# Machine Learning: Computadores e sua Capacidade de Aprender com os Dados

*por Alysson Machado*

- Vivemos em uma era **Big Data**, em que o meio digital está repleto de dados úteis disponíveis para análise;
- A ideia principal de se usar **Machine Learning** é transformar a grande quantidade de dados que existem atualmente em conhecimento;
- O principal objetivo de qualquer algoritmo de Machine Learning é encontrar **padrões** locais e fazer **predições** sobre estados futuros;

## Os Três Diferentes tipos de Aprendizado de Máquina

- **Aprendizado Supervisionado:**
1. Dados rotulados;
2. Feedback direto;
3. Prever resultado/futuro;

- **Aprendizado por reforço:**
1. Processo de decisão;
2. Sistema de recompensa;
3. Aprender séries de ações;

- **Aprendizado não Supervisionado:**
1. Sem rotulações;
2. Sem feedback;
3. Encontrar dados em estruturas ocultas;

# Fazendo Previsões com o Aprendizado Supervisionado

O principal objetivo do **aprendizado supervisionado** é ensinar um modelo a partir de dados de treinamento rotulados que nos permitem fazer previsões sobre o estado futuro.

Nele, é considerado uma base de dados usada para treinamento (**Modelo de Treinamento**), tais quais possuirá atributos **previsores** e, a partir disso, é possível obter um algoritmo treinado a responder com atributos **classe** (esses que serão a previsão do algoritmo). Entretanto, para supervisionar tais ações, é necessário ter um **Modelo de Teste** para avaliar a capacidade preditiva do algoritmo.

As principais categorias de algoritmos que utilizam aprendizado supervisionado é a **classificação** e a **regressão**.

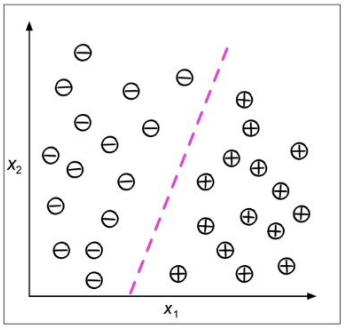
## Classificação para Prever Classes

**Classificação** é uma subcategoria de aprendizado supervisionado, em que o objetivo principal é classificar os atributos em determinadas categorias de novas instâncias, com base em observações anteriores utilizando os atributos previsores. Esses atributos classe são valores discretos e não ordenados que podem ser entendidos como associações de grupo das instâncias.

Um exemplo adequado seria um avaliador de e-mails spam e não spam, em que a partir de uma série de condições esse e-mail possa ser classificado em alguma das duas possíveis classes.

Outra forma de fazer classificação é considerar **multiclasses**, em que é possível obter uma gama de atributos **meta classe**. Como exemplo teríamos um algoritmo que classifica letras do alfabeto e números escritos manualmente para caracteres digitais. Tal mecanismo é usado com grande intensidade em diversas tecnologias da google e em mais uma série de softwares e aplicativos disponíveis na área de ensino.

Além disso, classificar elementos não é algo que seja 100% certeiro, qualquer algoritmo feito em Machine Learning possuirá uma certa **precisão** de acertos e erros. O principal objetivo de um desenvolvedor na área de machine learning é, muito além de obter a predição, garantir que ela será a mais adequada para garantir uma capacidade preditiva alta.




Ao lado, é possível analisar graficamente como ocorre a classificação. No exemplo, foram separados dois elementos: um com sinal negativo e outro com sinal positivo. As classes seriam **X1** e **X2**.

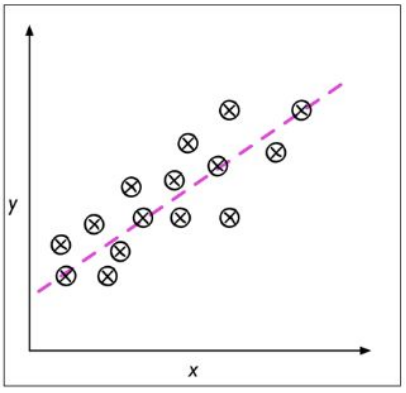
## Regressão para Prever Resultados Contínuos

Um segundo tipo de aprendizado supervisionado é a previsão de resultados contínuos, também chamada de **análise de regressão**.

Nesse algoritmo, será recebido várias variáveis preditivas e uma variável de resposta contínua. O objetivo do algoritmo que trabalha com análise de regressão é encontrar uma relação entre os dados e o valor contínuo.

Um exemplo adequado é realizar a estimativas de notas para estudantes em um teste de matemática. Para isso, poderia ser utilizado uma base de dados que relacionasse o tempo disponível para se dedicar ao estudo de algum teste e os possíveis resultados obtidos. Desse modo, poderíamos usá-los como dados de treinamento para aprender um modelo que possuirá o tempo de estudo para prever as notas dos futuros alunos que planejam fazer esse teste.

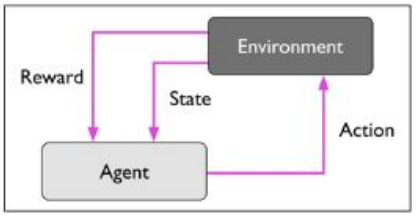
***O termo regressão foi elaborado por Francis Galton em seu artigo "Regressão à mediocridade na estatura hereditária em 1886". Galton descreveu o fenômeno biológico de que a variação de altura em uma população não aumenta ao longo do tempo. Ele observou que a altura dos pais não é repassada aos filhos, mas, em vez disso, a altura dos filhos está regredindo em relação à média da população.***




Ao lado é possível analisar o método de regressão graficamente, em que baseado em variáveis preditoras **x** e variáveis de resposta **y**, é ajustada uma linha reta que minimiza a distância quadrática média (**problema de mínimos quadráticos**) entre os elementos. Desse modo, a partir da intercepção e inclinação obtidas com o modelo de treinamento, é possível prever a variável de resultados de dados novos.

# Resolvendo Problemas com o Aprendizado por Reforço

Um outro tipo de aprendizado de máquina é o aprendizado por reforço. O objetivo desse aprendizado é desenvolver um sistema (**agente**) que melhore seu desempenho com base na sua **interação com o ambiente** e nos **sinais de recompensa** que o ambiente oferece.

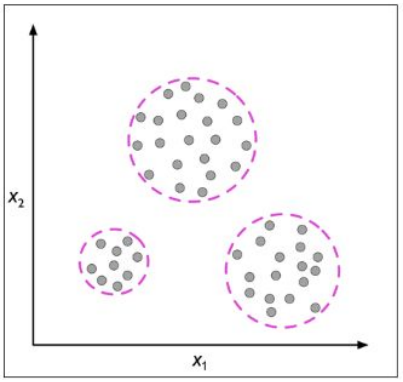



Através de sua interação com o ambiente, um agente pode usar o aprendizado por reforço para aprender uma série de ações que maximizam essa recompensa por meio de uma abordagem exploratória de **tentativa e erro** ou **planejamento deliberativo**.

Cada estado pode ser associado a uma recompensa positiva ou negativa, e uma recompensa pode ser definida como a realização de um objetivo geral, como vencer ou perder um jogo de xadrez, por exemplo.

# Encontrando Subgrupos por Agrupamento

O **agrupamento** é uma técnica de análise de dados exploratória que nos permite organizar uma pilha de informações em subgrupos significativos (também chamados **clusters**) sem ter conhecimento prévio de sua participação no grupo. Cada cluster que surge durante a análise define um grupo de objetos que compartilham um certo grau de similaridade, mas são mais diferentes dos objetos de outros clusters.

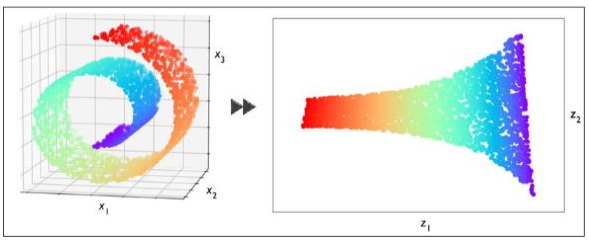



Ao lado, é possível observar um gráfico explanatório sobre como funciona o agrupamento. Nele, os elementos foram agrupados em três principais grupos, cada elemento das clusters compartilham um grau de semelhança, mas as clusters são totalmente independentes uns dos outros.

# Redução da Dimensionalidade para Compressão dos Dados

Outro subcampo da aprendizagem não supervisionado é a **redução da dimensionalidade**. Frequentemente, estamos trabalhando com dados de alta dimensionalidade que podem representar um desafio para o espaço de armazenamento limitado e o desempenho computacional dos algoritmos de aprendizado de máquina. A redução não supervisionada da dimensionalidade é uma abordagem comumente usada no **pré-processamento** de recursos para **remover o ruído dos dados**, o que também pode degradar o desempenho preditivo de certos algoritmos ao **compactar** os dados em um subespaço dimensional menor, mantendo a maioria das informações relevantes.

Às vezes, a redução de dimensionalidade também pode ser útil para a visualização de dados, por exemplo, um conjunto de recursos de alta dimensão pode ser projetado em espaços de recursos de uma, duas ou três dimensões, para visualizá-lo através de gráficos de dispersão e histogramas em 3D ou 2D.




A figura ao lado mostra um exemplo em que a redução não linear de dimensionalidade foi aplicada para compactar um rolo suíço 3D em um novo subespaço de recurso 2D.

# Terminologia Básica e Notações

Todos os dados são armazenados em matrizes, em que as colunas representam os campos de dados e as linhas os diversos registros para cada elemento. **Álgebra Linear** é uma disciplina muito importante na área de Machine Learning e Inteligência Artificial no geral.

- Desse modo, teremos a seguinte notação para os atributos previsores:

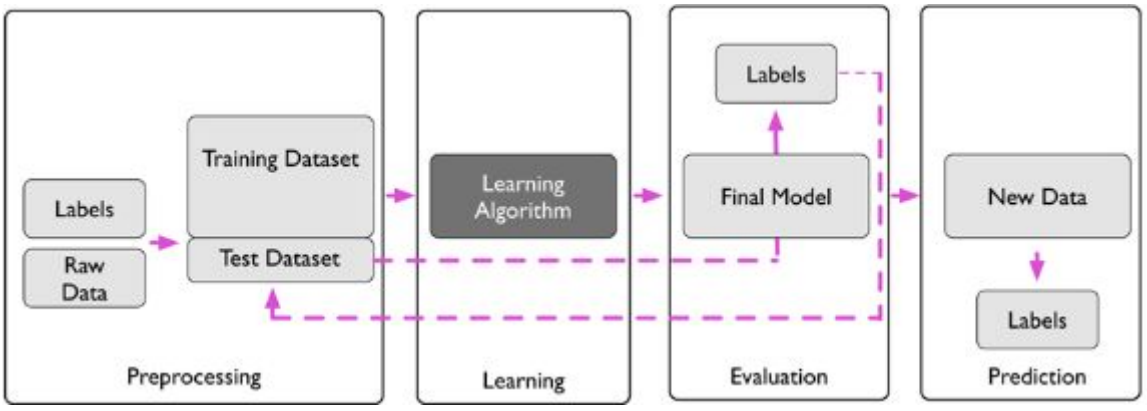
$$\begin{bmatrix} x_1^{(1)} & x_2^{(1)} & x_3^{(1)} & x_4^{(1)} \\ x_1^{(2)} & x_2^{(2)} & x_3^{(2)} & x_4^{(2)} \\ \vdots & \vdots & \vdots & \vdots \\ x_1^{(150)} & x_2^{(150)} & x_3^{(150)} & x_4^{(150)} \end{bmatrix}$$

- e a seguinte notação para as classes:

$$\boldsymbol{y} = \begin{bmatrix} y^{(1)} \\ \dots \\ y^{(150)} \end{bmatrix}$$

# Roteiro para Construção de Algoritmos em Machine Learning

- O ciclo geral de qualquer algoritmo em Machine Learning está representado abaixo:




Todo processo começa no **pré-processamento** da base de dados que será utilizada para montar o algoritmo. Nessa etapa, é importante fazer a correção de alguns problemas nos dados, tais como: **remoção de valores faltantes**, **correção de valores inconsistentes**, **redução da dimensionalidade** e de **ruídos** são os mais recorrentes.

Após concluir a etapa do pré-processamento, é necessário delimitar corretamente quais são os atributos **previsores** e quai são os atributos **classe**, assim como dividir a base de dados em um **modelo de treinamento** e outro de **teste**. O modelo de treinamento será crucial para ensinar o algoritmo as possíveis previsões feitas com base em determinadas entradas.

Com o modelo final do algoritmo pronto, será necessário utilizar o modelo de teste dos dados para verificar a capacidade de predição do algoritmo. Será usado os previsores desse modelo para obter saídas de dados por parte do algoritmo. Logo após, é necessário comparar esses dados com as classes do modelo de teste, visando analisar a precisão de predição do algoritmo.

Caso a precisão seja boa, o algoritmo está pronto, mas se esse não for o caso, diversas mudanças devem ser feitas a fim de obter o melhor resultado possível (as mudanças podem ser na parte de pré-processamento, estrutura do algoritmo ou escolha de novos métodos).

# Referência


- RASCHKA, Sebastian; MIRJALILI, Vahid. **Python Machine Learning, 2nd Ed**. Packt Publishing, 2017.