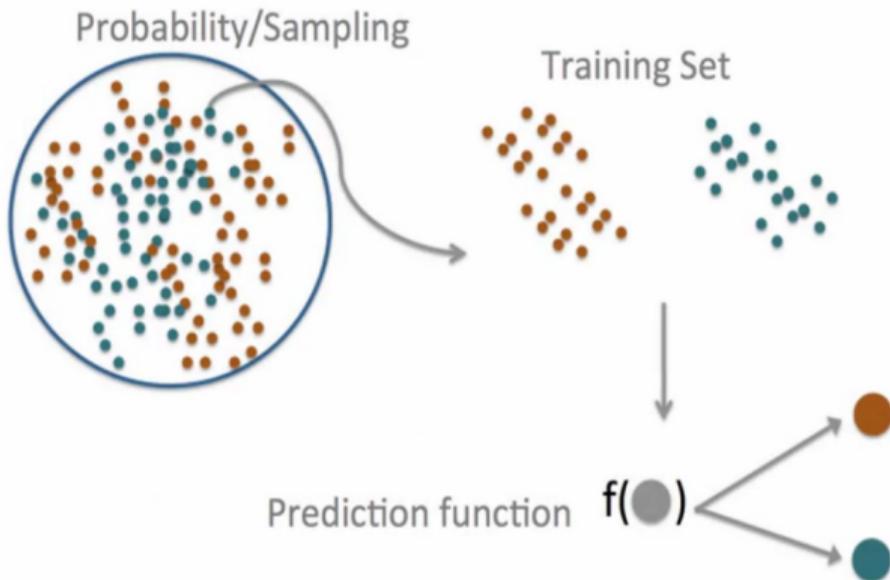




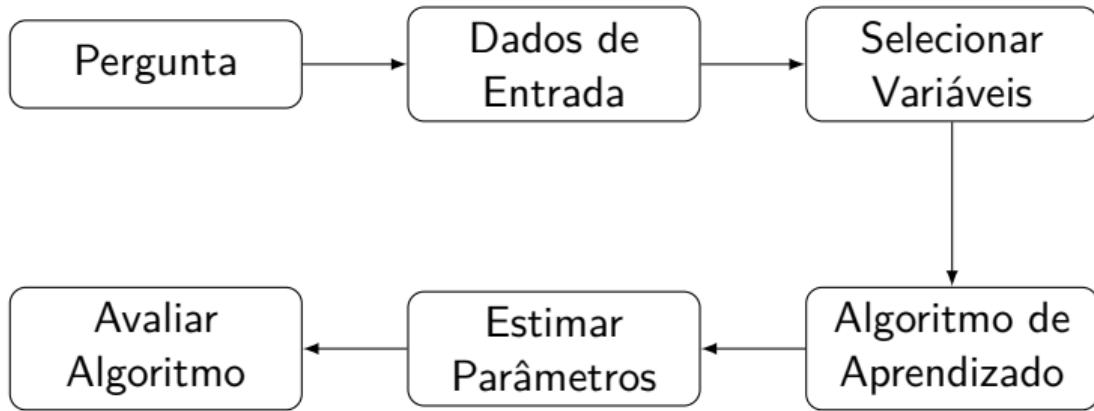
**População**



**Amostra**



# Etapas



# Exemplo

- Pergunta: Posso detectar automaticamente se um email é SPAM ou não?

- Pergunta: Posso detectar automaticamente se um email é SPAM ou não?
- De forma concreta: Posso utilizar características quantitativas de uma email para classifica-lo como SPAM?

Dear Feifei,

can you send me the address, so I can send you the invitation.

Thanks, Rob.

Dear Feifei,

can you send me the address, so I can send you the invitation.

Thanks, Rob.

Vamos analisar a frequência da palavra *you*

Dear Feifei,

can you send me the address, so I can send you the invitation.

Thanks, Rob.

Vamos analisar a frequência da palavra *you*

$$\text{Frequência de } you = \frac{2}{17} = 0.1176 = 11,76\%.$$

- Vamos instalar um pacote, que possui um banco de dados quantitativo de SPAM's.

```
> install.packages("kernlab")
> library(kernlab)
> data(spam)
```

- Esse banco de dados, coletado pela *Hewlett-Packard Labs*, contém informações de 4601 e-mails, que foram classificados como spam ou não spam.

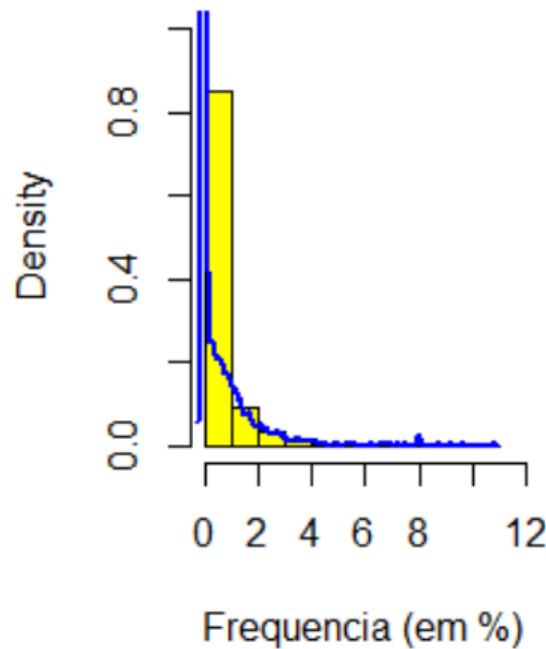
- Possui 57 variáveis, indicando a frequência (em %) de certas palavras, números e símbolos, em cada email:
  - ① As primeiras 48 contém a frequência de algumas palavras e números.
  - ② Por exemplo, a variável num650 indica a frequência do número 650.
  - ③ As variáveis 49-54 indicam a frequência de caracteres, como pontos ou parênteses.
  - ④ As variáveis 55-57 contém informações sobre letras em maiúsculo (média, maior comprimento e quantidade).
  - ⑤ A variável 58 indica se o email analisado é "spam" ou "nonspam".

- Como exemplo, vamos estudar a frequência da palavra *your* em emails considerados spam e não spam.

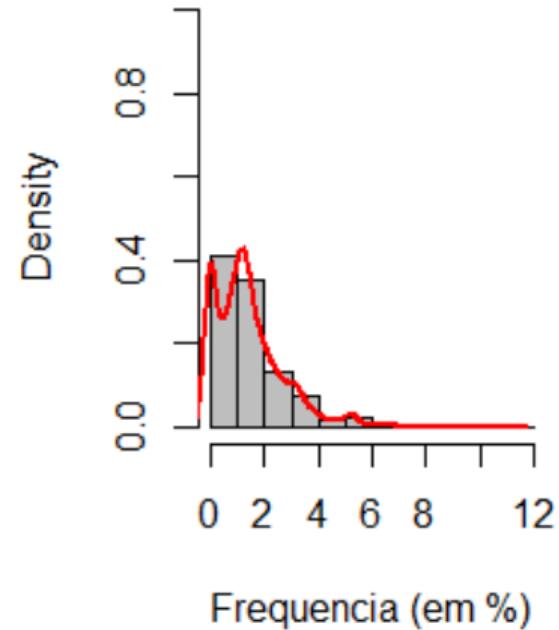
- Como exemplo, vamos estudar a frequência da palavra *your* em emails considerados spam e não spam.
- Para isso, vamos construir o histograma e linha de densidade da frequência de aparecimento dessa palavra, separando em não SPAM e SPAM.

# Histograma da Frequêcia de *your*

**Não SPAM**



**SPAM**



- Ideia: criar um critério para identificar SPAM baseado na frequência de aparecimento da palavra *your*.
- Ou seja, queremos um **C** tal que

$$\begin{cases} \text{Se a frequência de } \textit{your} > \mathbf{C} \Rightarrow \text{classifica como SPAM} \\ \text{Se a frequência de } \textit{your} \leq \mathbf{C} \Rightarrow \text{não classifica como SPAM} \end{cases}$$

- Ideia: criar um critério para identificar SPAM baseado na frequência de aparecimento da palavra *your*.
- Ou seja, queremos um **C** tal que
  - { Se a frequência de *your* > **C** ⇒ classifica como SPAM
  - { Se a frequência de *your* ≤ **C** ⇒ não classifica como SPAM
- O gráfico das densidades sugere que **0.5** é um bom candidato a valor de **C**

- Vamos criar um algoritmo de classificação, e aplicar nos nossos dados-teste.  
> prediction <- ifelse(spam\$your>0.5, "spam", "nonspam")

- Vamos criar um algoritmo de classificação, e aplicar nos nossos dados-teste.  
> prediction <- ifelse(spam\$type>0.5, "spam", "nonspam")
- Vamos avaliar nosso algoritmo de classificação, baseado nos resultados apresentados.  
> table(prediction,spam\$type)/length(spam\$type)

	Não SPAM	SPAM
Não SPAM	0,4590306	0,101717
SPAM	0,1469246	0,292328

Taxa de acerto  $\approx 0.4590306 + 0.2923278 \approx 75,14\%$ .