



# Aprendizado de Máquina

---

Fabício Olivetti de França

Universidade Federal do ABC

## 1. Introdução

# Introdução

---

- Extração de conhecimento.
- Automatização de tarefas.
- Tomada de Decisões.

Extração de informações *escondidas* a partir de dados coletados.

- Padrões de comportamento de usuários de uma rede social.
- Detecção de Fraudes.
- Que sintomas caracterizam determinada doença?

Tarefas que necessitam de muita mão de obra ou são inviáveis de serem feitas por humanos.

- Estacionar um carro sem motorista.
- Classificar tweets em críticas ou elogios.
- Estimar limites de crédito para clientes de banco.

Automatizar o processo de tomar decisões acertadas e/ou que maximizam uma certa função de recompensa.

- Próxima jogada no Xadrez.
- Quais ações investir.
- Detecção de possíveis ameaças em vôos.

O Aprendizado de Máquina consiste em um modelo matemático representável por uma função  $f(x)$  que transforma as entradas sensoriais do problema em uma resposta.

As entradas podem ser:

- Uma sequência de texto.
- Uma imagem.
- Resultados de exames médicos.
- Medições de sensores de um robô.
- Estado atual do tabuleiro de Xadrez.



O Aprendizado de Máquina consiste em um modelo matemático representável por uma função  $f(x)$  que transforma as entradas sensoriais do problema em uma resposta.

A saída seria:

- Se é uma crítica ou um elogio.
- A que grupo de imagens ela pertence.
- Possíveis doenças do paciente.
- A próxima ação do robô.
- A próxima jogada.

Existem três tipos principais de Aprendizado de Máquina:

- Aprendizado Supervisionado
- Aprendizado Não-Supervisionado
- Aprendizado por Reforço

Além desses existem outras formas que não serão cobertas nesse curso por limitações de tempo.

Quando o objetivo do aprendizado está bem definido:

- Quero saber se o texto faz uma crítica ou um elogio.
- Quero saber o preço de um imóvel.
- Quero saber o diagnóstico entre as doenças conhecidas.

Dividido entre problema de **regressão** e **classificação**.

Na regressão desejamos obter uma função  $f : \mathbb{R}^d \rightarrow \mathbb{R}$ , ou seja, dado um conjunto de variáveis de medição, desejo obter uma saída de valor real.

- **Modelo matemático descritivo:** o que faz com que um dado sistema se comporte de certa maneira? Se eu alterar um valor de entrada, o que acontecerá?
- **Entender fenômenos:** quais leis governam um certo fenômeno observável?

Na classificação buscamos por  $f : \mathbb{R}^d \rightarrow C$ , ou seja, dado um conjunto de variáveis de medição, desejo obter um valor do conjunto finito  $C$ .

- **Rotular dados em classes:** em qual categoria esse objeto em particular pertence?
- **Automatizar tarefa:** que ação, dentre as possíveis ações, devo realizar nesse instante?

Quando queremos entender melhor nossos dados ou agrupá-los.

- Quero transformar meus atributos em combinações de variáveis que façam mais sentido.
- Quero visualizar meus dados  $n$ -dimensionais em um plano  $2D$ .
- Quero segmentar meus clientes em grupos de perfis similares.

Compreende as tarefas de **agrupamento**, **aprendizado de representação**, e **redução de dimensionalidade**.

No agrupamento buscamos dividir nossos dados amostrados em diferentes grupos, similar com a tarefa de classificação desejamos uma função  $f : \mathbb{R}^d \rightarrow C$  em que  $C$  é o conjunto de possíveis grupos formados.

Para o aprendizado de representação, desejamos encontrar uma função  $f : \mathbb{R}^d \rightarrow \mathbb{R}^{d'}$  que transforma nossas variáveis.

As variáveis transformadas pertence a outro espaço de dimensão  $d' \neq d$  que faz mais sentido para as tarefas de classificação ou regressão.



Na redução de dimensionalidade, queremos encontrar  $f : \mathbb{R}^d \rightarrow \mathbb{R}^{d'}$ , mas com  $d' < d$ .

Dessa forma os dados podem ser representados e classificados em um espaço de dimensão menor, isso tanto pode ajudar na tarefa de aprendizado supervisionado como também pode nos permitir visualizar melhor nossos dados para os casos de  $d' \leq 3$ .

- **Descobrir padrões similares:** que conjunto de objetos se comportam de forma similar?
- **Descobrir pontos representativos:** que objetos são bons representantes de um certo grupo?
- **Descobrir como os dados são gerados:** qual a função que gera esses dados?

- **Transformar atributos para a máquina:** como os atributos originais podem ser combinados para facilitar tarefas de Aprendizado de Máquina?
- **Transformar atributos para o humano:** como os atributos podem ser combinados e transformados para ajudar o entendimento pelo ser humano?

- **Visualizar meus dados:** como projetar os dados em um plano  $2D$ ?
- **Facilitar o trabalho de aprendizado:** eliminar variáveis ruidosas, combinar variáveis que agregam informação.

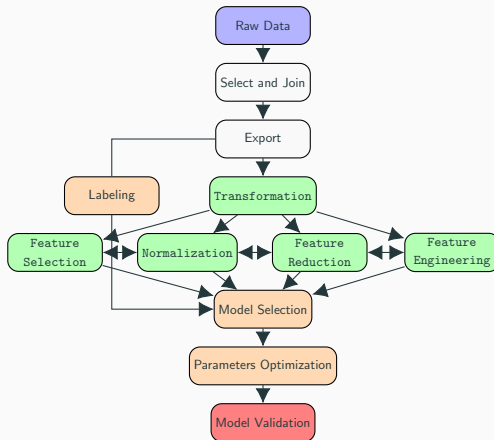
Quando não sabemos exatamente a resposta correta que o sistema deve fazer ou a resposta só pode ser mensurada após uma sequência de respostas do sistema.

- Dadas as últimas jogadas, qual a próxima peça de Xadrez que devo mover?
- Qual sequência de ações o robô deve fazer para desarmar uma bomba?

A pesquisa, desenvolvimento e aplicação de tarefas de Aprendizado de Máquina seguem um *pipeline* comum para todas as tarefas:

- Obtenção de dados.
- Tratamento dos atributos para uma forma computacional.
- Rotulação dos dados.
- Escolha do modelo da tarefa e modelo de aprendizado.
- Otimização dos parâmetros do algoritmo.
- Validação.
- Interpretação.

# Pipeline



Coleta dos dados da tarefa a ser realizada:

- Transações de compras na internet
- Dados de navegação (Google, Facebook)
- Dados de satélite
- Experimentos com genes



Maior parte dos algoritmos de Aprendizado de Máquina assumem uma entrada de dados bem definida.

Geralmente um vetor multidimensional.

Muitas bases de dados de interesse não possuem esse formato:

- Textos
- Imagens
- Som
- Vídeos

Necessidade de transformar tipos desestruturados em uma estrutura vetorial.

Cuidado para não perder informação no meio do caminho!

Nas tarefas de Aprendizado Supervisionado, é necessário rotular uma parte dos dados obtidos para a construção de um modelo de classificação/regressão baseado em exemplos.

Regra dos 10 para estimar a quantidade mínima de dados rotulados necessário para um bom modelo de predição: número de atributos multiplicados por 10.

Idealmente teremos um número suficiente para entrar exemplos de combinações diferentes de valores de atributos.

De posse dos dados e com eles já tratados, precisamos escolher o que queremos extrair de informação.

Após essa escolha, precisamos escolher o algoritmo a ser utilizado <sup>1</sup>.

---

<sup>1</sup>Existem algoritmos para escolher algoritmos!

Uma vez que o algoritmo foi escolhido, dependendo da escolha existem diversos parâmetros que o usuário deve ajustar para obter o melhor resultado.

Em muitos desses algoritmos a escolha desses parâmetros pode fazer muita diferença <sup>2</sup>.

---

<sup>2</sup>Existem algoritmos para escolher esses parâmetros!

Uma vez que o modelo foi construído, é necessário verificar se ele está respondendo corretamente ou qual a taxa de acerto que obtemos com tal modelo.

Para isso separamos parte da nossa base rotulada e calculamos o erro. O erro deve ser comparado com um *baseline* para ter certeza que o modelo funciona!

Em muitas tarefas pode ser necessário interpretar ou entender o que o modelo aprendeu:

- O carro autônomo não vai entrar em modo GTA?
- Por que ele recomendou determinada vaga de emprego para uma pessoa?
- Que tipo de produto maximizará minhas vendas?

Essa é uma área de pesquisa atual e de grande interesse das empresas que trabalham com Inteligência Artificial.



A área de Aprendizado de Máquina é de grande interesse prático para empresas de diversas áreas.

Embora existam diversos algoritmos que apresentam boa performance, o recente aumento na capacidade de obtenção de dados requer aprimoramento destes.

Apesar de esses algoritmos existirem em diversos pacotes prontos para serem utilizados, o passo-a-passo desde a obtenção dos dados até os resultados não é trivial.