Introdução a Aprendizagem De Máquina

Pós-graduação em Ciência de Dados e Machine Learning
Módulo 3 - Data Mining e Machine Learning

Professor Msc. Ricardo José Menezes Maia

Abordagem no curso

Para cada tópico de Machine Learning

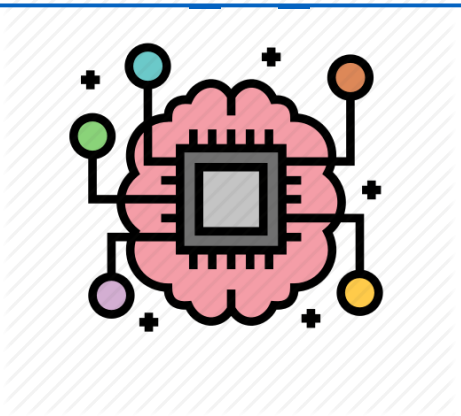
- Leitura recomendada
  - Utilizaremos Introduction to Statical Learning de Gareth James
- Breve resumo da teoria
- Demonstrações do algoritmo no Python
- Projeto
- Solução do Projeto
- Exercícios
- Avaliação final será um projeto proposto pelo aluno para resolver algum problema utilizando os algoritmos abordados no curso.

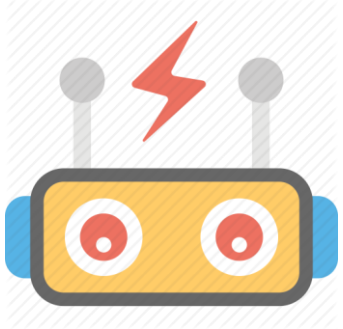
Arthur Sammuel (1959)

“Aprendizado de Máquina é o campo de estudo que permite o aprendizado à computadores sem que estes sejam explicitamente programados.”
https://pt.wikipedia.org/wiki/Arthur_Samuel

Tom Mitchel (1977)

“É dito que um programa de computador aprende com a experiência E a respeito de uma tarefa T e com uma performance P, se sua performance em T, medida por P, melhora através da experiência E”.
https://en.wikipedia.org/wiki/Tom_M._Mitchell





O que é?

- Machine Learning é um método de análise de dados que automatiza o processo de criação de modelos.
- Usando algoritmos que iterativamente aprendem dos dados, Machine Learning permite que computadores encontrem padrões escolhidos nos dados sem terem sido programados para isso.

O que é?

- Também conhecido como aprendizado automático explora o estudo e construção de algoritmos que podem aprender e fazer previsões sobre dados. Esses algoritmos operam construindo um modelo a partir de entradas amostrais com a finalidade de fazer previsões ou decisões guiadas pelos dados ao invés de simplesmente seguindo inflexíveis e estáticas instruções programadas. Enquanto que na inteligência artificial existem dois tipos de raciocínio (o indutivo, que extrai regras e padrões de grandes conjuntos de dados, e o dedutivo).

- O aprendizado de máquina só se preocupa com o **indutivo**. A indução é uma forma de inferência lógica que permite obter conclusões a partir de um conjunto de exemplos.

- Relembrando que um dos exemplos mais clássicos de raciocínio dedutivo é:
  - “Todos os homens são mortais. Sócrates é um homem. Portanto, Sócrates é mortal…”.
  - A conclusão deste raciocínio é: Sócrates é mortal. E ela deriva de duas premissas: Todos os homens são mortais e Sócrates é um homem. Dessa forma, está sendo aplicada a lei da lógica de predicados

O que é?

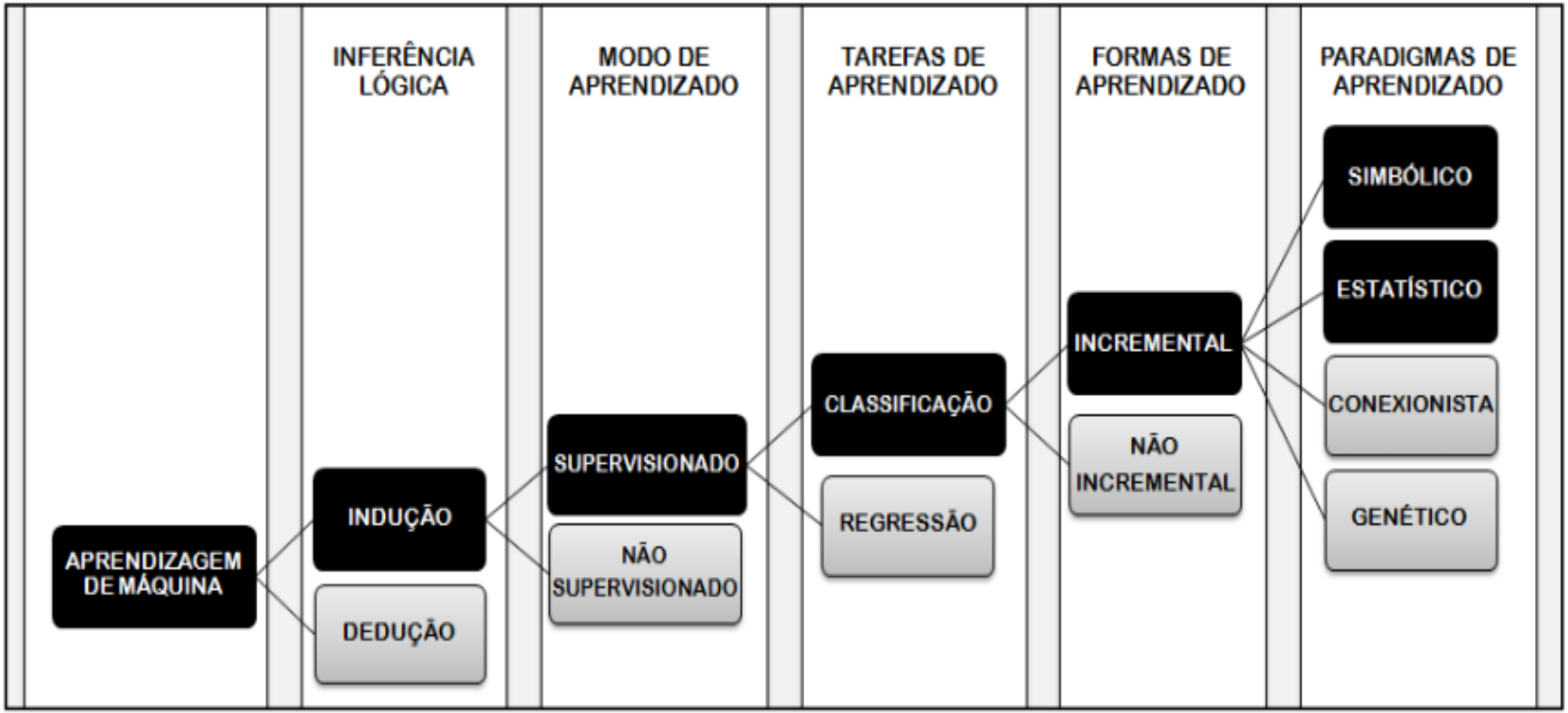
- Fundamentalmente, Machine Learning (ou Aprendizado de Máquina) é a utilização de algoritmos para extrair informações de dados brutos e representá-los através de algum tipo de modelo matemático. Usamos então este modelo para fazer inferências (previsões) a partir de outros conjuntos de dados. E existem muitos algoritmos que permitem fazer isso, cabendo a um Cientista de Dados escolher o algoritmo que melhor se encaixa em cada tipo de problema a resolver.
- Aprendizado é um conceito difícil de ser definido, mas também de grande importância para que possamos partir para a construção de sistemas inteligentes dotados da capacidade de aprendizado. Se eu perguntar a 10 pessoas qual é o conceito de aprendizado, terei 10 respostas diferentes. Pare por 1 minuto e pergunte a si mesmo o que é aprendizado. Vamos tentar definir aqui o que entendemos por aprendizado.
- Aprendizado é a capacidade de se adaptar, modificar e melhorar seu comportamento e suas respostas, sendo portanto uma das propriedades mais importantes dos seres ditos inteligentes, sejam eles humanos ou não.
- Ou seja, estamos tentando reproduzir nas máquinas, o mesmo processo de aprendizagem dos seres humanos e fazemos isso através de algoritmos de Machine Learning que em última instância nada mais são do que Matemática e Estatística.

• Para que é usado ?

- Para que é usado ?

- Detecção de fraudes
- Pequisas na Web
- Anúncios automáticos na internet
- Predição de falhas em equipamentos
- Modelos de precificação de ativos financeiros
- Detecção de invasores de rede
- Sistemas de recomendação (Netflix, Spotify)
- Segmentação de clientes
- Analise de sentimentos em textos
- Reconhecimento de padrões em Imagens
- Filtro de spam em emails.

Conceitos


APRENDIZAGEM DE MÁQUINA
INFERÊNCIA LÓGICA
INDUÇÃO
DEDUÇÃO
MODO DE APRENDIZADO
SUPERVISIONADO
NÃO SUPERVISIONADO
TAREFAS DE APRENDIZADO
CLASSIFICAÇÃO
REGRESSÃO
FORMAS DE APRENDIZADO
INCREMENTAL
NÃO INCREMENTAL
PARADIGMAS DE APRENDIZADO
SIMBÓLICO
ESTATÍSTICO
CONEXIONISTA
GENÉTICO

Conceitos

•Indução: é uma forma de inferência lógica que permite obter conclusões a partir de um conjunto de exemplos.

•Dedução: humanos usam raciocínio indutivo para deduzir nova informação a partir de informação relacionada logicamente. Prolog Lisp

• Aprendizado supervisionado: As observações no conjunto de treinamento são acompanhadas por “labels” indicando a classe que ela pertence. Novas ocorrências são classificadas com base no conjunto de treinamento.

- •Classificação: rótulos para valores discretos
- •Regressão: Rótulos para valores contínuos

•Aprendizado não-supervisionado(clusterização): o indutor analisa os exemplos e tenta determinar se alguns deles podem ser agrupados de alguma maneira, formando agrupamento ou clusters.

•Incremental é um método de aprendizagem de máquina , em que os dados de entrada são continuamente usados para estender o conhecimento do modelo existente ou seja, para treinar ainda mais o modelo. Representa uma técnica dinâmica de aprendizado supervisionado e aprendizado não supervisionado que pode ser aplicada quando os dados de treinamento se torna disponível gradualmente ao longo do tempo ou seu tamanho está fora dos limites de memória do sistema.

•Não incremental necessita de que todos os exemplos de treinamento, simultaneamente, estejam disponíveis para que seja induzido um conceito. É vantajoso usar esses algoritmos para problemas de aprendizado onde todos os exemplos estão disponíveis e, provavelmente, não irão ocorrer mudanças.

•Simbólico: Buscam aprender construindo representações simbólicas (expressão lógica, árvore de decisão regras)

•Estatístico: Buscam métodos estatísticos (Aprendizado bayesiano)

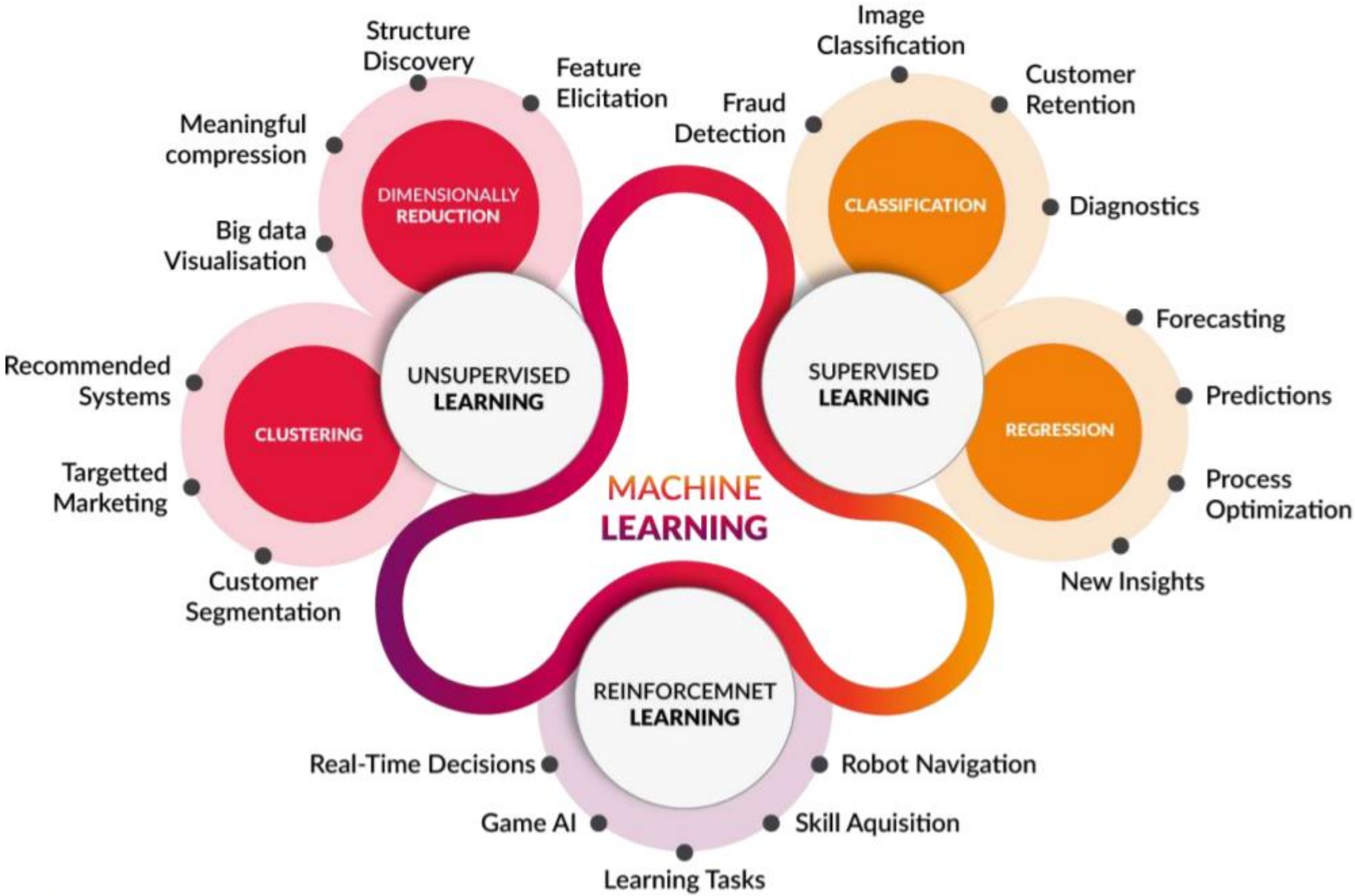
•Baseado em exemplos: **lazy** learning algorithm (Raciocínio Baseado em Casos, Nearest Neighbors)

•Conexionista: Modelo inspirado no modelo biológico do sistema nervoso (Redes Neurais)

•Evolucionistas: Teoria de Darwin (Algoritmos Genéticos)

Modos de Aprendizagem

- Aprendizagem Supervisionada
  - São apresentadas ao computador exemplos de entradas e saídas desejadas, fornecidas por um "professor". O objetivo é aprender uma regra geral que mapeia as entradas para as saídas.
  - Você tem parâmetros rotulados que são usados para construir o modelo e tentar predizer os demais rótulos, baseados em parâmetros apenas.
- Aprendizado não supervisionado
  - Nenhum tipo de etiqueta é dado ao algoritmo de aprendizado, deixando-o sozinho para encontrar estrutura nas entradas fornecidas. O aprendizado não supervisionado pode ser um objetivo em si mesmo (descobrir novos padrões nos dados) ou um meio para atingir um fim.
  - Você possui os parâmetros sem rótulos e quer encontrar subgrupos dentro de dados que possuam algum tipo de semelhança.
- Aprendizado por reforço robótica
  - Um programa de computador interage com um ambiente dinâmico, em que o programa deve desempenhar determinado objetivo (por exemplo, dirigir um veículo). É fornecido, ao programa, feedback quanto a premiações e punições, na medida em que é navegado o espaço do problema. Outro exemplo de aprendizado por reforço é aprender a jogar um determinado jogo apenas jogando contra um oponente.
  - Algoritmos que aprendem a executar ações baseados em experiências do mesmo com algum meio.


MACHINE LEARNING
UNSUPERVISED LEARNING
SUPERVISED LEARNING
REINFORCEMNET LEARNING
DIMENSIONALLY REDUCTION
CLUSTERING
CLASSIFICATION
REGRESSION
Structure Discovery
Feature Elicitation
Meaningful compression
Big data Visualisation
Recommended Systems
Targetted Marketing
Customer Segmentation
Image Classification
Customer Retention
Fraud Detection
Diagnostics
Forecasting
Predictions
Process Optimization
New Insights
Real-Time Decisions
Robot Navigation
Game AI
Skill Aquisition
Learning Tasks

Supervised Learning

Supervised Learning

Entradas e saidas

- Você tem parâmetros rotulados que são usados para construir o modelo e tentar predizer os demais rótulos, baseados em parâmetros.
- Por exemplo: você tem características técnicas de peças de equipamentos que falharam “F” e não falharam “NF” e quer predizer o comportamento das demais peças

Output é fornecido ao algoritmo

Supervised Learning

Supervised Learning

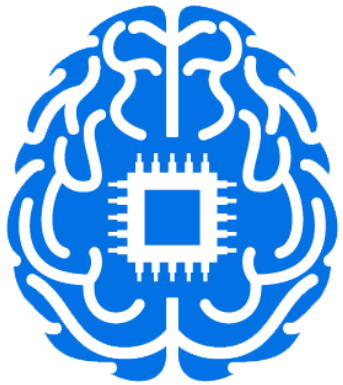
- O algoritmo de aprendizado recebe entradas com as saídas corretas e ajusta o seu modelo de forma iterativa para que o mesmo se adapte as condições apresentadas no **conjunto de dados de treino**.
- Então o algoritmo irá conferir a precisão do modelo criado usando o **conjunto de dados de teste**.

Supervised Learning

Aprendizado Supervisionado

Os **OUTPUTS** são fornecidos ao algoritmo!

O modelo aprende com estes dados.



... E fornece respostas a partir de novas entradas**.**

Supervised Learning

Classificação

Valores
Discretos

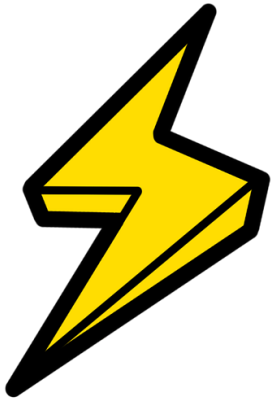
Regressão

Valores
Contínuos

Supervised Learning

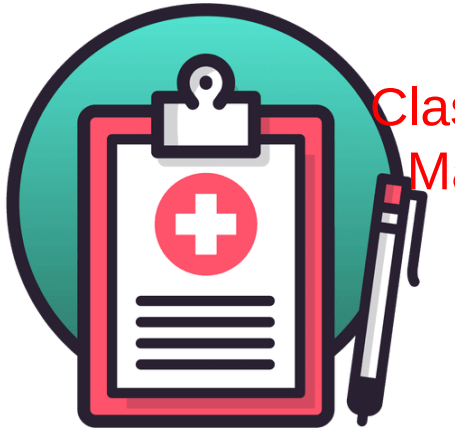
Previsão de preços de casas

Previsão de consumo de energia elétrica



Classificação de imagens através de padrões

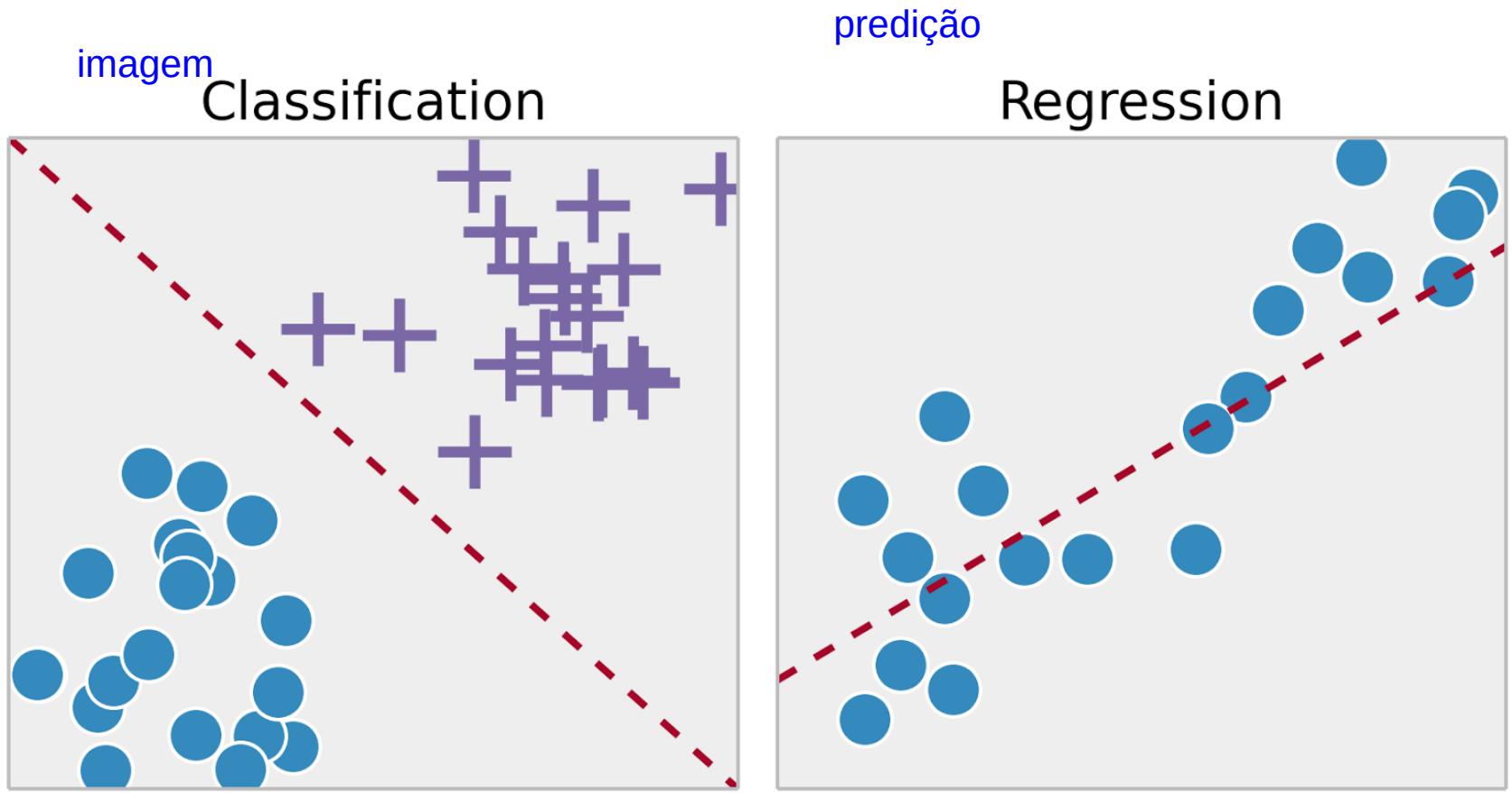
Quantos clientes irão migrar para a concorrência?



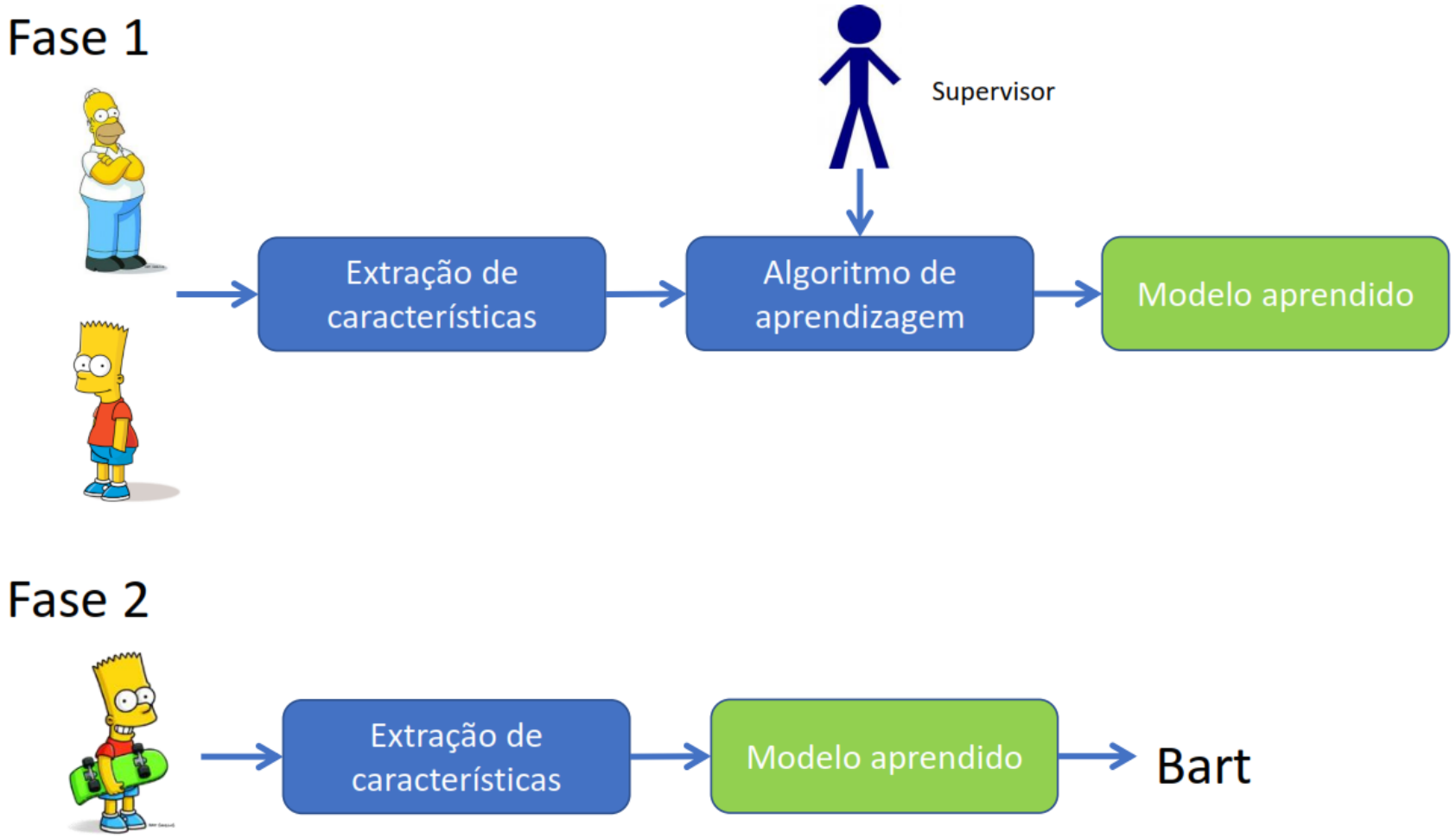
Classificar tumores como Malignos ou Benignos

Supervised Learning

Aprendizado Supervisionado



Aprendizagem supervisionada



Unsupervised Learning

Decisões similares

Unsupervised Learning

Com base no texto ele inferir qual o tribunal 1. 2. 3. Instância

- Técnicas populares incluem mapas auto-organizáveis, k-means clustering e singulares value decomposition
- Estes algoritmos também são usados para segmentar textos em tópicos, identificação de outliers em conjuntos de dados e recomendação de itens à clientes.

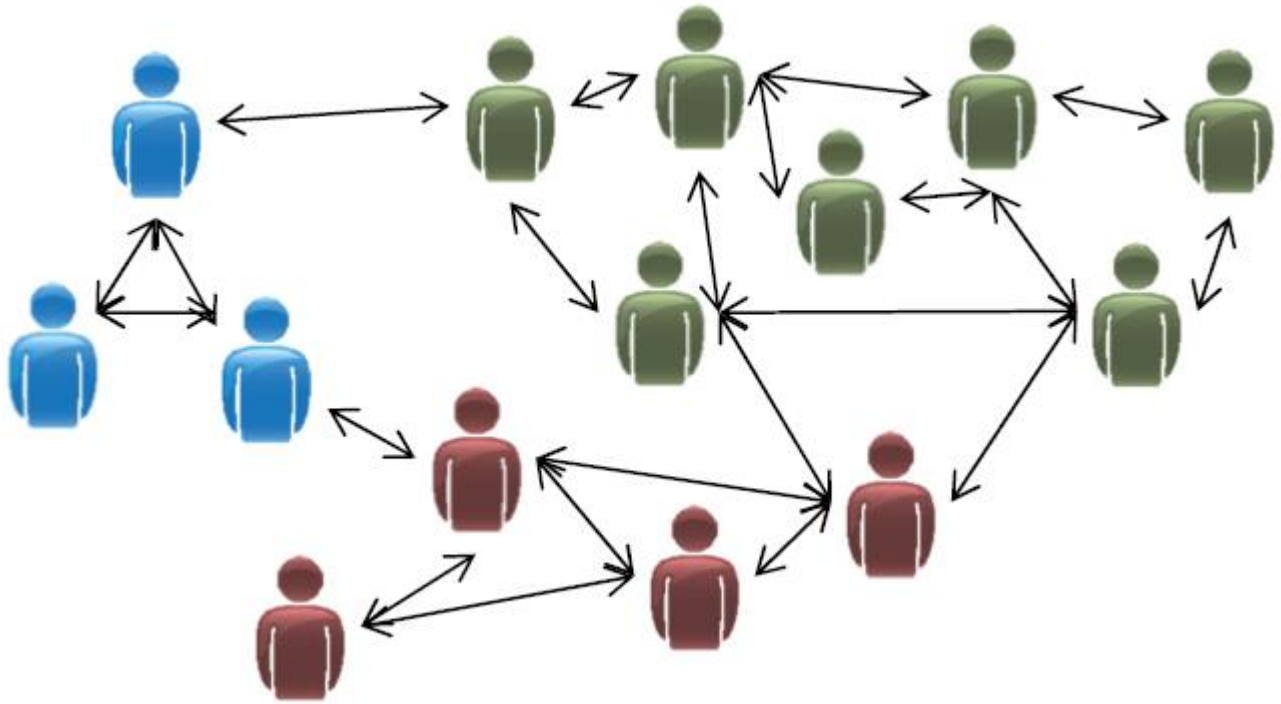
Unsupervised Learning

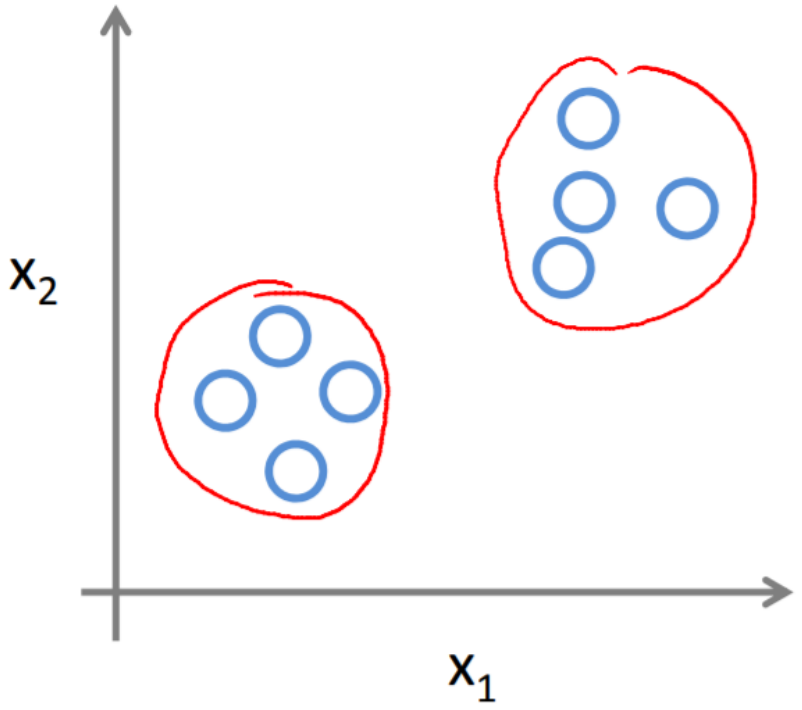
Unsupervised Learning

- É usado quando os dados não possuem classificações prévias.
- A resposta correta não é dita ao algoritmo, cabendo ele encontrar padrões nos dados e agrupá-los/classifica-los baseados em similaridades nos conjuntos de parâmetros.

Unsupervised Learning

Aprendizado Não-Supervisionado





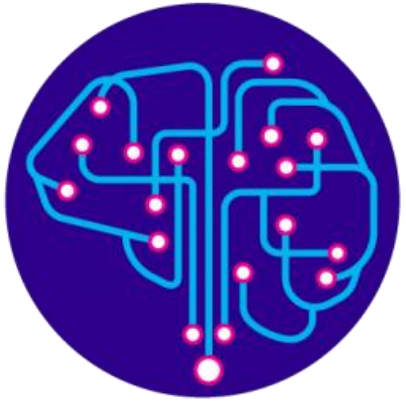
Unsupervised Learning

Aprendizado Não-Supervisionado

tst agrupou alguns termos para identificar
1. 2. 3. Instância

Os dados **não são rotulados!**

O **OUTPUT** não é fornecido ao modelo...



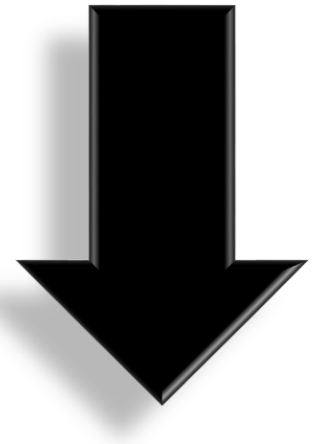
... Mesmo assim é retornada uma resposta.

Unsupervised Learning

Clusterização

Agrupamento

Redução de Dimensionalidade



Aprimoramento

Unsupervised Learning

Sistemas de Recomendação

Agrupamento de notícias semelhantes.

Classificação de clientes em subgrupos com padrões de compra semelhantes – ofertas diferenciadas

Unsupervised Learning

Aprendizado Não-Supervisionado

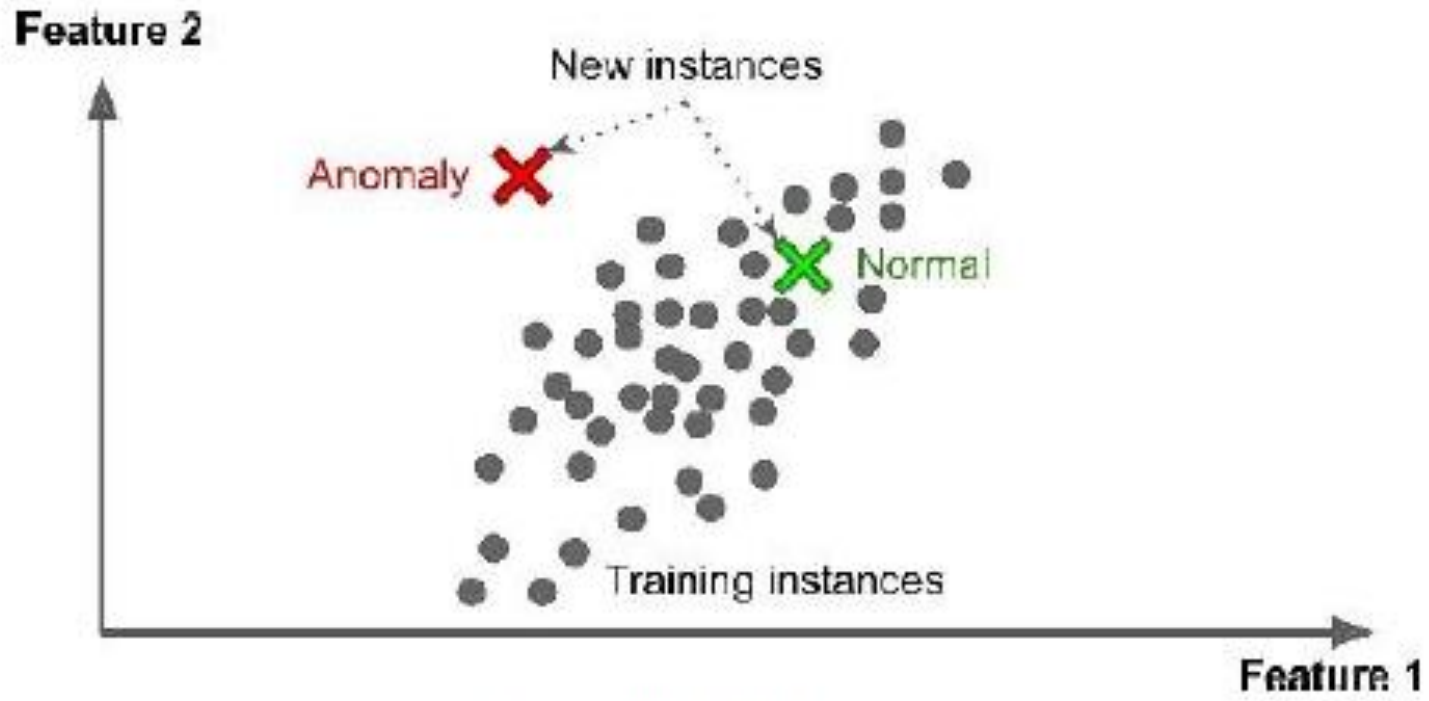
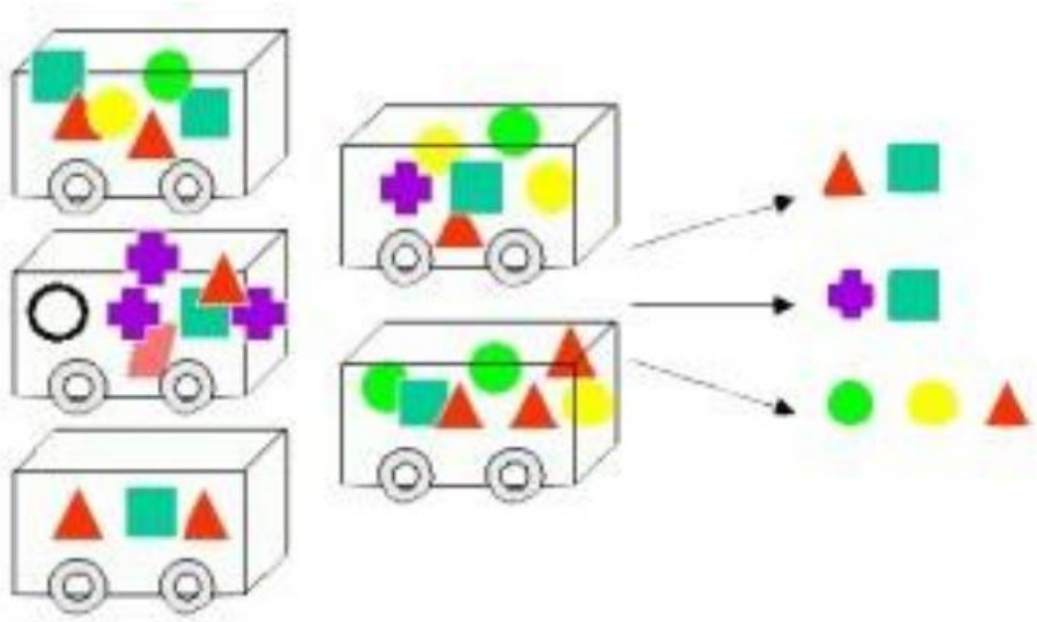
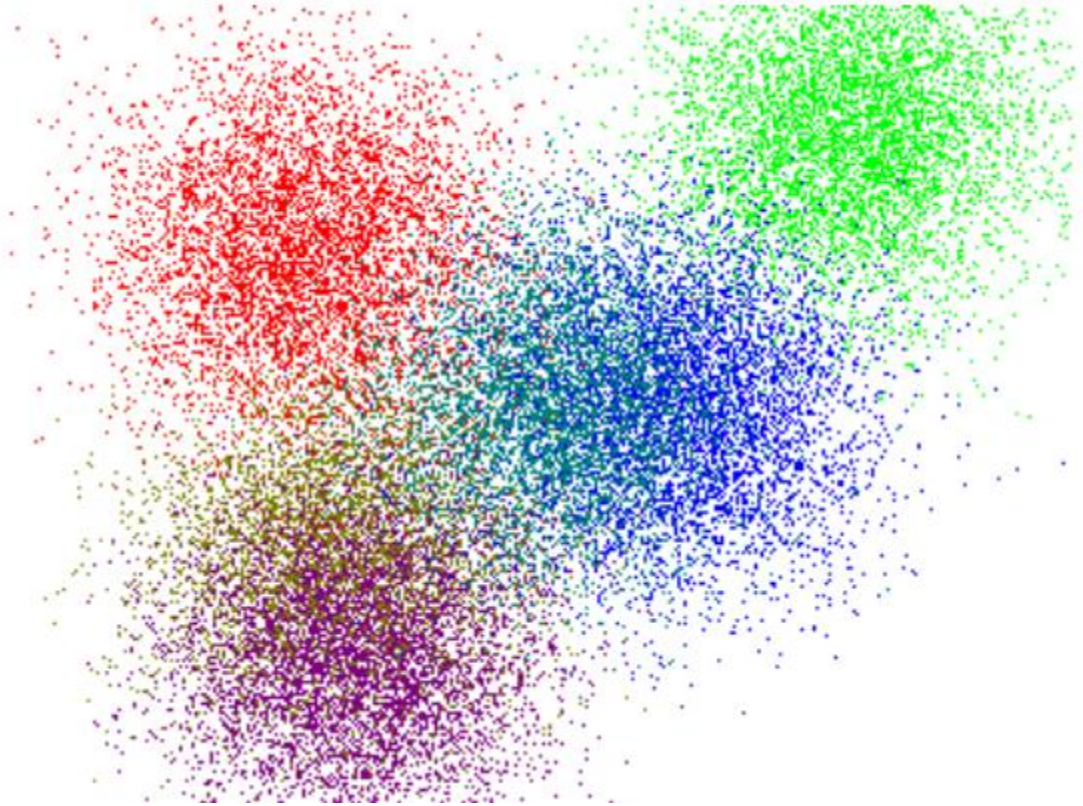


Figure 1-10. Anomaly detection

Aprendizagem não-supervisionada

- Analisar automaticamente os dados (associação, agrupamento)
- Necessita análise para determinar o significado dos padrões encontrados





Reinforcement Learning

Reinforcement Learning

- Este tipo de algoritmo é usado principalmente em robótica, jogos e navegação.
- Através deste método, o algoritmo aprende através de tentativa e erro quais pares de estado-ação obtém a maior recompensa no longo prazo.
- O objetivo do agente é escolher ações que maximizem a recompensa esperada dada uma determinada quantidade de tempo.
- O agente irá, dessa forma, criar uma política de tomada de decisões, baseadas no seu atual estado.

Reiforcement Learning

Reinforcement Learning

- Exemplo com um robô que dirige:
  - Faz o algoritmo aprender com tentativa e erro, onde apresenta para o algoritmo uma série de estados.
  - “Estou a 80 km/h e tenho uma conversão para a direita a 30 metros” este conjunto de informação é um estado. Para o estado há um conjunto de possibilidades, pois o robô pode virar a direita ou a esquerda X graus, pode manter a direção na mesma velocidade, pode acelerar ou frear o carro.
  - O robô vai interagindo com o meio ao longo do tempo de forma interativa e vai encontrando e maximizando qualquer são as atitudes que ele tem que tomar para cada par de estados que maximiza a recompensa dele no longo prazo que compensa no caso pode se chegar ao destino ou não.
  - O objetivo do agente de escolha do agente é escolher ações que maximiza a recompensa esperada dada a capacidade de tempo para o gente. Dessa forma criar uma política de tomada de decisões baseado no seu estado.
  - Ele cria uma política de decisões ação estado e maximiza isso se consegue encontrar a melhor forma de agir em relação ao tempo.

Reinforcement Learning

Aprendizado por Reforço

Reinforcement Learning

Aprendizado por Reforço

Observa o ambiente e toma determinadas ações!

Reinforcement Learning

Recompensas

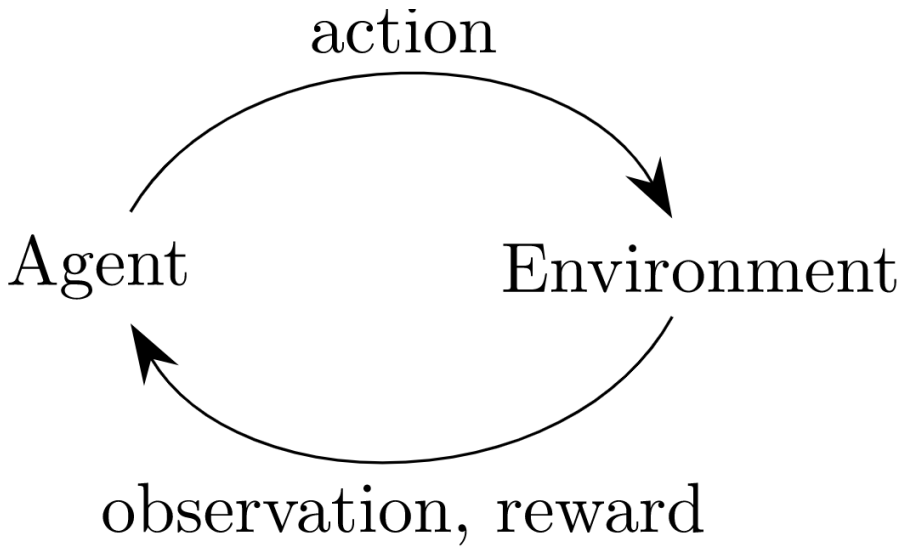
Penalidades

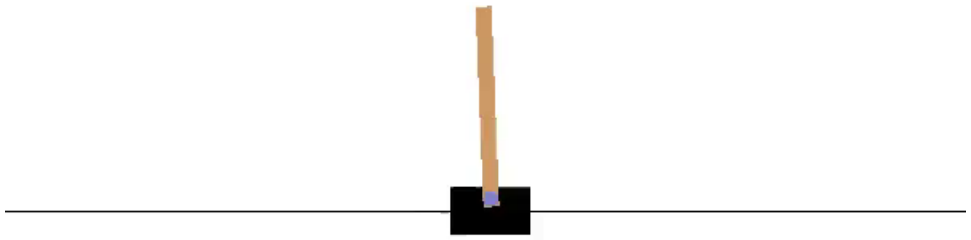
Reinforcement Learning

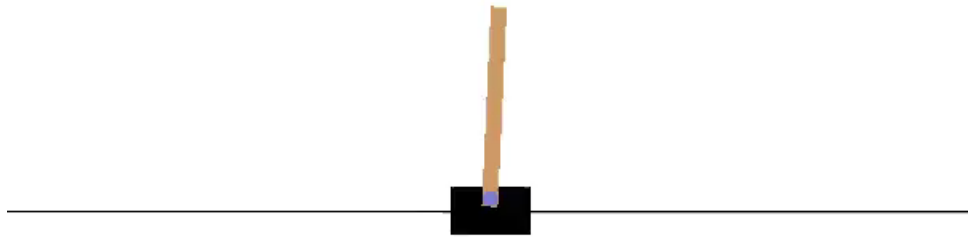
https://www.linkedin.com/feed/update/urn:li:activity:6416849908929691648

Reinforcement Learning

https://gym.openai.com/docs/







Aprendizagem por reforço

- Aprender com as interações com o ambiente (causa e efeito)
- Aprender com sua própria experiência
- Robô coletando lixo aprendendo a andar em um ambiente
- Controle automatizado de elevadores

Modos

Aprendizagem de Máquina		
Supervisionada	**Não supervisionada**	**Reforço**
Classificação	Associação	
Regressão	Agrupamento	
	Detecção de desvios	
	Padrões sequenciais	
	Sumarização	

Capacidade e generalização

Aprender normal

A pessoa com oculos, algoritmo consegui identificar

- **Capacidade e generalização são as duas qualidades que gostaríamos que nossos modelos de AM adquirissem**: a primeira lhe da força para aprender as regularidades nos dados em que treinamos o modelo; a segunda faz com que ele consiga generalizar o que aprendeu para dados novos. Infelizmente essas duas forças então em polos opostos, de forma que ter mais de uma geralmente significa perder mais da outra. A seguir, vamos detalhar bem como esse *tradeoff* acontece e como ponderar essas duas forças.

Teorema No Free Lunch

- O problema que temos é complexo: temos que aprender uma regra geral para a relação entre x e y, mas utilizando apenas alguns exemplos. Como aponta Goodfellow et al (2016), isso é contraditório. Do ponto de vista lógico, não podemos inferir regras gerais a partir de uma quantidade limitada de exemplo. Nós então focamos em um problema mais simples: aprender uma regra que está aproximadamente correta na maioria dos casos. É importante lembrar que tudo o que estimamos, estimamos com erro!
- Devemos então escolher algum algoritmo para aprender essa regra provavelmente aproximadamente correta (PAC).
- No entanto, o teorema " **No Free Lunch** " (Wolpert, 1996) nos diz que nenhum algoritmo de Aprendizado de Máquina é melhor do que outro universalmente. Felizmente, só precisamos de um algoritmo que seja melhor no nosso problema particular, e isso é possível de encontrar.

Teorema No Free Lunch

- Por que é necessário introduzir tantas abordagens estatísticas diferentes de aprendizagem, em vez de apenas um único melhor método?
- Não há almoço grátis nas estatísticas: nenhum método domina todos os outros em todos os conjuntos de dados possíveis. Em um conjunto de dados específico, um método específico pode funcionar melhor, mas algum outro método pode funcionar melhor em um conjunto de dados semelhante, mas diferente.
- Portanto, é uma tarefa importante decidir, para qualquer conjunto de dados, qual método produz os melhores resultados. Selecionar a melhor abordagem pode ser uma das partes mais desafiadoras da realização de aprendizado estatístico na prática.

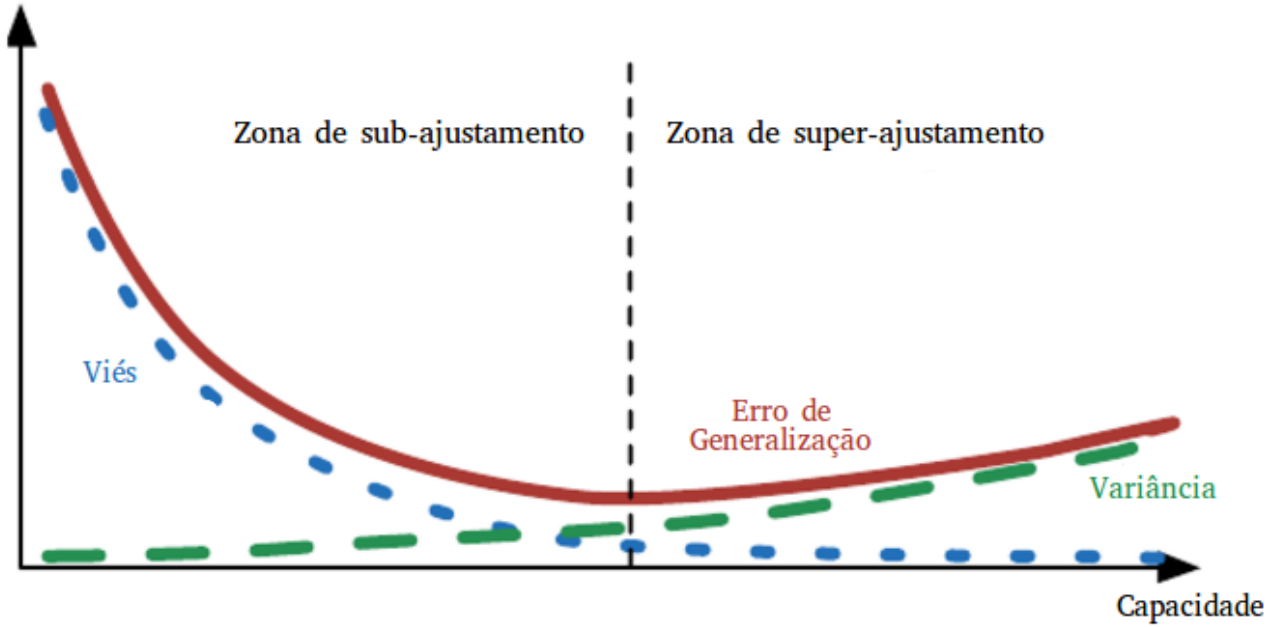
Balanço Viés-variância

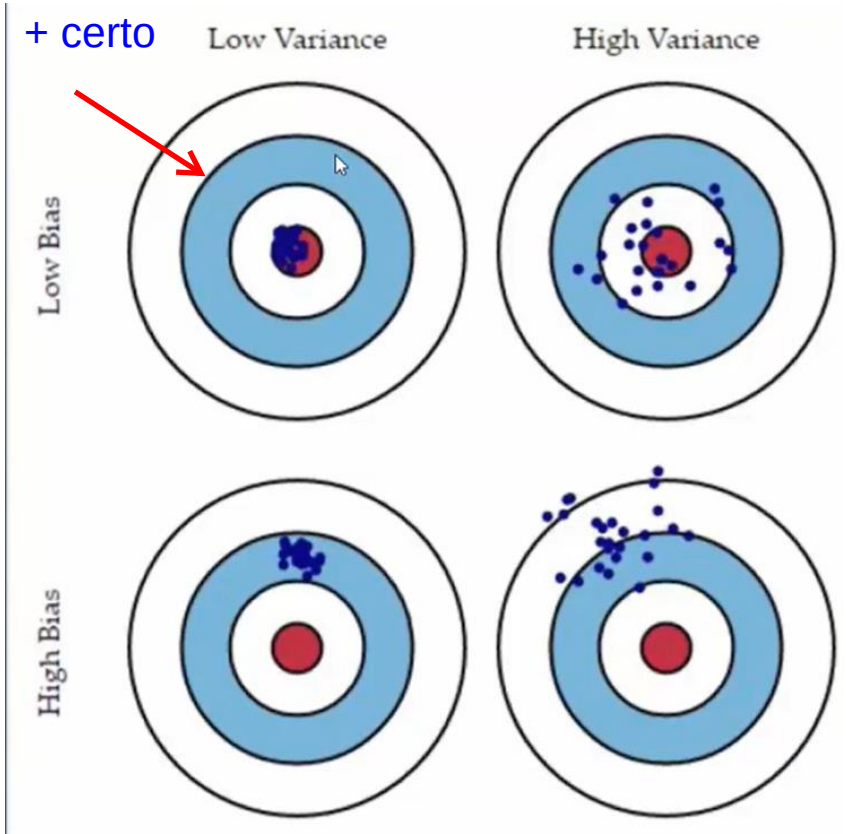
Tendência do modelo acertar ou errar

O modelo ganha viés a medida em que a média das predições se afasta do valor correto. É a tendência que o modelo tem de aprender errado por não levar em consideração toda informação necessária, então um modelo com viés alto é um modelo que pode não se ajustar bem aos dados. O viés (bias) de um modelo é a diferença entre a predição esperada e o modelo correto que nós tentamos prever para determinados pontos de dados.

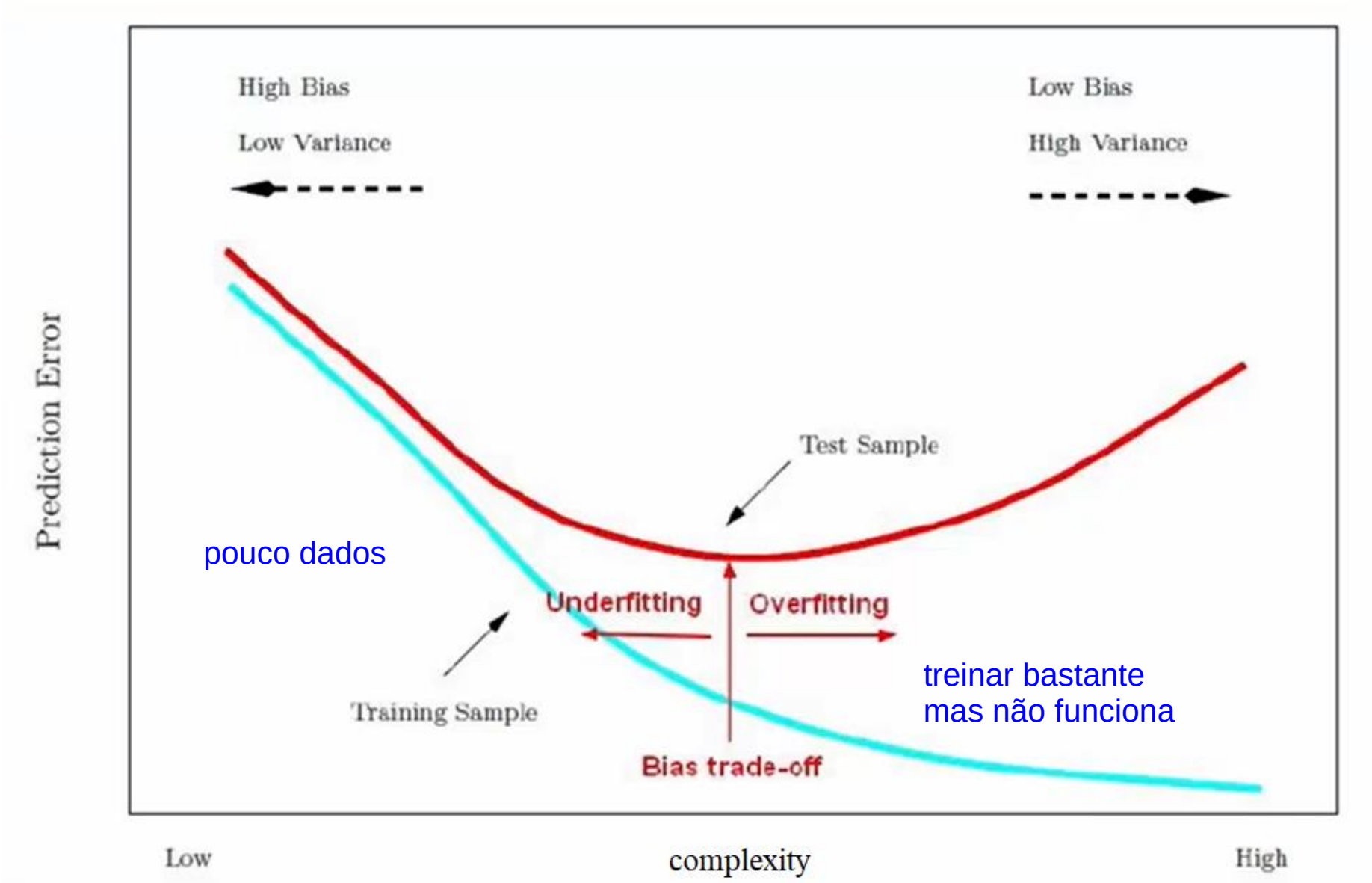
A variância é o que acontece com o modelo para cada novo ponto de dados. Ou seja, tem a ver com a sua estabilidade em resposta a novos exemplos. O mesmo também ganhará variância quanto mais diferentes forem as tentativas entre si. A variância (*variance*) de um modelo é a variabilidade da previsão do modelo para determinados pontos de dados.

Quanto mais simples o modelo, maior o viés e, quanto mais complexo o modelo, maior a variância.





Balanço Viés-variância

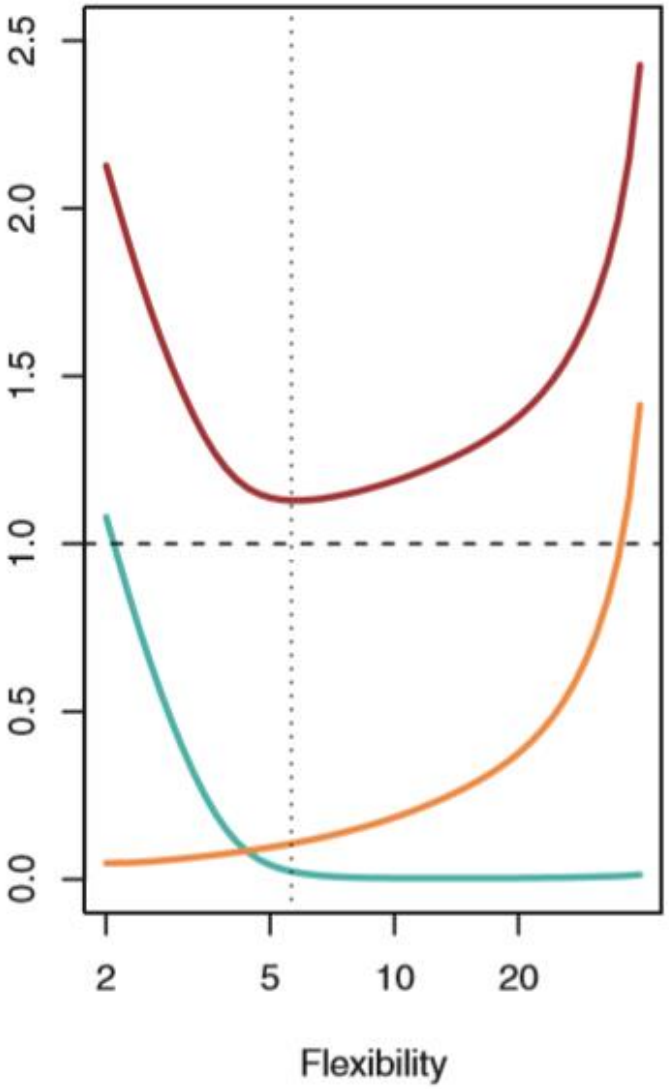


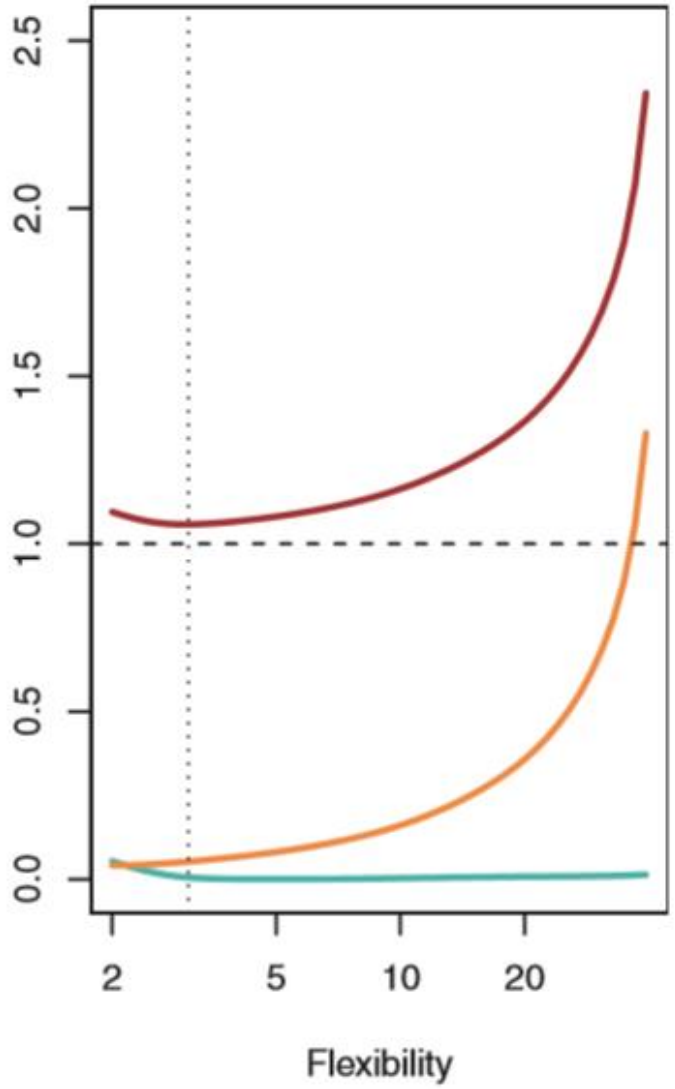
Balanço Viés-variância

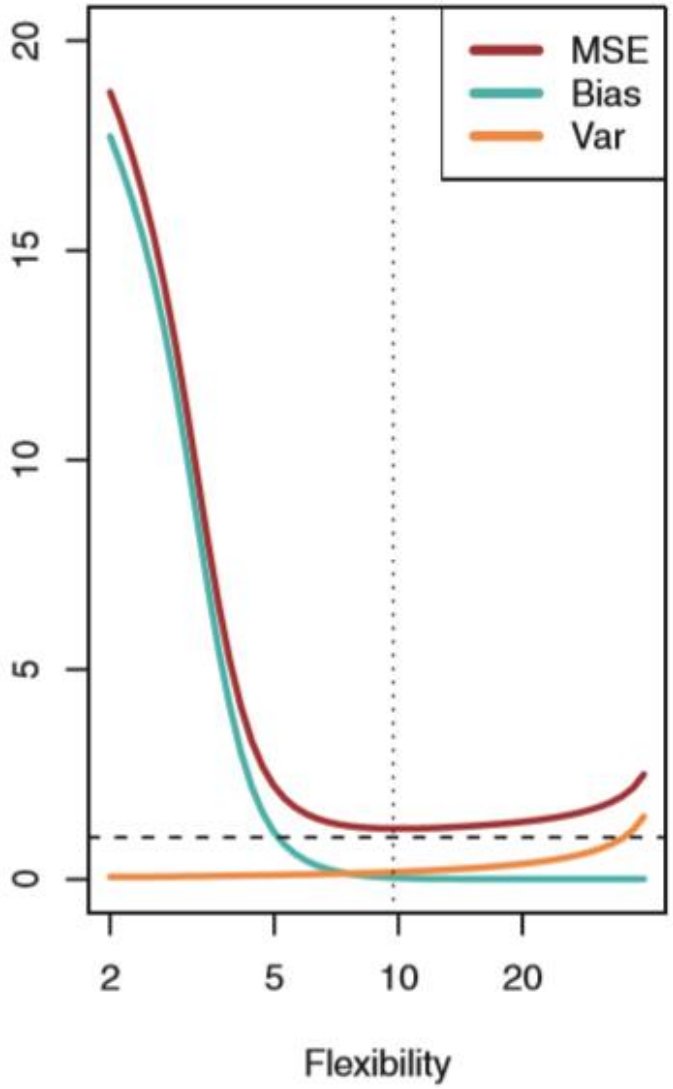
Viés - verde

Variancia - laranja

MSE - Erro



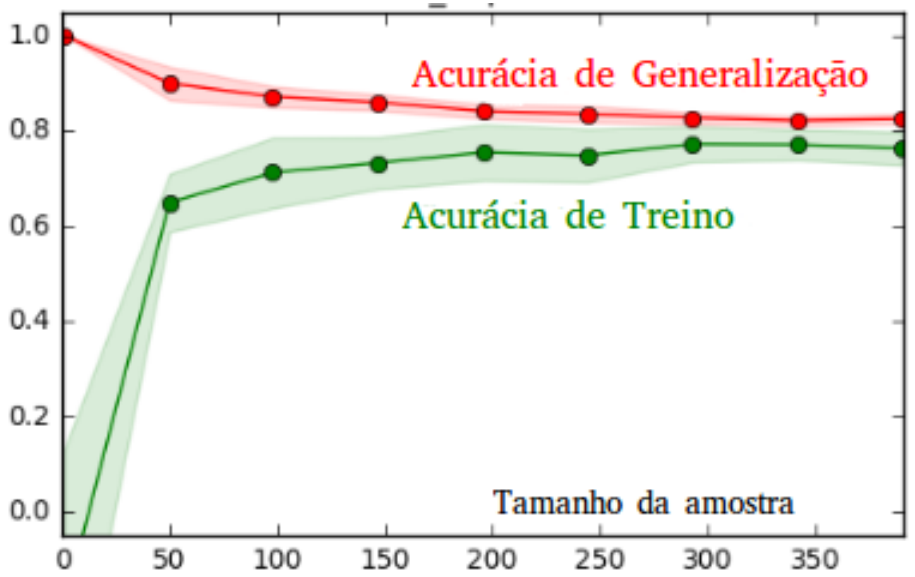




Balanço Viés-variância

Ponto de ajuste

- Um detalhe muito importante é que a variância pode ser diminuída aumentando a quantidade de dados de treinamento, algo que não pode ser feito com o viés. Pense no algoritmo utilizado para gerar o terceiro modelo, por exemplo. Se utilizássemos muitos dados de treinamento, eventualmente a sua capacidade não seria mais suficiente para passar por todos os pontos, ele então deixaria de aprender ruído para se concentrar em aprender o sinal do processo gerador de dados. Em se tratando do primeiro modelo, não importa quantos dados utilizássemos para treinar, ele não aprenderia o sinal correto pois não tem capacidade para isso. No geral, erro de variação e erro de treinamento convergem quando o tamanho da amostra de treinamento cresce:
- Como não conseguimos aumentar a quantidade de dados infinitamente, a grande questão de Aprendizado de Máquina é como ajustar a capacidade de forma a balancear variância e viés, de forma que nem sobre-ajustamos nem sub-ajustamos o modelo
- O balanço viés-variância determina o ponto que estamos apenas adicionando ruído ao nosso modelo a medida que adicionamos complexidade á ele.



Cross Validation

Técnica para verificar se o modelo faz uma boa generalização

- Cross-Validation é uma das melhores técnicas para saber se o seu modelo generaliza bem. Por exemplo, se você já participou de alguma competição no Kaggle, já percebeu que a sua pontuação durante a competição é diferente da pontuação que você recebe após o desafio acabar. Muitas das vezes esse score pode aumentar, mas também diminuir. **Esse é um tipo de problema que você pode resolver com o Cross-Validation**.
- **Cross-Validation — ou Validação Cruzada — é uma técnica que visa entender como seu modelo generaliza**, ou seja, como ele se comporta quando vai prever um dado que nunca viu. Mas você pode estar se perguntando: “Não é pra isso que o conjunto de teste serve?”, e a resposta é “Sim”. A propósito, o conjunto de teste muita das vezes é chamado de conjunto de validação — validation set.
- Então, pra quê a gente precisa de Cross-Validation? Justamente para criar diferentes conjuntos de treino e teste, treinar o modelo e ter certeza de que ele está performando bem. Nesse caso, **ao invés de usarmos apenas um conjunto de teste para validar nosso modelo, utilizaremos N outros a partir dos mesmos dados**.
- Existem vários métodos de Cross-Validation, como Holdout, Leave-one-out e K-Fold

Cross Validation

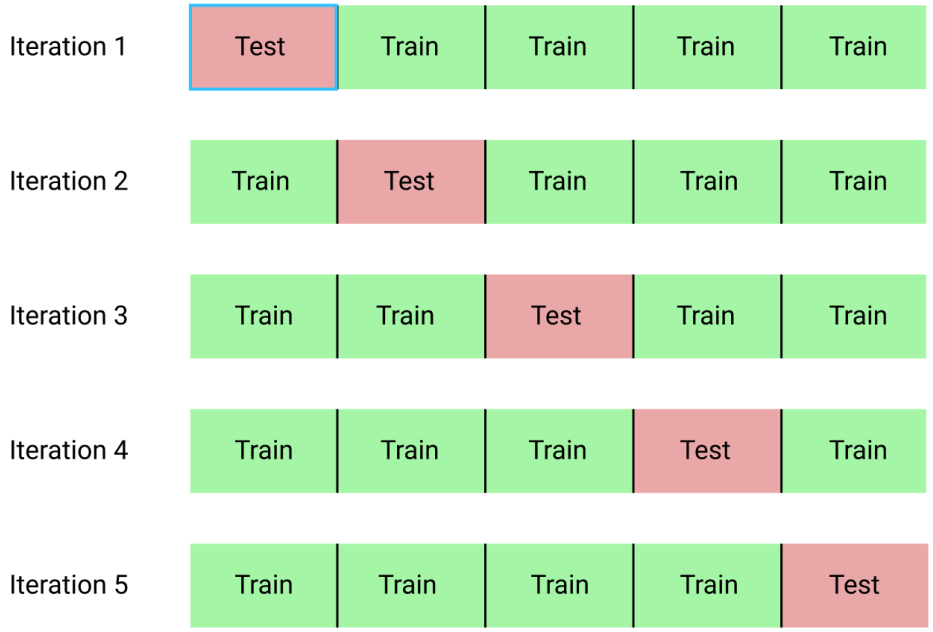
- Sabemos que não podemos confiar em métricas de erro computadas nos sets de treinamento para escolher entre modelos, pois podemos sempre reduzir esse erro a zero simplesmente aumentando a capacidade do algoritmo de aprendizado.
- Há um método extremamente simples para selecionar o modelo com menor erro de generalização - o que realmente nos interessa.
- Após coletar os dados, nós vamos separá-los em duas sub-amostras.
- Uma delas, a sub-amostra de treino, será utilizada para aprender um modelo; a outra sub-amostra será utilizada para medir a performance do modelo, que será uma estimativa do erro de generalização. Para escolher a capacidade do modelo, nós dividiremos a amostra de treino mais uma vez: em um novo set de treino e em um set que chamaremos de validação.
- Nós então **treinaremos o modelo no set de treino**, **ajustaremos a capacidade do algoritmo de aprendizado com base na performance no set de validação, e reportaremos uma estimativa final do erro de generalização conforme a performance no set de teste.**
- É importante que o set de teste seja observado apenas uma única vez, na hora de reportar a estimativa de erro final. Se múltiplas tentativas forem feitas e comparadas com base no erro no set de teste, esta medida de erro não será confiável como uma estimativa do erro de generalização e o modelo provavelmente performará pior do que o esperado, quando utilizado na prática.

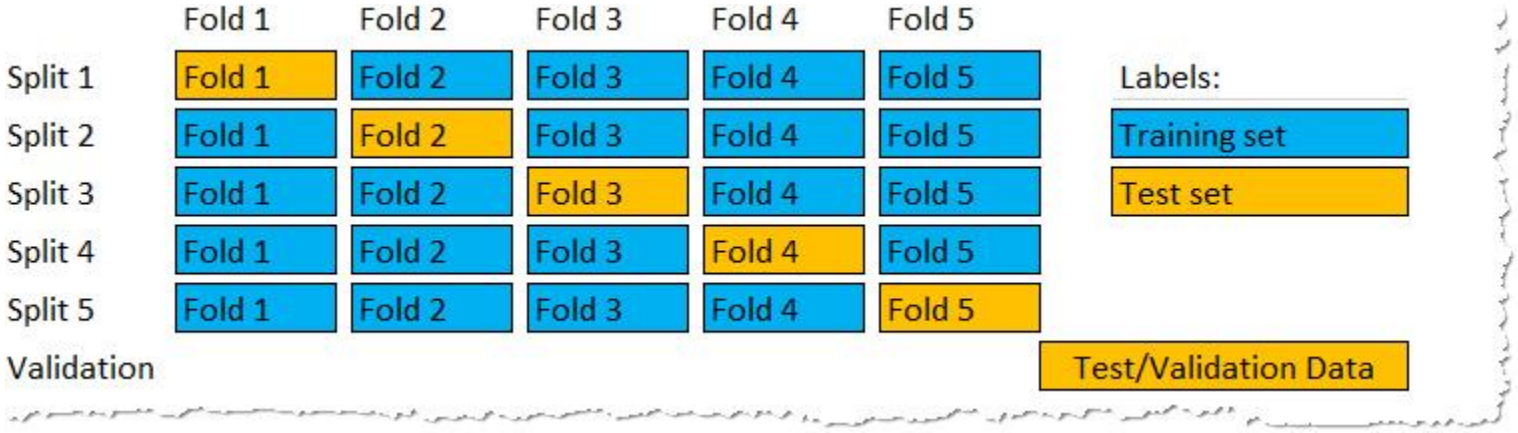
K-fold Cross Validation

dividir em sub amostra

e uma delas é test

no final tem um score - com dados que vc nunca viu





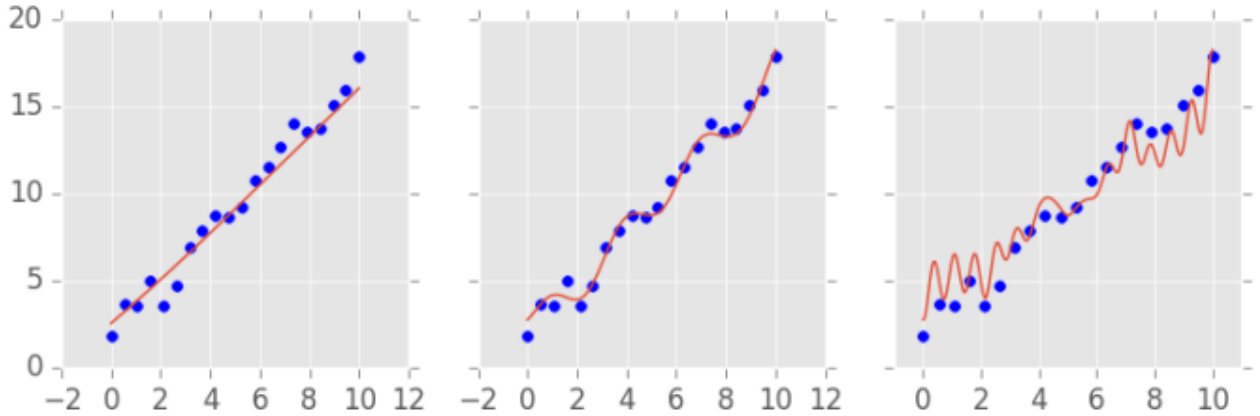
Dados que eu nunca vi é o test

O método K-Fold de Cross-Validation irá dividir seu modelo K vezes — daí o nome K-Fold. Em nosso exemplo acima, o dataset foi dividido em 5 partes, para cada uma dessas partes, o modelo irá usar 4 partes (K-1) para treinar, enquanto usará 1 parte para validar. Ao final do processo, quando o modelo iterar/treinar 5 vezes, você terá um verdadeiro score de como seu modelo está generalizando, geralmente ao tirar a média e desvio padrão de todos os treinos realizados.
Como você pode estar imaginando, esse processo faz com que o treino de seu modelo demore um pouco mais, mas é crucial se você quer ter certeza de que seus dados estão generalizando bem.

Objetivo é criar um algoritmo com MENOR vies - Menor erro

Overfitting e Underfitting

•Essas novas observações, que não foram utilizadas para treinar o modelo, são chamadas de set de teste. Com elas, é possível descobrir o erro de generalização do nosso modelo e definir qual deles é o melhor. Agora está claro que o modelo do meio, com capacidade média, é o melhor. Para ser mais específico, o modelo do meio é capaz de explicar 99% da variação no lucro (de acordo com a medida R2), ao passo que o primeiro modelo explica 96% e o último, apenas 92%.

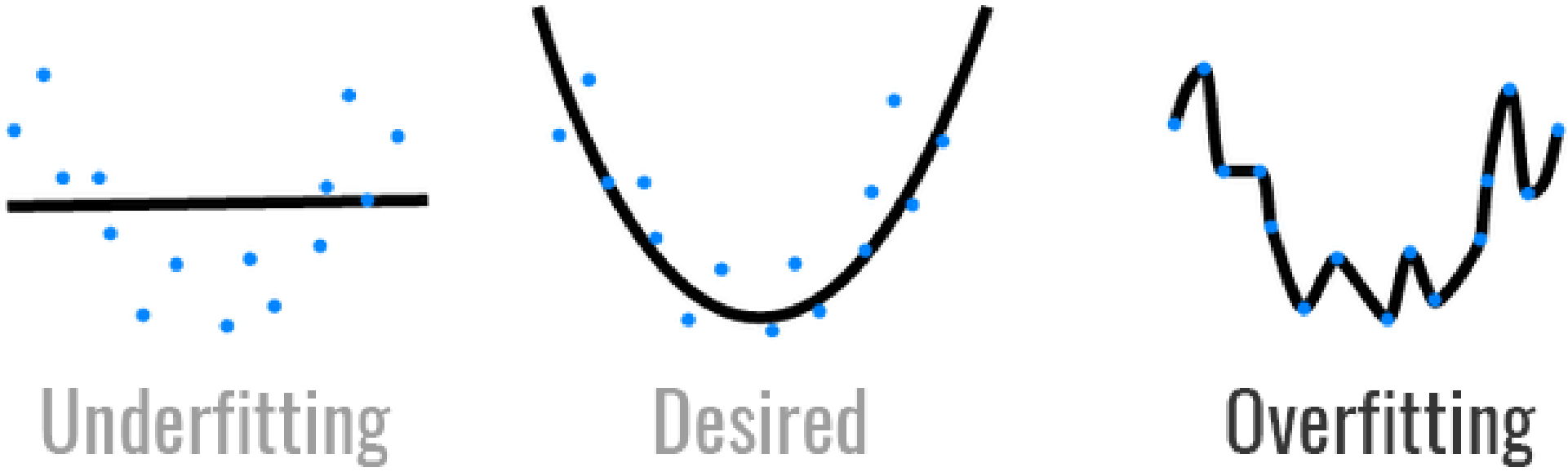


•Em Aprendizado de Máquina, dizemos que o primeiro modelo esta sofrendo de sub-ajustamento, isto é, ele tem uma capacidade tão baixa que não consegue aprender as regularidades presentes nem no set de treinamento. Nesse cenário, a performance nos sets de treino e de teste costumam ser baixas. O terceiro modelo, por sua vez, sofre de sobre-ajustamento, isto é, tem uma capacidade tão alta que além de aprender as regularidades, aprendeu também o ruído presente nos dados de treino. Nesse cenário, o erro no set de treino é extremamente baixo, mas o erro de generalização - erro no set de teste - é alto. O modelo do centro parece ter aprendido bem as regularidades dos dados sem aprender os ruídos no set de treino. Nesse cenário, os erros no set de treino e teste são parecidos e satisfatórios.

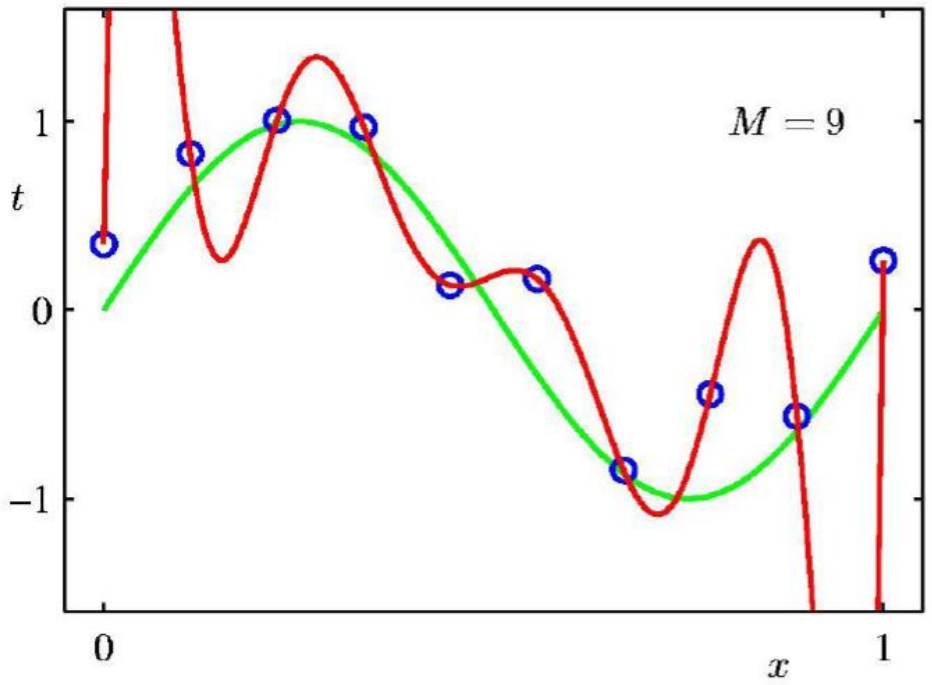
Overfitting e Underfitting

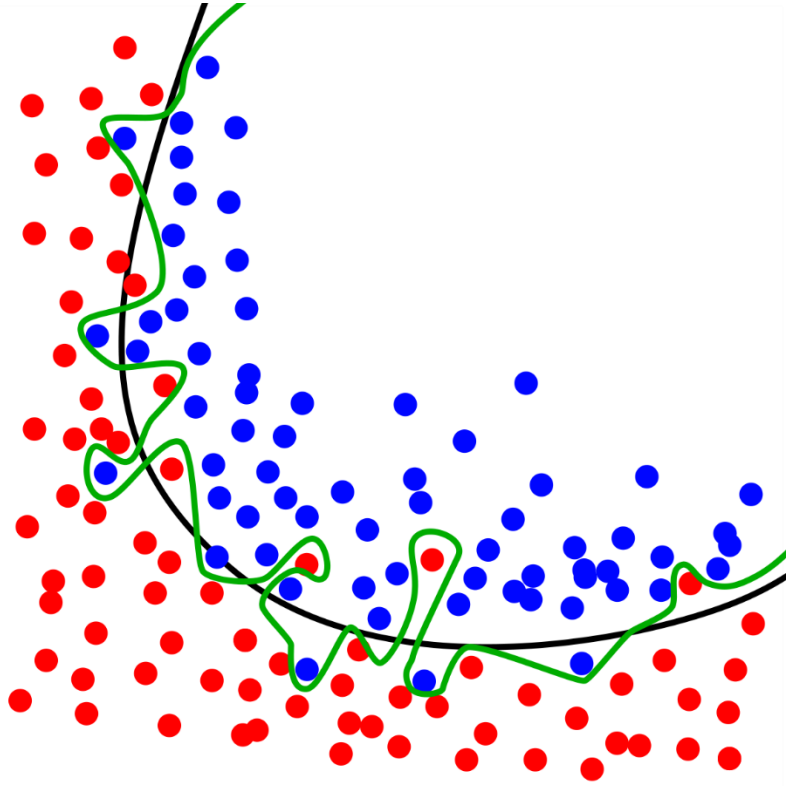
pouco treino

Passar uma situação diferente



Overfitting e Underfitting





Dimensão VC

Teoria VC (Vapnik-Chervonenkis). Atualmente é considerada a melhor teoria para estimação de parâmetros de amostras finitas, estudo de dependência funcional e de aprendizado preditivo.

A teoria de VC abrange quatro partes importantes na sua implementação:

- Teoria da Consistência dos processos de aprendizagem;
- Teoria da Taxa de convergência dos processos de aprendizagem;
- Teoria da Minimização do Risco Estrutural;
- Teoria da Otimização.

Dimensão VC Vapnik-Chervonenkis

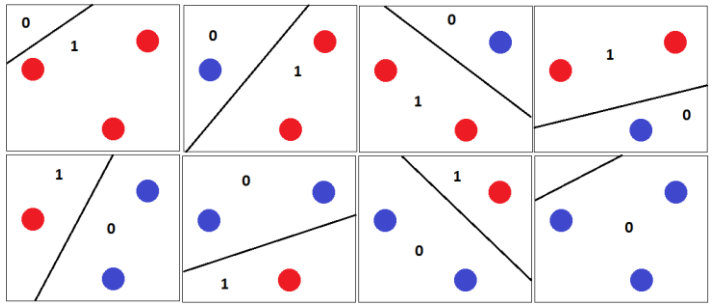
$$P(\sup_{f \in F} |Remp(f) - R(f)| A = \pi r^2 \geq \epsilon) \leq 2N(f, 2n) ex^{(-2n\epsilon^2)}$$

Remp(f) erro ou risco medido em uma amostra

R(f) erro ou risco esperado para dados nunca vistos

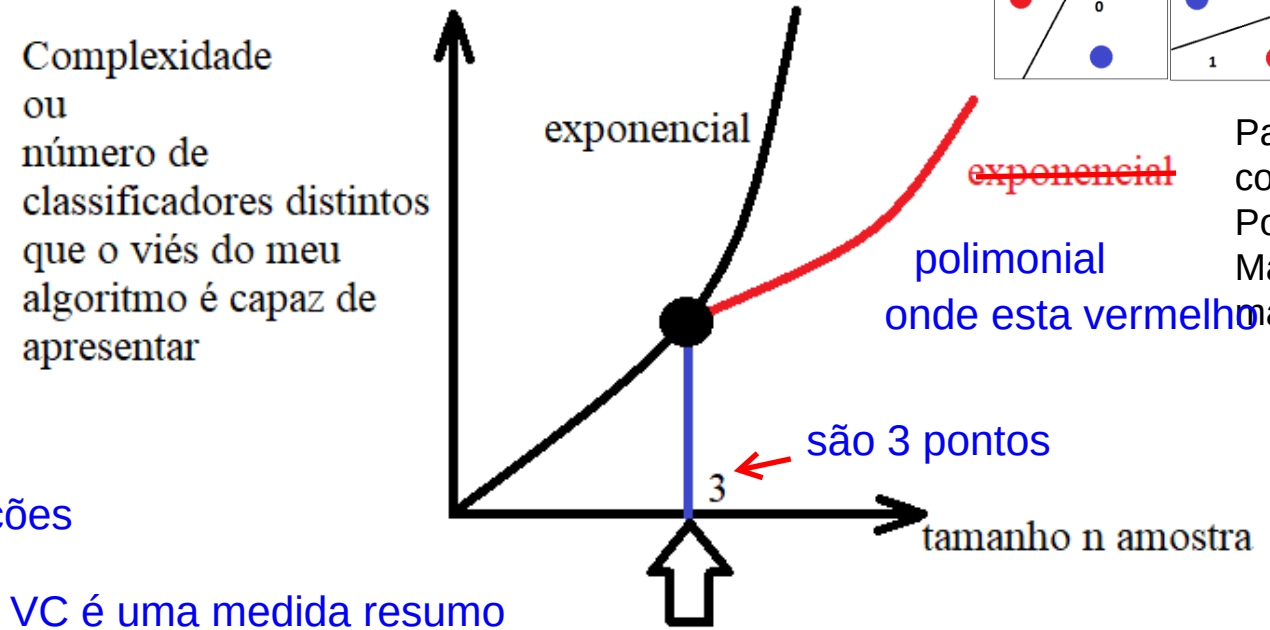
limite de dados para aprender

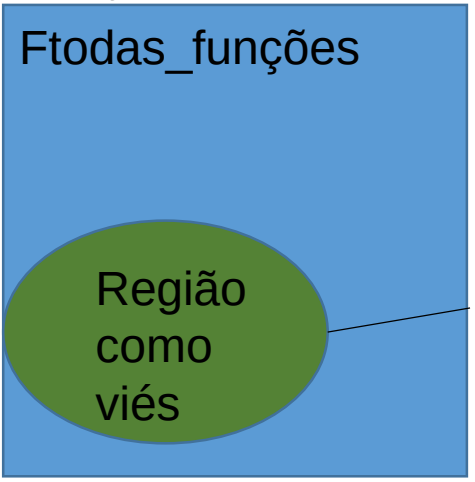
coeficiente de shattering



Para 3 pontos atingiu todas combinações em espaço R2.
Pois 2^3=8
Mas para 4 pontos será no máximo 14 possibilidades

VC(Algoritmo1)=3
VC(Algoritmo2)=
Significa que o espaço de funções admissíveis conhecido como viés é mais complexo, ou seja contém mais funções que conseguem cortar o espaço





Ftodas_funções

Universo de funções

Medida resumo da complexidade deste espaço

VC é uma medida resumo

VC é um natural que define o tamanho de amostra para o qual até determinado gera o número exponencial de classificadores e a partir daqui é polinomial. VC é uma medida resumo que se for ilimitada então o comportamento do algoritmo é exponencial em termo do viés então o aprendizado não será possível. Significa que se crescer um tamanho de amostra além do tamanho da dimensão VC o aprendizado pode começar a ocorrer.

Limites de aprendizagem: risco empírico versus risco estrutural.

- A abordagem tradicional para separar os dados é usar uma função, como um polinômio, e então ajustar seus parâmetros para separar os dados de treinamento, agrupando-os em uma das classes.
- Durante a fase de treinamento, se aumentado o grau do polinômio é possível reduzir o erro nos dados, o que levaria a um melhor aprendizado; No entanto, esta estratégia pode levar ao *overfitting*, resultando em uma capacidade reduzida de generalização nos dados futuros. Uma alternativa nos modelos tradicionais é a redução significativa do grau do polinômio, porém isto pode gerar um erro nos dados de treinamento, chamado *underfitting*.
- O risco empírico pode ser reduzido à zero ao custo de uma função de decisão extremamente complexa. A distribuição dos dados de treinamento pode não ser complexa de ser classificada, porém, características como “ruídos” podem fazer com que o processo de aprendizado seja muito mais complexa que a realidade.
- Na teoria de aprendizado estatístico existem condições matemáticas para a escolha de um classificador com desempenho desejado para dados de treinamento, minimizando o erro estrutural do processo, ou seja, evitando um *overfitting* ou um *underfitting*. A teoria da Minimização do Risco Estrutural formaliza o conceito de controle de complexidade e minimização do risco empírico.
- Desta forma, se o objetivo é minimizar o erro da classificação, a máquina deve conseguir minimizar tanto o risco empírico quanto o termo de complexidade, o que nos permite chegar ao *well-trained*. Ou seja, o modelo ajustado para o conjunto de dados que será analisado.

Métodos

Aprendizagem de Máquina	
Métodos Preditivos	**Métodos Descritivos**
Classificação	Associação
Regressão	Agrupamento
	Detecção de desvios
	Padrões sequenciais
	Sumarização

Métodos Preditivos

- Classificação
- Regressão

Classificação: definir classes para cada registro

- Marketing direto
- Insatisfação de clientes
- Risco de crédito
- Filtros de SPAM
- Separação de notícias
- Reconhecimento de voz
- Reconhecimento de face
- Previsão de doenças

Bom pagador ou não

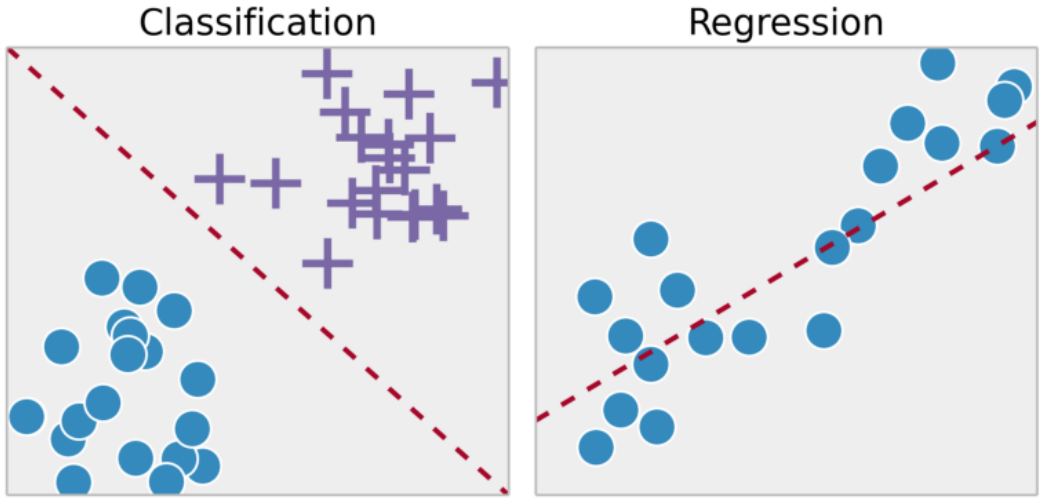
é spam ou não

Regressão

Valores

tendência é valorizar ou não

- Gastos propaganda -> valor de venda
- Temperatura, umidade e pressão do ar -> velocidade do vento
- Fatores externos -> valor do dólar
- Resultados do exame -> probabilidade de um paciente sobreviver
- Risco de investimento
- Gastos no cartão de crédito, histórico -> limite
- Valores anteriores -> valores de produtos



Classificação *versus* Regressão

- Na Classificação você tem rótulos
- Na regressão você tem números

- Por exemplo, enquanto “Risco de Crédito” na Classificação pode ser rotulado como, “*alto*”, *“médio”* ou *“baixo”* , em “Risco de Investimento”, utilizando a Regressão, a saída seria um valor. Por exemplo, 95% de possibilidade de retorno do investimento.

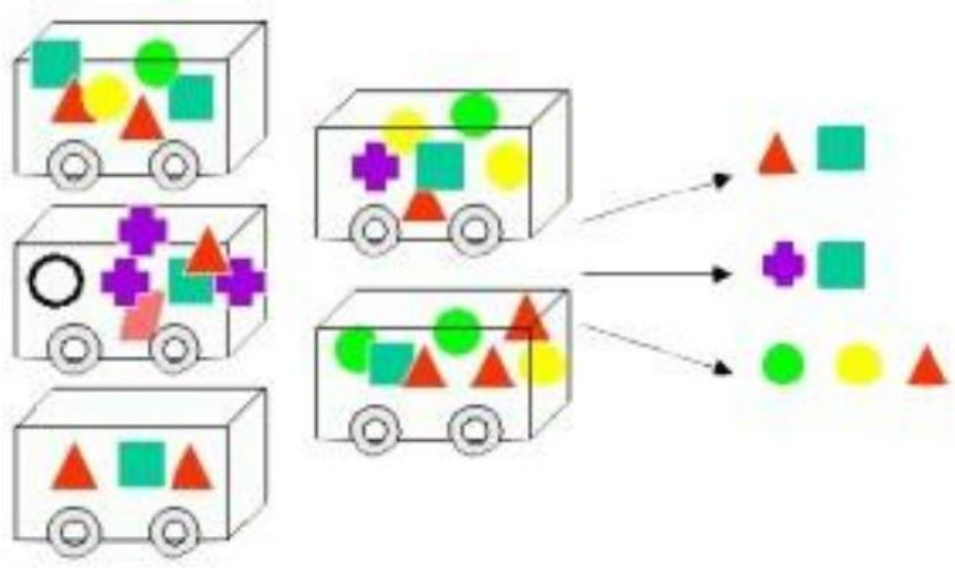
Métodos Descritivos

- Associação
- Agrupamento
- Detecção de desvios
- Padrões sequenciais
- Sumarização

Associação

frauda com cerveja

- Prateleiras de mercado
- Promoções com itens que são vendidos em conjunto
- Planejar catálogos das lojas e folhetos de promoções
- Controle de evasão em universidades

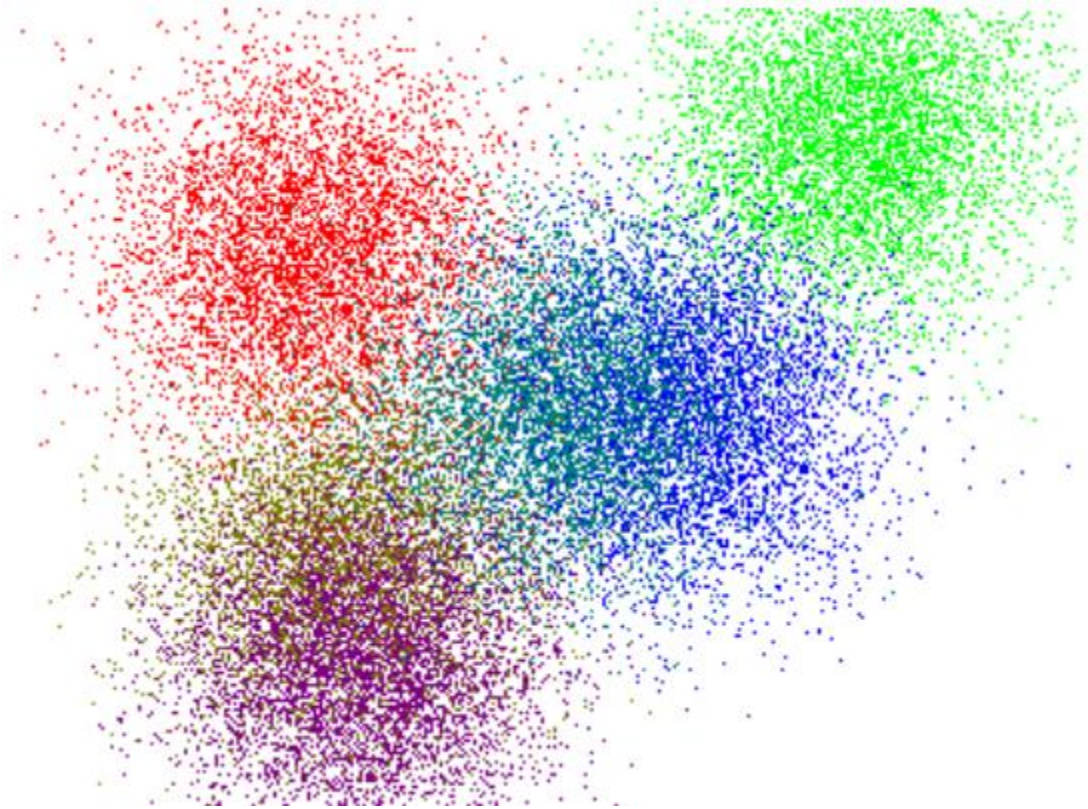


Agrupamento

Agrupar as emendas - objetivo rotular

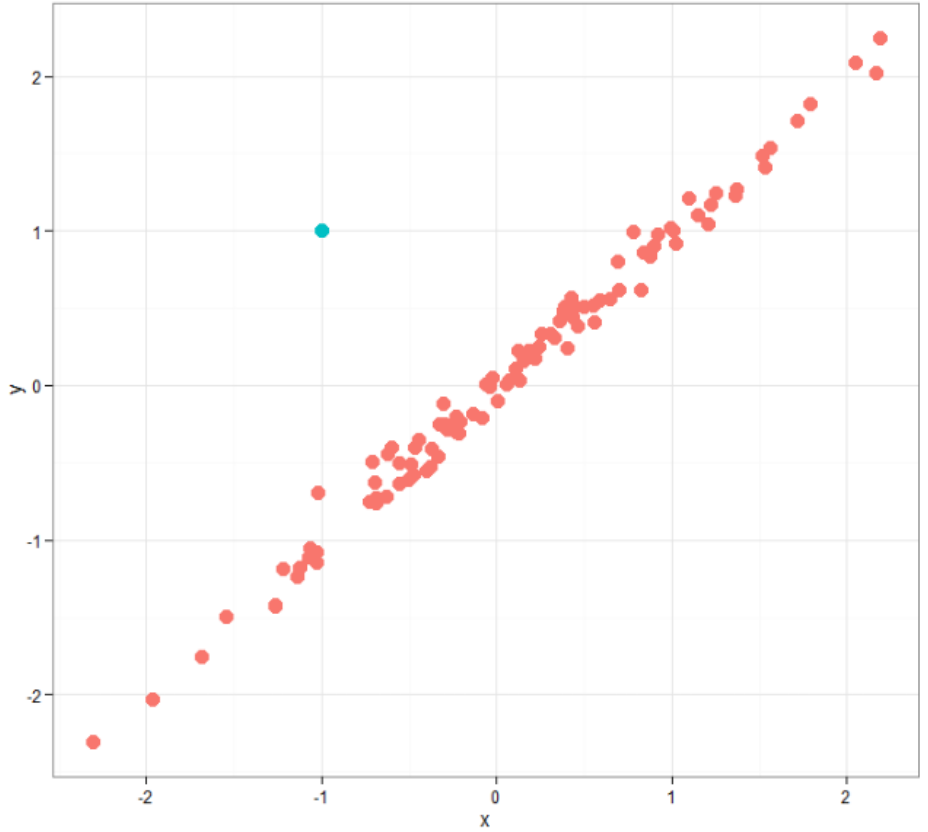
Washington post - juntou as noticias

- Segmentação de mercado
- Encontrar grupos de clientes que irão comprar um produto (mala direta)
- Agrupamento de documentos/notícias
- Agrupamento de produtos similares
- Perfis de clientes (NetFlix)
- Análise de redes sociais



Detecção de Desvios

- Fraude em cartão de crédito
- Intrusão em redes
- Uso de energia elétrica, água ou telefone
- Desempenho de atletas (doping)
- Monitorar máquinas em um data center

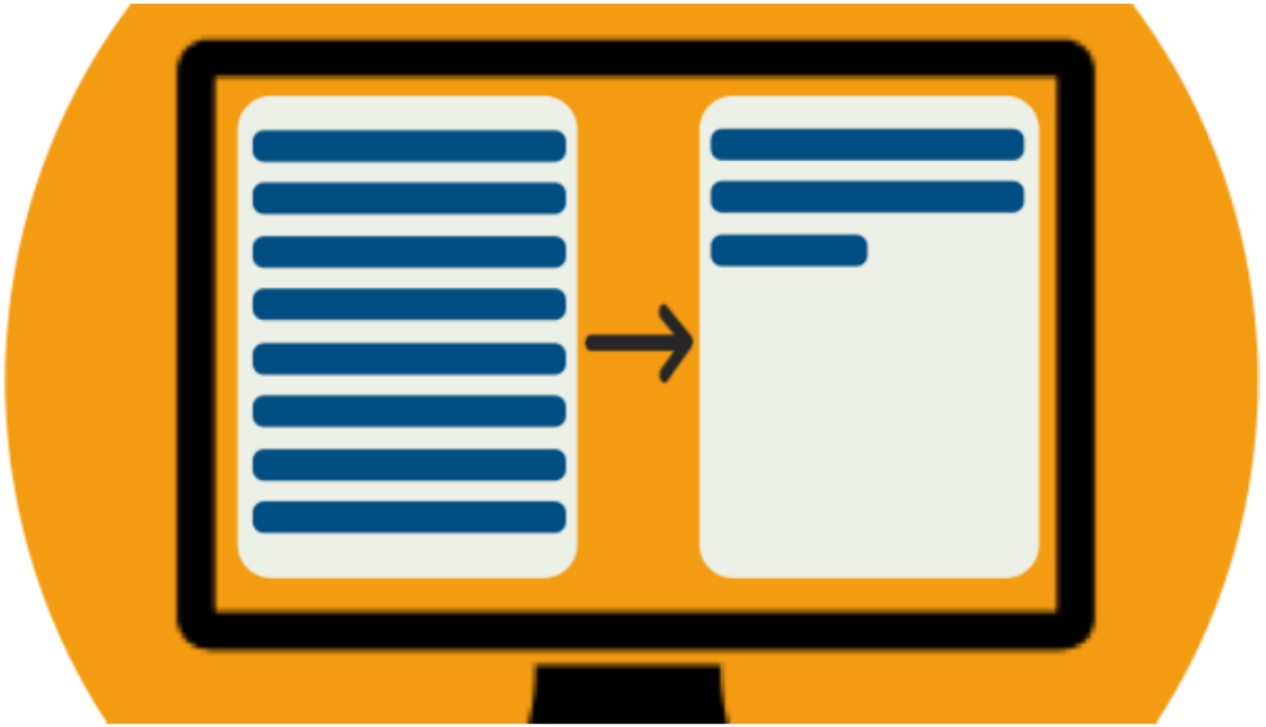


Descoberta de padrões sequenciais

- Livrarias, loja de equipamentos de atletismo, computadores
- Marketing direcionado para pessoas que tem maiores chances de adquirir um novo produto
- Prevenção de doenças
- Navegação em sites

Sumarização

- São ouvintes do programa homens na faixa de 45 a 60 anos, com nível superior e que trabalham na área de tecnologia
- Segmentação de mercado



Métodos Preditivos

Classificação (risco de crédito)

História do crédito	Dívida	Garantias	Renda anual	Risco
Ruim	Alta	Nenhuma	< 15.000	Alto
Desconhecida	Alta	Nenhuma	>= 15.000 a <= 35.000	Alto
Desconhecida	Baixa	Nenhuma	>= 15.000 a <= 35.000	Moderado
Desconhecida	Baixa	Nenhuma	< 15.000	Alto
Desconhecida	Baixa	Nenhuma	> 35.000	Baixo
Desconhecida	Baixa	Adequada	> 35.000	Baixo
Ruim	Baixa	Nenhuma	< 15.000	Alto
Ruim	Baixa	Adequada	> 35.000	Moderado
Boa	Baixa	Nenhuma	> 35.000	Baixo
Boa	Alta	Adequada	> 35.000	Baixo
Boa	Alta	Nenhuma	< 15.000	Alto
Boa	Alta	Nenhuma	>= 15.000 a <= 35.000	Moderado
Boa	Alta	Nenhuma	> 35.0000	Baixo
Ruim	Alta	Nenhuma	>= 15.000 a <= 35.000	Alto

Treinamento

História do crédito	Dívida	Garantias	Renda anual
Ruim	Alta	Adequada	< 15.000
Desconhecida	Alta	Adequada	< 15.000
Desconhecida	Baixa	Nenhuma	> 35.000
Boa	Alta	Adequada	>= 15.000 a <= 35.000

Métodos Preditivos

Classificação (venda de livros)

Sexo	País	Idade	Comprar
M	França	25	Sim
M	Inglaterra	21	Sim
F	França	23	Sim
F	Inglaterra	34	Sim
F	França	30	Não
M	Alemanha	21	Não
M	Alemanha	20	Não
F	Alemanha	18	Não
F	França	34	Não
F	França	34	Não
M	França	55	Não
M	Inglaterra	25	Sim
M	Alemanha	48	Sim
F	Inglaterra	23	Não

Treinamento

Sexo	País	Idade
M	França	38
F	Inglaterra	25
M	Alemanha	55
F	França	20

Métodos Preditivos

Classificação (prever o esporte)

Cor dos olhos	Casado	Sexo	Cabelo	Esporte
Castanho	Sim	M	Longo	Futebol
Azul	Sim	M	Curto	Futebol
Castanho	Sim	M	Longo	Futebol
Castanho	Não	F	Longo	Aeróbica
Castanho	Não	F	Longo	Aeróbica
Azul	Não	M	Longo	Futebol
Castanho	Não	F	Longo	Aeróbica
Castanho	Não	M	Curto	Futebol
Castanho	Sim	F	Curto	Aeróbica
Castanho	Não	F	Longo	Aeróbica
Azul	Não	M	Longo	Futebol
Azul	Não	M	Curto	Futebol

Treinamento

Cor dos olhos	Casado	Sexo	Cabelo
Castanho	Sim	M	Curto
Castanho	Não	M	Longo
Azul	Não	F	Longo
Azul	Sim	M	Longo

Métodos Preditivos

Classificação (jogar tênis)

Tempo	Temperatura	Humidade	Vento	Jogar tênis
Ensolarado	Quente	Alta	Fraco	Não
Ensolarado	Quente	Alta	Forte	Não
Nublado	Quente	Alta	Fraco	Sim
Chuvoso	Moderada	Alta	Fraco	Sim
Chuvoso	Agradável	Normal	Fraco	Sim
Chuvoso	Agradável	Normal	Forte	Não
Nublado	Agradável	Normal	Forte	Sim
Ensolarado	Moderada	Alta	Fraco	Não
Ensolarado	Agradável	Normal	Fraco	Sim
Chuvoso	Moderada	Normal	Fraco	Sim
Ensolarado	Moderada	Normal	Forte	Sim
Nublado	Moderado	Alta	Fraco	Sim
Nublado	Quente	Normal	Fraco	Sim
Chuvoso	Moderado	Alta	Forte	Não

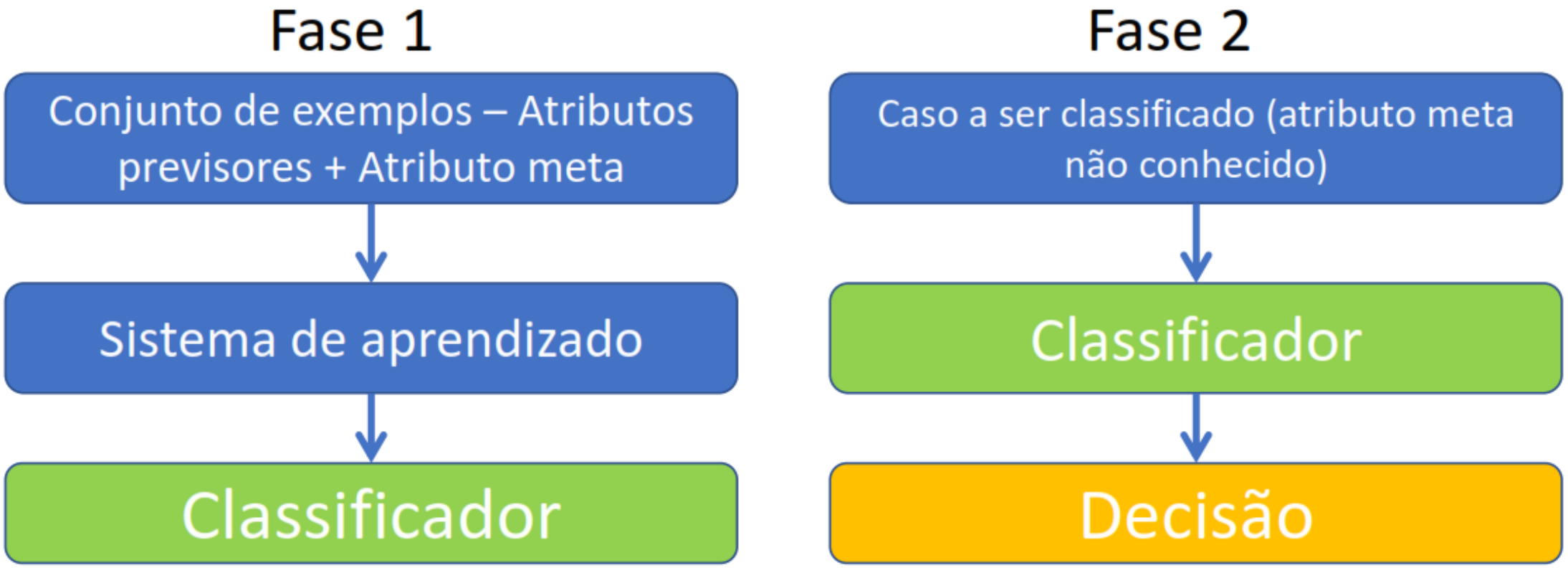
Treinamento

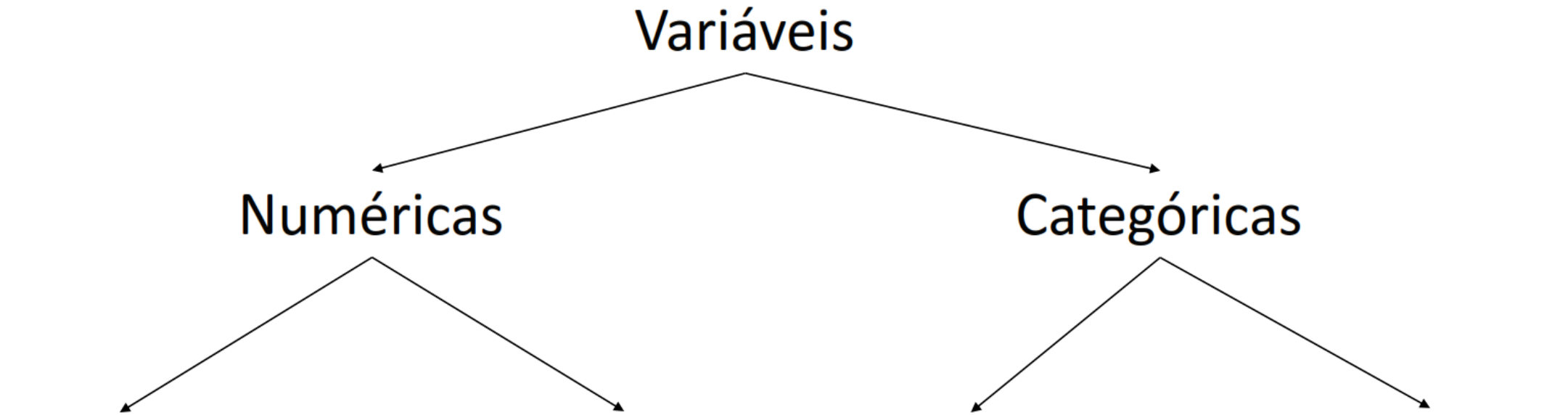
Tempo	Temperatura	Humidade	Vento
Ensolarado	Moderada	Normal	Forte
Chuvoso	Agradável	Normal	Fraco
Nublado	Quente	Normal	Forte
Nublado	Agradável	Alta	Forte

Classificação

- Cada registro pertence a uma classe e possui um conjunto de atributos previsores
- Objetiva-se descobrir um relacionamento entre os atributos previsores e o atributo meta
- O valor do atributo meta é conhecido (aprendizagem supervisionada)

Representação da classificação (método indutivo)



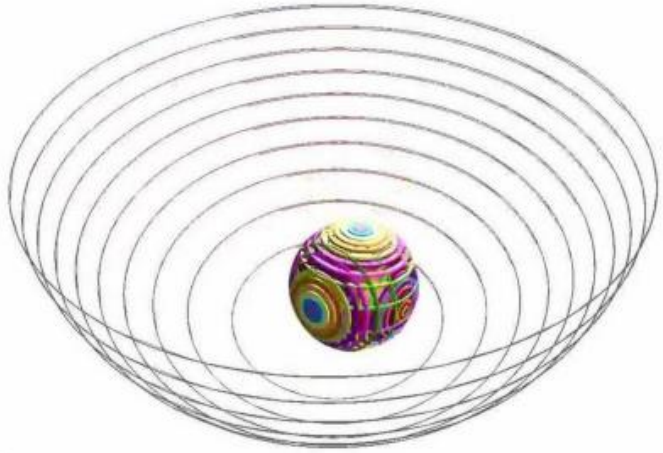

Variáveis
Numéricas
Categóricas
Contínua
Discreta
Nominal
Ordinal
Números reais
Temperatura, altura, peso, salário
Conjunto de valores finito (inteiros)
Contagem de alguma coisa
Dados não mensuráveis
Sem ordenação: cor dos olhos, gênero
Categorizado sob uma ordenação
Tamanho P, M e G

Erro

Risco de Crédito			
	Alto	Moderado	Baixo
Alto	28	7	3
Moderado	6	32	2
Baixo	5	8	25

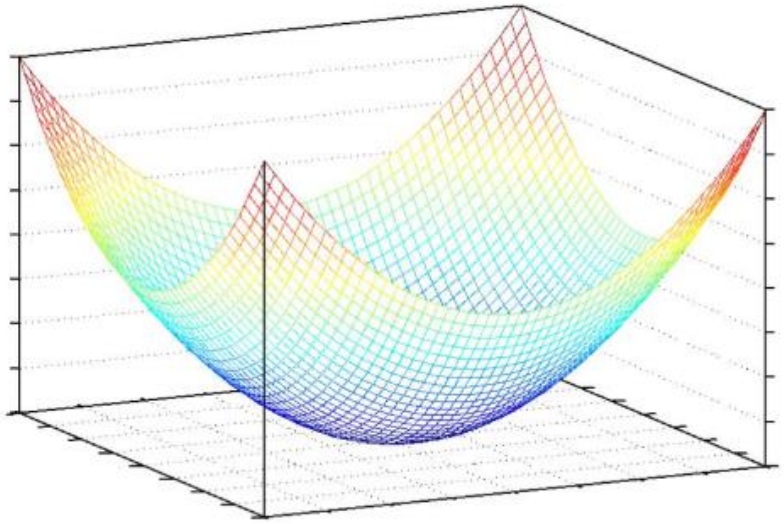
Para medirmos a percentagem de acerto devemos fazer o somatório da diagonal principal
Acertos = 85
Erros = 31
Total de registros = 116
Percentual de acerto = (85/116) * 100 = 73.27
Percentual de erro = (31/116) * 100 = 26.73

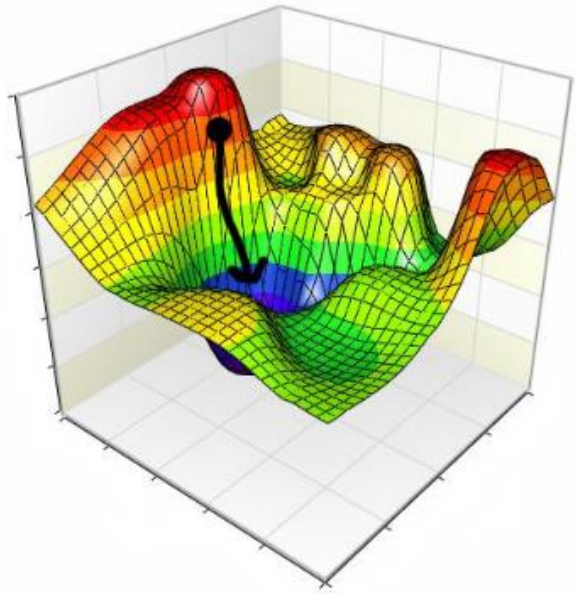
Gradiente



$\min C(w_1, w_2 \ldots w_n)$

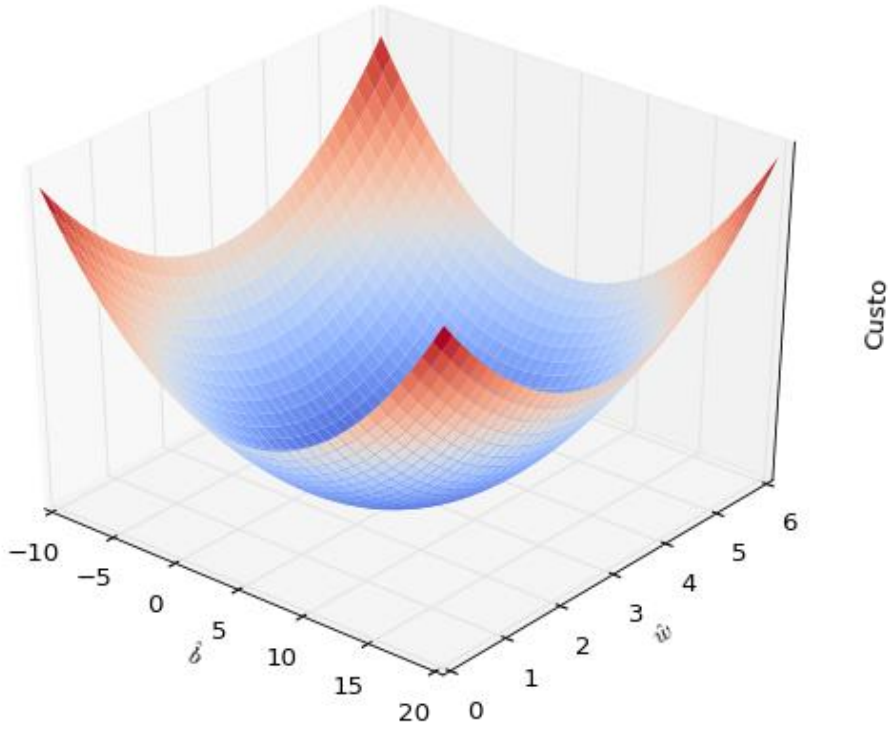
Calcular a derivada parcial para mover para a direção do gradiente





O **método do gradiente** (ou **método do máximo declive**) é um método numérico usado em otimização. Para encontrar um mínimo (local) de uma função usa-se um esquema iterativo, onde em cada passo se toma a direção (negativa) do gradiente, que corresponde à direção de declive máximo. Pode ser encarado como o método seguido por um curso da água, na sua descida pela força da gravidade.

Gradiente



A ideia pro trás dos métodos iterativos de otimização é bastante simples: nós começamos com algum chute razoável para os valores de b e w e vamos atualizando-os na direção certa até que chegamos no valor mínimo da nossa função custo, em que *w* é o vetor com os parâmetros, incluindo b.

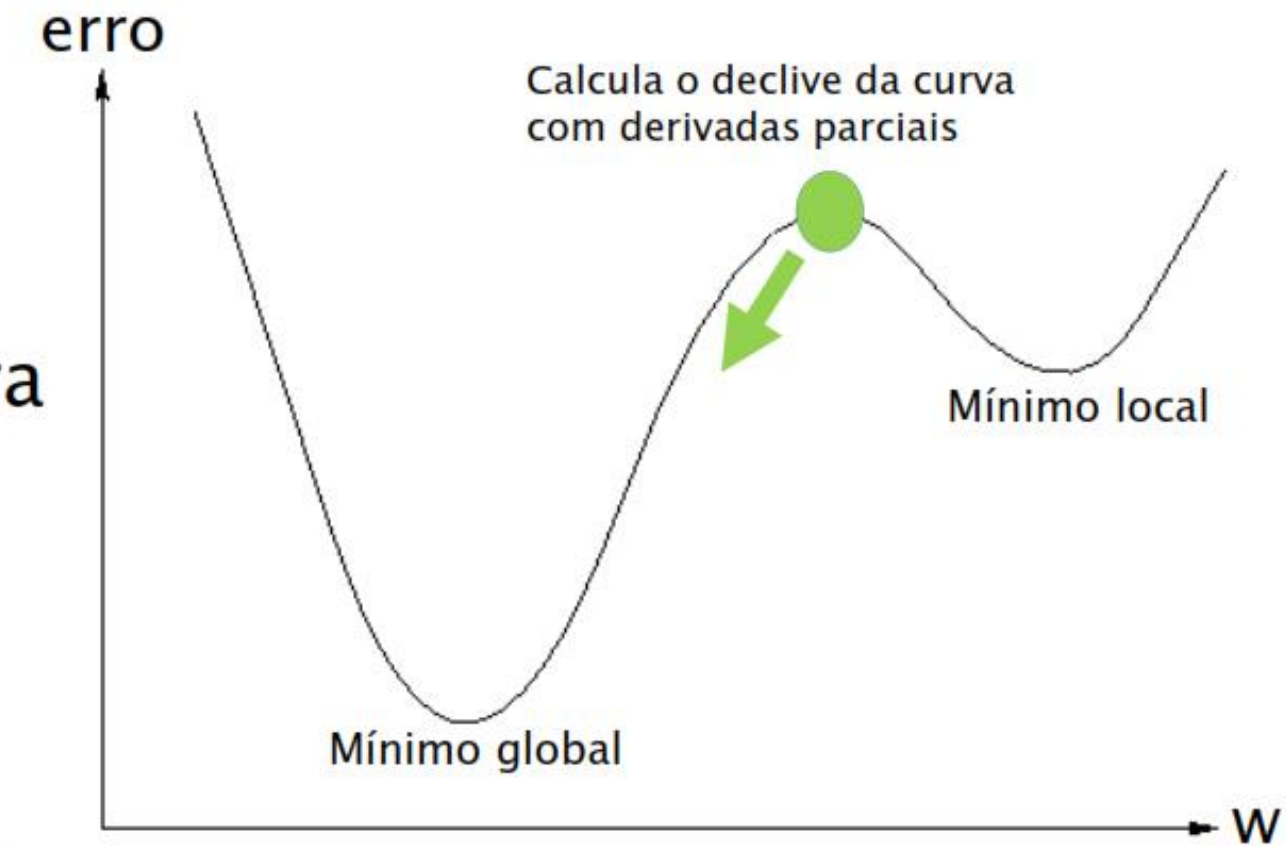
O gradiente dessa função é simplesmente um vetor de derivadas parciais, que dão a inclinação dessa tigela em cada ponto e em cada direção:

$$\nabla(L) = \left[\frac{\partial L}{\partial \hat{b}}, \frac{\partial L}{\partial \hat{w}}\right]$$

Se nós seguirmos na direção oposta do gradiente, então chegaremos no ponto de mínimo

Gradiente

- Encontrar a combinação de pesos que o erro é o menor possível
- Outras maneiras
  - Força bruta
  - Simmulated anealing
  - Algoritmos genéticos
- Gradiente é calculado para saber quanto ajustar os pesos



• Matriz de confusão

- A matriz de confusão (confusion matrix) é usada para termos uma cenário mais completa quando estamos avaliando o desempenho de um modelo. Uma matriz de confusão é uma tabela que indica os erros e acertos do seu modelo, comparando com o resultado esperado (ou etiquetas/labels).
- Verdadeiro positivo e falso positivo

		Classe esperada	
		Gato	Não é gato
Classe prevista	Gato	25 Verdadeiro Positivo	10 Falso Positivo
	Não é gato	25 Falso Negativo	40 Verdadeiro Negativo

Métricas

- Precisão
- Recall
- Acurácia
- Especificidade
- F1 Score
- Suporte
- Confiança
- Raiz do Erro Quadrático Médio (RMSE)
- Erro médio quadrático
- Erro médio absoluto (MAE)

Métricas

Precisão = **TP / (TP + FP)** Entre todos que acho verdadeiro realmente achei tantos verdadeiros. Quão precisas são as predições positivas. Dentre todas as classificações de classe Positivo que o modelo fez, quantas estão corretas;

Recall = Revocação/Sensibilidade = **TP / (TP + FN)** Entre todos verdadeiros achei tantos corretos. Cobertura da amostra positiva real. Dentre todas as situações de classe Positivo como valor esperado, quantas estão corretas.

Specificity Especificidade = **TN / (TN + FP).** Cobertura da amostra negativa real

Acurácia = **(TP + TN) / (P + N)**. Entre todos exemplos que achei positivo/correto achei tantos negativos. Desempenho geral do modelo. Dentre todas as classificações, quantas o modelo classificou corretamente;

F1 Score = 2 / (recall^(-1) + precisao^(-1)) = 2 * (precisao*recall)/(precisao+recall) = **2TP/(2TP+FP+FN)** Métrica híbrida útil para classes desequilibradas

Métricas

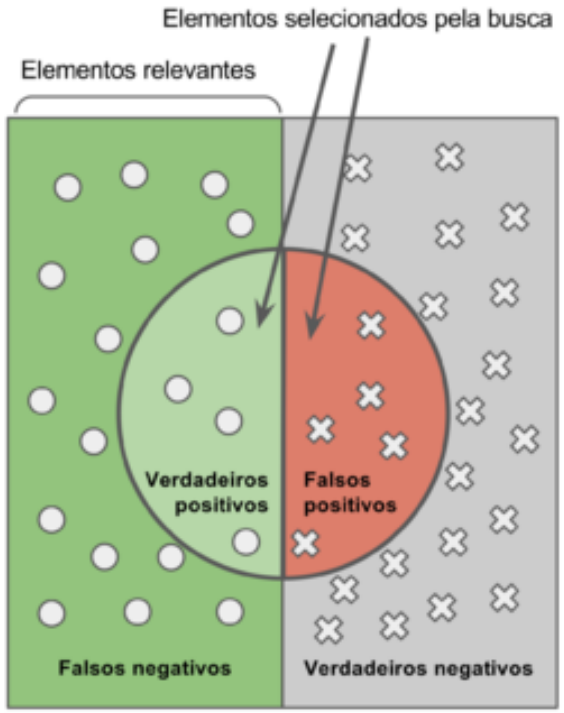
A **acurácia** é uma boa indicação geral de como o modelo performou. Porém, pode haver situações em que ela é enganosa. Por exemplo, na criação de um modelo de identificação de fraudes em cartões de crédito, o número de casos considerados como fraude pode ser bem pequeno em relação ao número de casos considerados legais. **Para colocar em números, em uma situação hipotética de 280000 casos legais e 2000 casos fraudulentos, um modelo simplório que simplesmente classifica tudo como legal obteria uma acurácia de 99,3%.** Ou seja, você estaria validando como ótimo um modelo que falha em detectar fraudes.

A **precisão** pode ser usada em uma situação em que os **Falsos Positivos são considerados mais prejudiciais que os Falsos Negativos**. Por exemplo, ao classificar uma ação como um bom investimento, é necessário que o modelo esteja correto, mesmo que acabe classificando bons investimentos como maus investimentos (situação de Falso Negativo) no processo. Ou seja, o modelo deve ser preciso em suas classificações, pois a partir do momento que consideramos um investimento bom quando na verdade ele não é, uma grande perda de dinheiro pode acontecer.

O ***recall*** pode ser usada em uma situação em que os **Falsos Negativos são considerados mais prejudiciais que os Falsos Positivos.** Por exemplo, o modelo deve de qualquer maneira encontrar todos os pacientes doentes, mesmo que classifique alguns saudáveis como doentes (situação de Falso Positivo) no processo. Ou seja, o modelo deve ter alto *recall*, pois classificar pacientes doentes como saudáveis pode ser uma tragédia.

O **F1-Score** é simplesmente uma maneira de observar somente 1 métrica ao invés de duas (precisão e *recall*) em alguma situação. É uma média harmônica entre as duas, que **está muito mais próxima dos menores valores do que uma média aritmética simples**. Ou seja, quando tem-se um F1-Score baixo, é um indicativo de que ou a precisão ou o *recall* está baixo.

Métricas

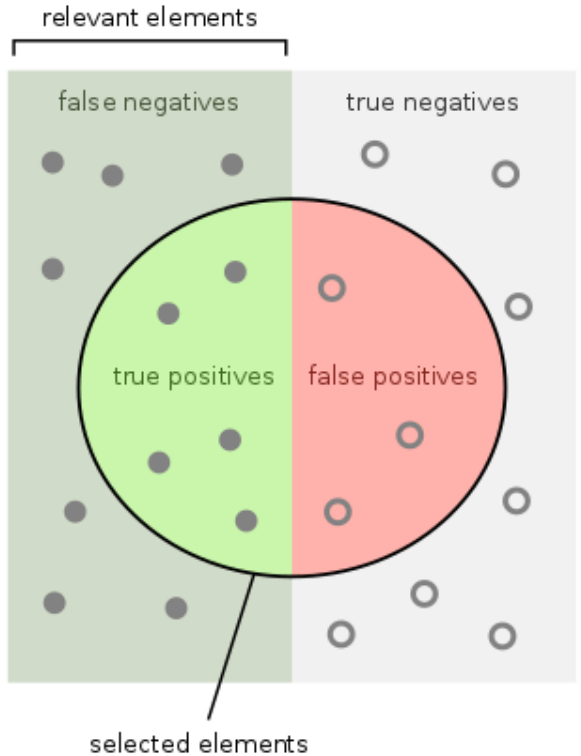


Precisão = $\frac{\text{Verdadeiros positivos}}{\text{Verdadeiros positivos} + \text{Falsos positivos}}$

"Quantos elementos selecionados são relevantes?"

Revocação = $\frac{\text{Verdadeiros positivos}}{\text{Verdadeiros positivos} + \text{Falsos negativos}}$

"Quantos elementos relevantes foram selecionados?"



How many relevant items are selected? e.g. How many sick people are correctly identified as having the condition.

Sensitivity= $\frac{\text{True positives}}{\text{True positives} + \text{False negatives}}$

How many negative selected elements are truly negative? e.g. How many healthy peple are identified as not having the condition.

Specificity = $\frac{\text{True negatives}}{\text{True negatives} + \text{False positives}}$

Métricas

Exemplo de base de dados com 4 itens e 5 transações.				
transação	leite	pão	manteiga	cerveja
1	1	1	0	0
2	0	1	1	0
3	0	0	0	1
4	1	1	1	0
5	0	1	0	0

Várias métricas podem ser utilizadas para avaliar as regras e identificar quais são interessantes. As restrições mais utilizadas são limiares mínimos de suporte e confiança.
O suporte de um conjunto X é definido como a proporção de transações da base de dados que contém esse conjunto. Suporte é uma definição dentro de quadro qual é a frequência da transação que ocorreu
sup({leite,pao})=2/5=0,4 sup({manteiga})=2/5=0,4
sup({leite,pao}=>{manteiga})=1/5= $\mathrm{conf}(X \Rightarrow Y) = \mathrm{sup}(X \cap Y)/\mathrm{sup}(X).$

A confiança de uma regra é definida.
Por exemplo, a regra {leite,pao} =>{manteiga} tem uma confiança de 0,2/0,4=0,5 na base de dados, o que significa que para 50% das transações que contém leite e pão a regra está correta.
A confiança pode ser interpretada como uma estimativa de probabilidade P(Y|X).

Métricas

Mean absolute error (erro absoluto médio) (MAE) é a média do valor absoluto dos erros:

$$\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \qquad \frac{1}{n}\sum|Real - Previsão|$$

Mean Squared Error (erro médio quadrático) (MSE) é a média dos erros quadrados:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

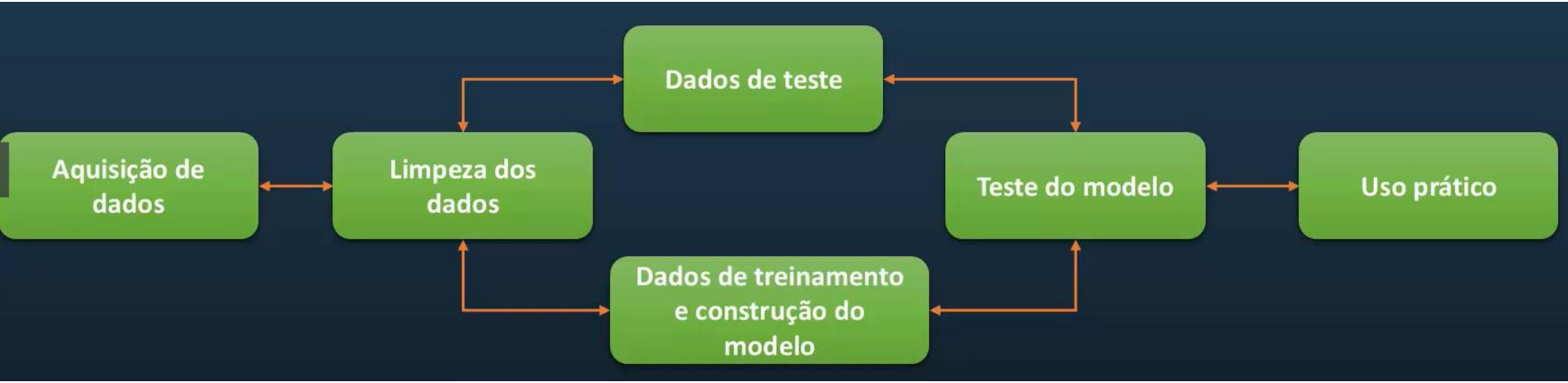
Root Mean Square Error (raiz do erro quadrático médio) (RMSE) é a raiz quadrada da média dos erros quadrados:

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Comparando estas métricas:

MAE é o mais fácil de entender, porque é o erro médio.
MSE é mais popular que o MAE, porque a MSE "puniria" erros maiores, o que tende a ser útil no mundo real.
RMSE é ainda mais popular do que MSE, porque o RMSE é interpretável nas unidades "y".

O processo básico em Machine Learning



O processo básico em Machine Learning

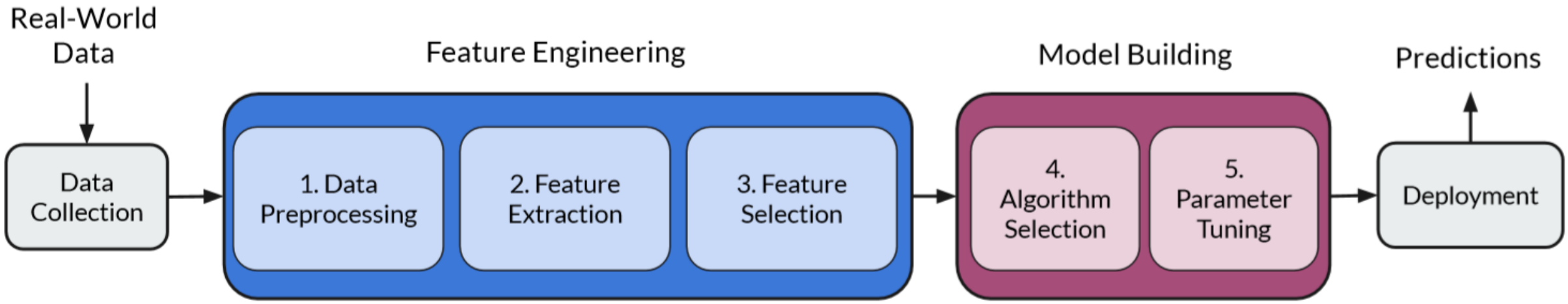
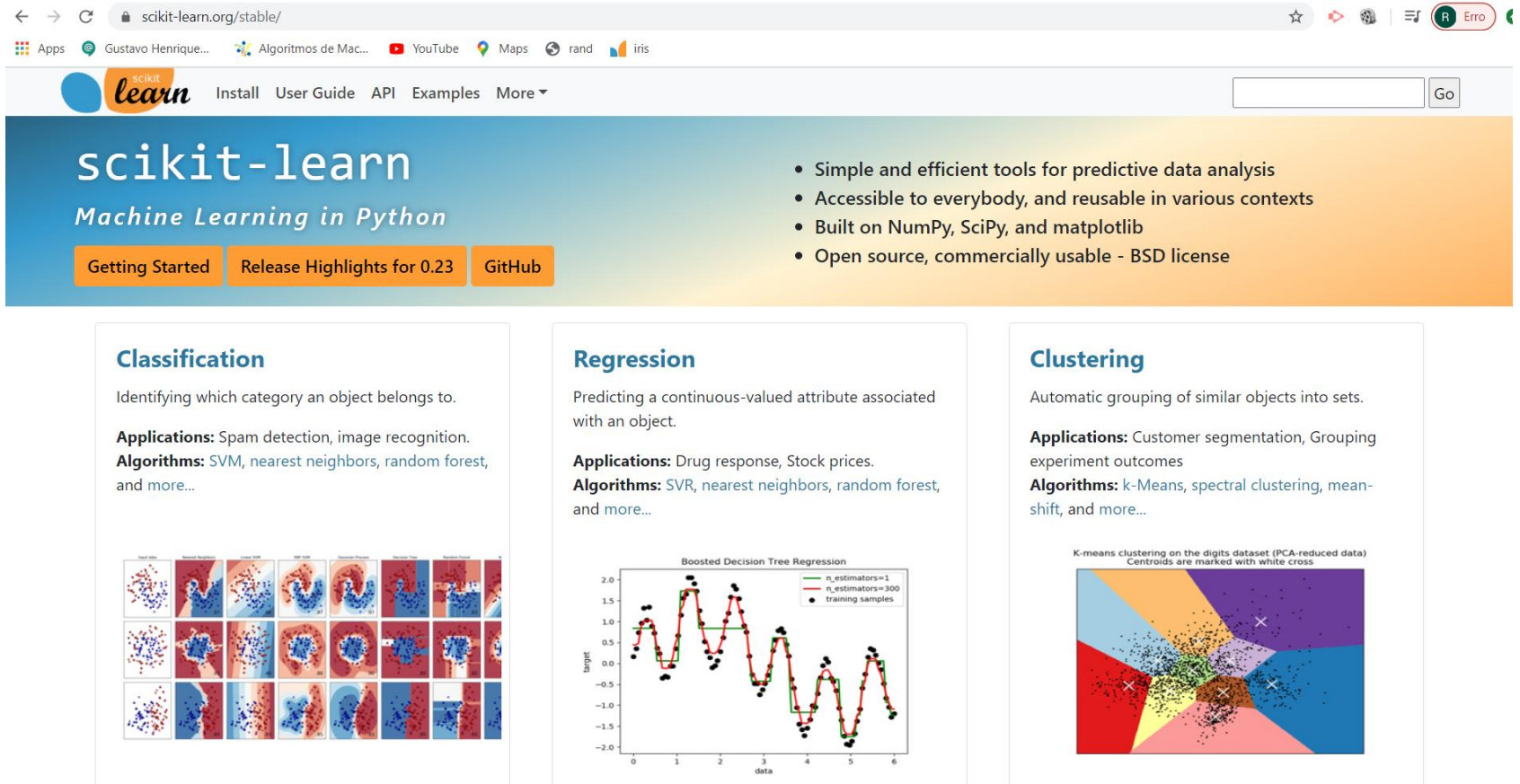


Figure 1: Typical Supervised Machine Learning Pipeline.

Bibliotecas em python

- https://scikit-learn.org/stable/
- pip install scikit-learn conda install scikit-learn



Referências

- ***Básica***
- James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R*, Springer Science+Business Media, New York. http://www-bcf.usc.edu/~gareth/ISL/index.html Download Livro
- Hastie, T., Tibshirani, R., Friedman, J. (2009). *Elements of Statistical Learning*, Second Edition, Springer Science+Business Media, New York. http://statweb.stanford.edu/~tibs/ElemStatLearn/
- Ethem Alpaydin. Introduction to Machine Learning: Adaptive Computation and Machine Learning series. MIT Press, 2014, ISBN 0262028182, 9780262028189.
- Shai Shalev-Shwartz, Shai Ben-David. Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press, 2014. ISBN 1107057132, 9781107057135
- Ulrike von Luxburg and Bernhard Scholkopf. STATISTICAL LEARNING THEORY: MODELS, CONCEPTS, AND RESULTS.

- ***Complementar***
- Christopher M. Bishop. Pattern Recognition and Machine Learning: Information Science and Statistics, Springer, 2006. ISSN 1613-9011. ISBN 0387310738, 9780387310732.