

# Classificação Texual e Análise de Sentimento

*Prof. Marcelo Pita*

Pós em Inteligência Artificial  
Processamento de Linguagem Natural



# Tarefas de mineração de dados

---

Mineração de dados pode ser analisada a partir da sua capacidade em realizar um conjunto de tarefas:

- **Descrição:** Os dados utilizados em uma análise podem descrever um comportamento ou tendência
- **Classificação:** A tarefa de classificação consiste em determinar a classe de um registro. Nesta tarefa, os algoritmos utilizados produzem modelos que descrevem as características de cada classe
- **Regressão:** De forma similar ao processo de classificação, a regressão procura predizer o valor de um registro a partir de um modelo gerado através de dados conhecidos

# Tarefas de mineração de dados

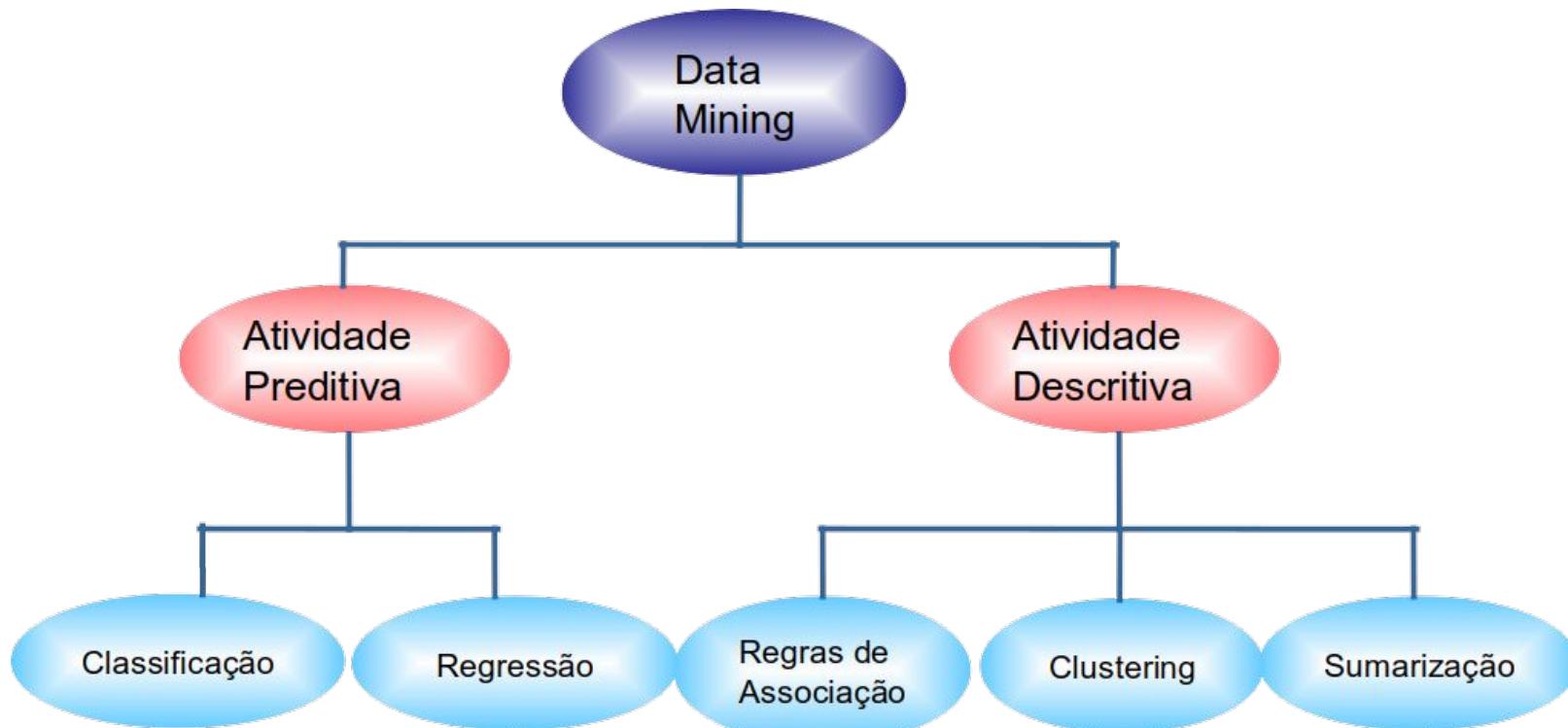
---

Mineração de dados pode ser analisada a partir da sua capacidade em realizar um conjunto de tarefas:

- **Predição:** Similar ao processo de classificação e regressão a tarefa de predição visa estimar o valor futuro de uma variável
- **Agrupamento:** Na tarefa de agrupamento registros similares são identificados. Cada grupo (cluster) é formado por um conjunto de registros similares entre si; entretanto, diferentes dos registros pertencentes aos demais grupos.
- **Associação:** A tarefa de associação consiste em identificar atributos relacionados. Em geral, a associação é expressa através de regras do tipo Se X então Y; em que, X e Y são conjuntos de atributos categóricos

# Tarefas de mineração de dados

---



# Tarefas de mineração de dados

---

Técnicas de MD também podem ser classificadas a partir da perspectiva de aprendizagem de máquina:

- **Aprendizado supervisionado:**
  - Neste tipo de aprendizagem existe um "professor" que avalia a resposta
  - Algoritmos para classificação, regressão.
- **Aprendizado não-supervisionado:**
  - Nesta forma de aprendizagem não existe "professor"
  - Algoritmos para agrupamento
- **Aprendizado por reforço:**
  - Aprendizagem por “castigo” e “recompensa”

# Treinamento de modelos

---

Para construir um modelo os dados são divididos em:

- *Treinamento*
- *Validação*
- *Teste*

A base de treinamento é utilizada para construir o modelo e deve ser representativa para permitir um efetivo processo de aprendizagem.

Deve, sempre que possível, cobrir todo espaço amostral (extremos).

Em geral os dados numéricos são normalizados:

- 0 - 1
- 0.2 – 0.8

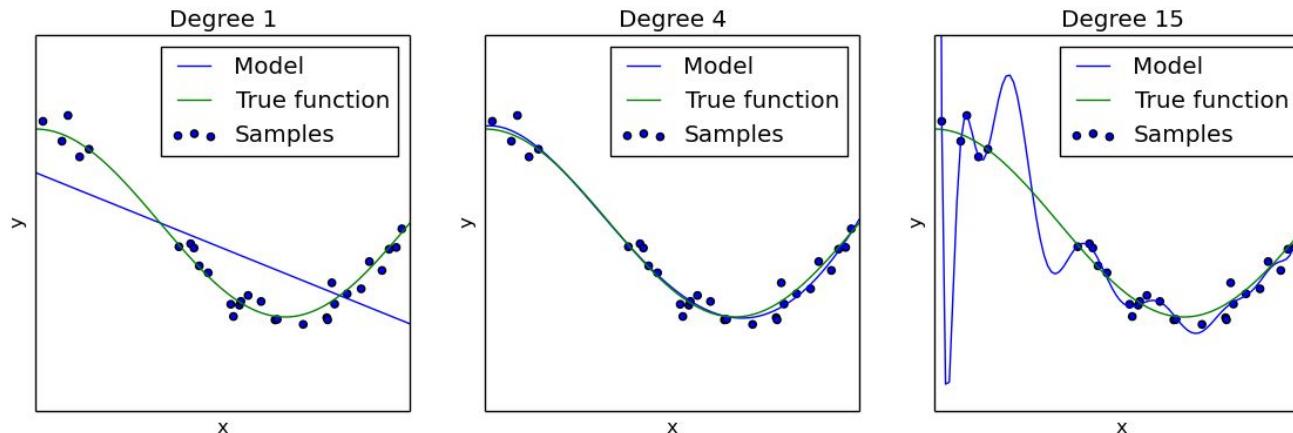
# Treinamento de modelos

---

Existem diversos critérios para divisão dos dados em treinamento e teste

- 70% - 30%
- 40% - 60%, etc
- *Validação cruzada*
  - Divilda os dados em N partes
  - Construa N-1 modelos com amostras de tamanho N-1 (treinamento)
  - Faça o teste com enésima parte

# Overfitting e Underfitting



- A função linear de grau 1 não é suficiente para um bom treinamento (*underfitting*)
- A função de grau 4 tem uma boa aproximação
- Funções de maior grau resultam em *overfitting*
  - Perda da capacidade de generalizar
  - Modelo aprende até os erros e se torna muito específico

# Problema de classificação

---

**Classificação** é a tarefa de predizer uma classe ou rótulo para um dado desconhecido.

Recebe como entrada o valor correto de uma função desconhecida para entradas específicas e tenta recuperar a função.

Dada uma coleção de exemplos de  $f$ , retornar uma função  $h$  que aproxima  $f$

- $f$  é a função-alvo
- $h$  é a hipótese
- $(x, f(x))$  são exemplos de entrada

# Problema de classificação

---

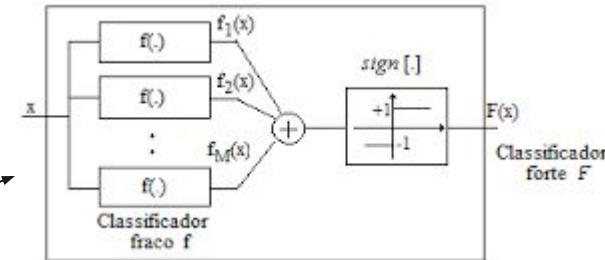
## Exemplos:

- *Detecção de spam*
  - $x$  : descreve o conteúdo de mensagens de e-mail
  - $f(x)$  : indica se a mensagem é spam ou não (em geral fornecido pelos usuários)
- *Detecção de fraude*
  - $x$  : características da transação
  - $f(x)$  : indica malha ou não malha
- *Diagnóstico de doenças*
  - $x$  : descreve o resultado de exames do paciente
  - $f(x)$  : indica se o paciente tem a doença ou não

# Modelos de classificadores

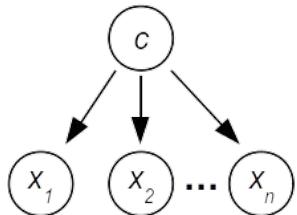
Existem diversas famílias de técnicas para classificação:

- Indução de regras
- Redes Bayesianas
- Redes Neurais
- Árvores de decisão
- KNN (k-Nearest Neighbors)
- SVM (support vector machine)
- Comitê de classificadores
- *Deep networks*



# Redes Bayesianas (Naïve Bayes)

O algoritmo **Naive Bayes** é uma rede Bayesiana aplicável em problema de classificação que possui a premissa ingênua (*naïve*) de independência completa dos nós filhos ( $X = \{x_1, \dots, x_n\}$ ) que representam as features dado um nó pai que representa a classe ( $Y = c$ ).



The formula for the posterior probability is shown:  $P(c | x) = \frac{P(x | c)P(c)}{P(x)}$ . The components are labeled as follows:

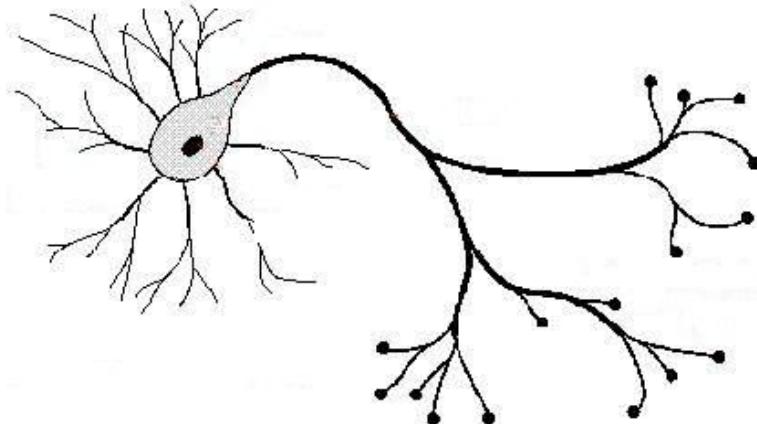
- Likelihood:  $P(x | c)$
- Class Prior Probability:  $P(c)$
- Posterior Probability:  $P(c | x)$
- Predictor Prior Probability:  $P(x)$

$$P(c | X) = P(x_1 | c) \times P(x_2 | c) \times \dots \times P(x_n | c) \times P(c)$$

# Redes Neurais

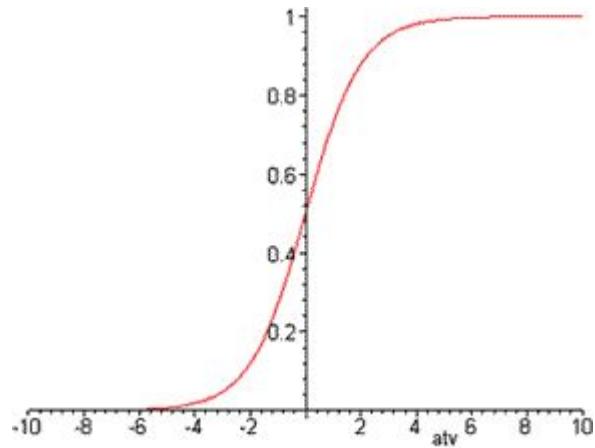
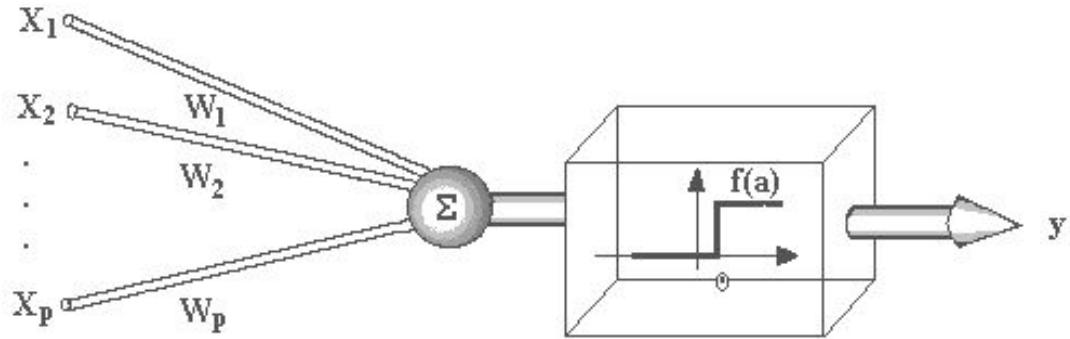
---

Conjunto de dispositivos que se intercomunicam, formando uma teia ou rede, possibilitando troca de informação entre si.



# Redes Neurais

Modelo de McCulloch-Pitts:

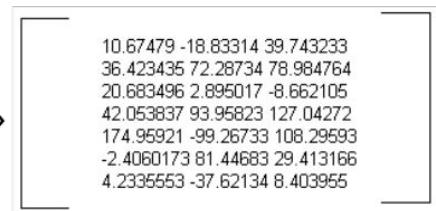
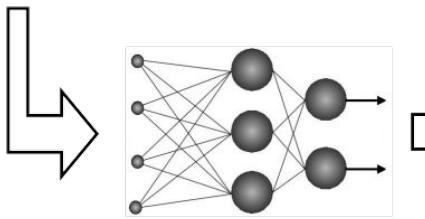


$$f(v) = \frac{1}{1+e^{-av}}$$

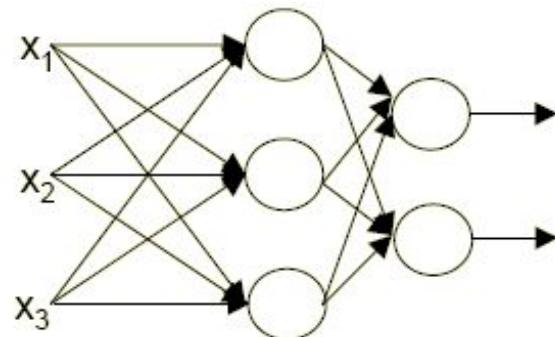
# Redes Neurais



- Segmentar a imagem
- Treinar a rede neural para identificar padrões

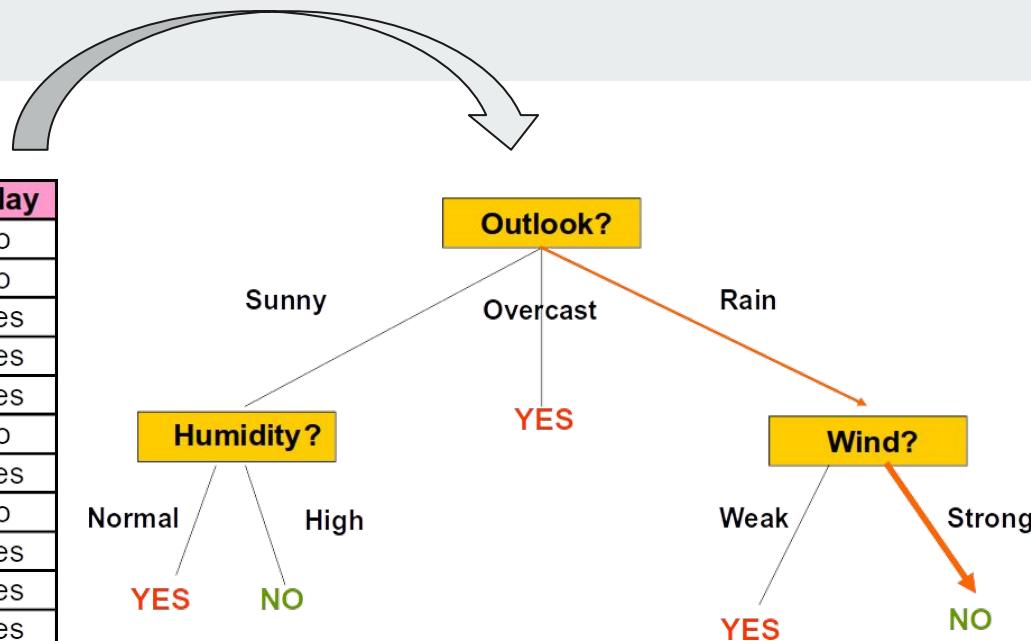


## Rede neural:

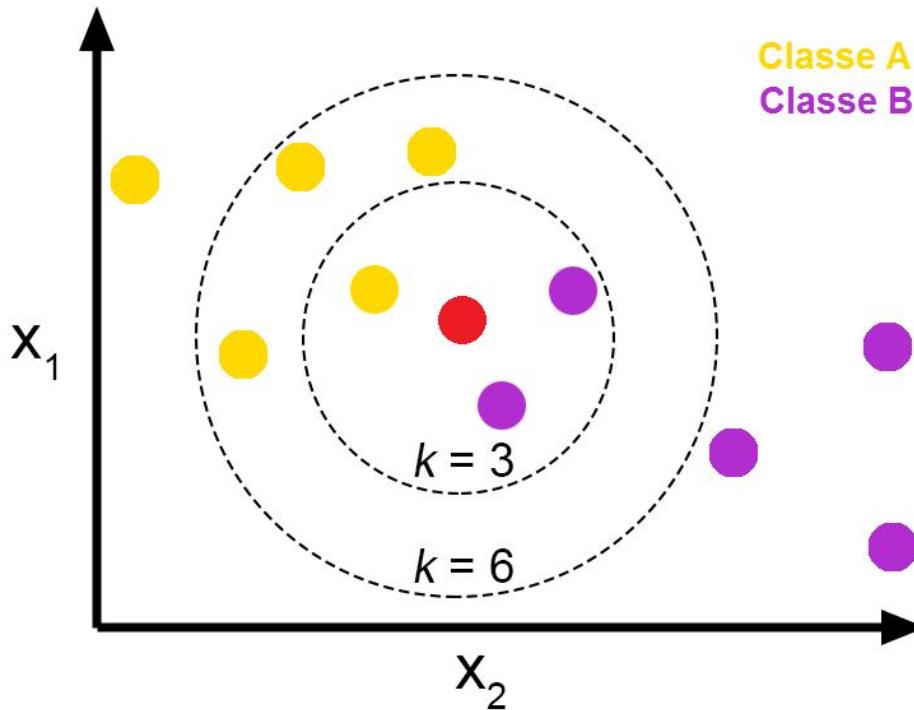


# Árvores de decisão

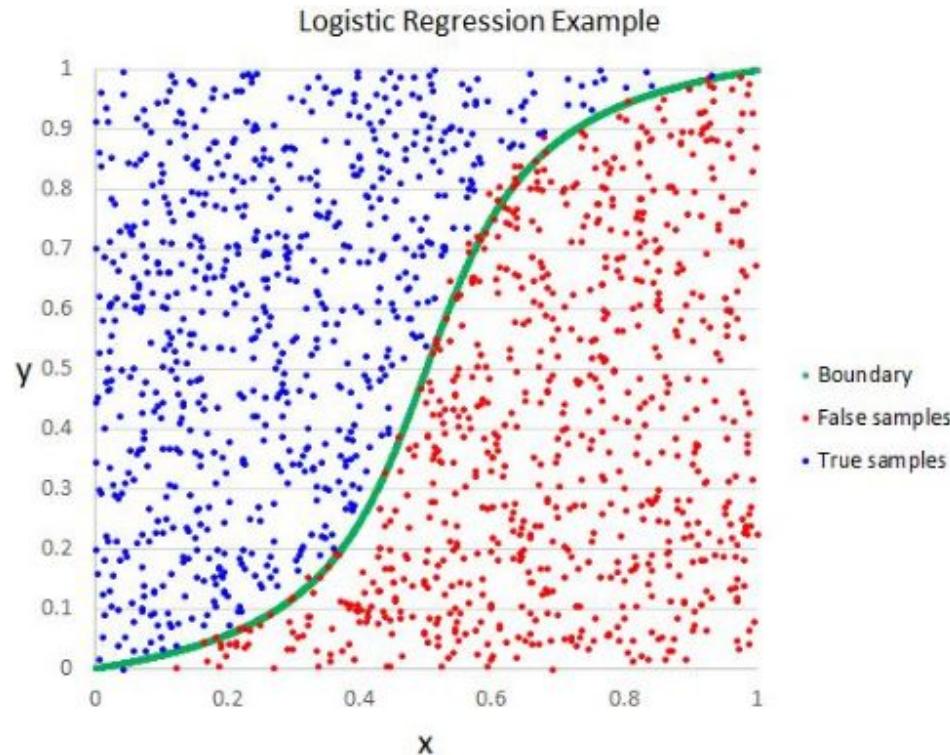
Instância	Outlook	Temperature	Humidity	Wind	Play
D1	sunny	hot	high	weak	no
D2	sunny	hot	high	strong	no
D3	overcast	hot	high	weak	yes
D4	rain	mild	high	weak	yes
D5	rain	cool	normal	weak	yes
D6	rain	cool	normal	strong	no
D7	overcast	cool	normal	strong	yes
D8	sunny	mild	high	weak	no
D9	sunny	cool	normal	weak	yes
D10	rain	mild	normal	weak	yes
D11	sunny	mild	normal	strong	yes
D12	overcast	mild	high	strong	yes
D13	overcast	hot	normal	weak	yes
D14	rain	mild	high	strong	no



# K-Nearest Neighbors

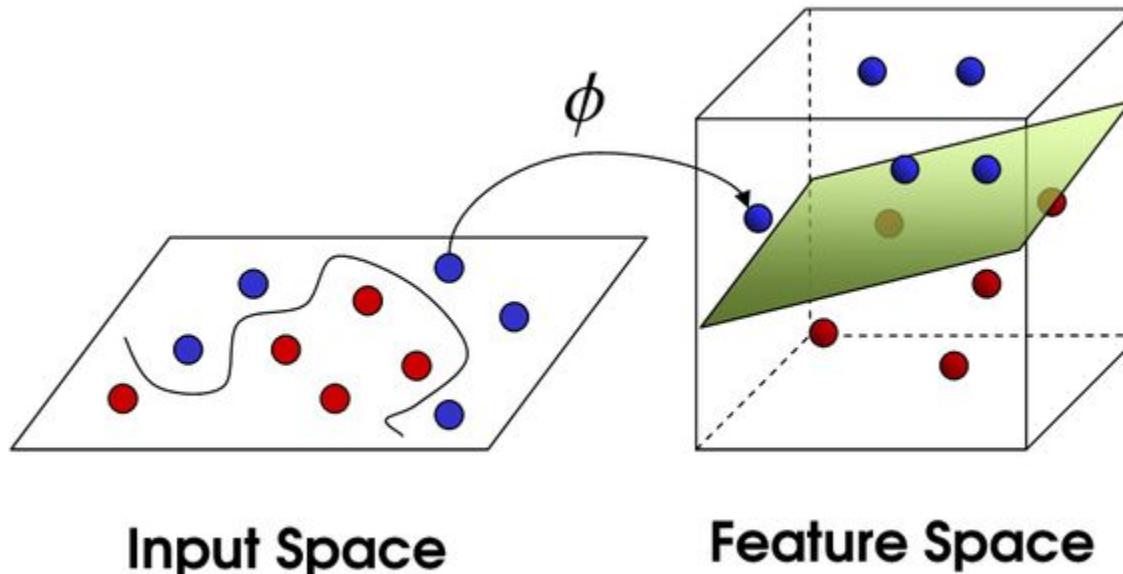


# Logistic Regression



# Support Vector Machine

---

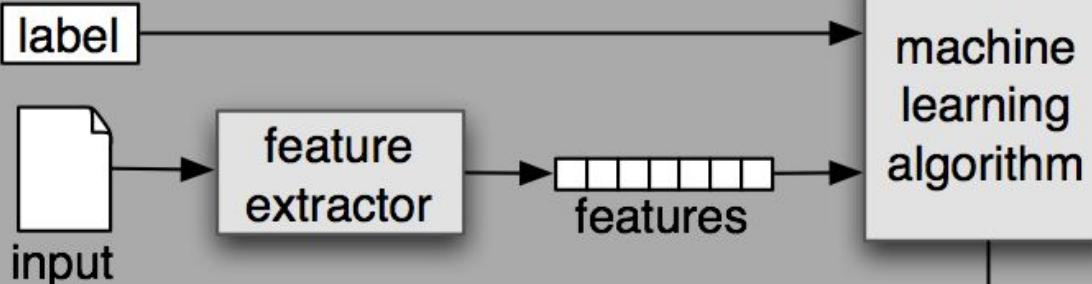


**Input Space**

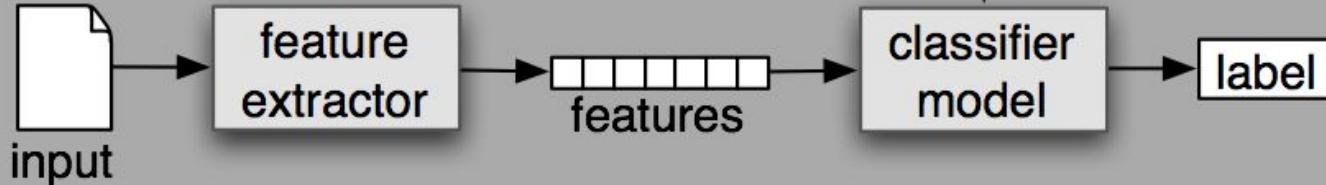
**Feature Space**

# Arquitetura básica de um classificador de texto

(a) Training



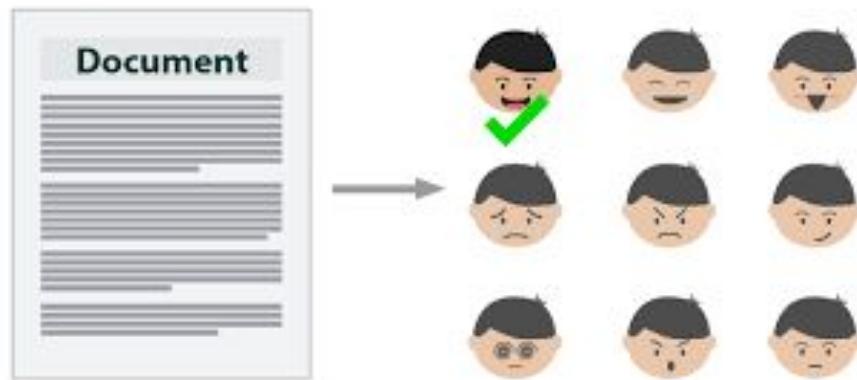
(b) Prediction



# Análise de sentimento

---

Análise de sentimento é a aplicação de rotinas de processamento de linguagem natural para a identificação e extração de informações subjetivas (sentimentos) de textos.



# Análise de sentimento

---

## Variações:

- Avaliações POSITIVAS ou NEGATIVAS
- Avaliações POSITIVAS, NEGATIVAS ou NEUTRAS
- Graus de satisfação sobre um determinado assunto

Também conhecida como **mineração de opinião**.

Muito usado por empresas para identificar reações emocionais da opinião pública na Web.



# Análise de sentimento

---

O que veremos aqui:

- Análise de sentimento por classificação textual

Mas o que é também interessante explorar:

- Nem todas as opiniões têm o mesmo peso
- Opinião de uma celebridade tem mais peso do que a de um professor.
- Usar *análise de redes sociais* para incluir modelos de influência

# Análise de sentimento

---

## Aplicações:

- Monitoramento de redes sociais
- Monitoração de percepção de marca
- Suporte ao usuário
- Análise de feedbacks de usuários
- Pesquisa de mercado

# Análise de sentimento

---

## Análise de sentimentos SEM aprendizagem de máquina



Classificar sentimento de texto com base na *polaridade de palavras*.

# Análise de sentimento com classificação textual

---

<Prática>