
Aprendizado

Detecção de Spam

Vamos pensar num algoritmo para resolver o problema de detecção de spams sem aprendizado de máquina!



Detecção de Spam

Vamos pensar num algoritmo para resolver o problema de detecção de spams sem aprendizado de máquina!

- 1) Identificar as características do spam. Palavras ou frases que costumam aparecer muito. Campo remetente, etc
- 2) Escrever o código de detecção para cada um dos padrões observados
- 3) Testar o programa até que ele estivesse bom o bastante

Figura 1: Abordagem Tradicional

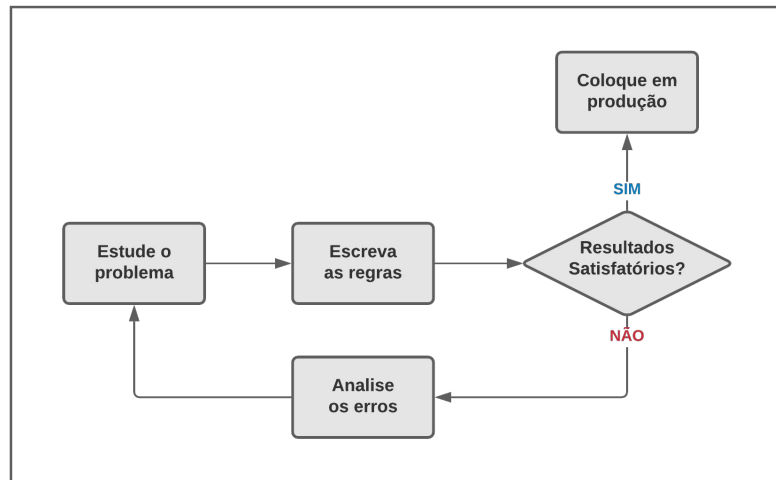


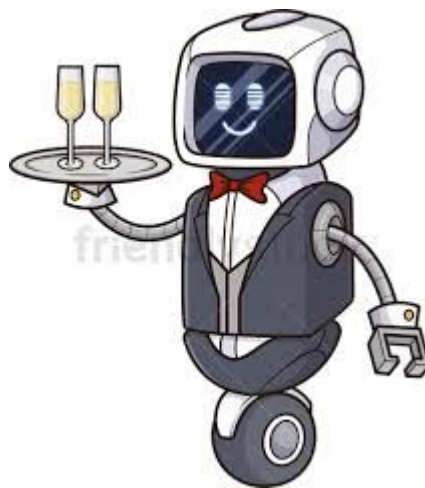
Imagem: Caíque Filipini (feito com Lucidchart)

Aprendizado de máquina

“Aprendizado de máquina é o campo de estudo que possibilita aos computadores a habilidade de aprender sem explicitamente programá-los”
[Arthur Samuel, 1959]

Um programa de computador aprende pela **experiência E** em relação a algum tipo de **tarefa T** e alguma medida de **desempenho P** se o seu desempenho em T, conforme medido por **P**, melhora com a experiência E. [Tom Mitchell, 1997]

Aprendizado de máquina



Aprendizado de máquina

- Desenvolver algoritmos que permitam ao computador aprender tarefas **específicas** a partir de exemplos de treinamento
- Aprendizado quer dizer que o computador não memoriza os exemplos vistos, mas pode generalizar
- Idealmente, o computador pode usar os exemplos para extrair uma regra geral de como a tarefa específica pode ser executada corretamente

Aprendizado de máquina

- Filtro de spam

[History of email spam - Wikipedia](#)

Aprendizado de máquina

- Uma abordagem baseada em AM aprende automaticamente quais palavras e frases são bons indicadores de spam

Figura 2: Abordagem com Machine Learning

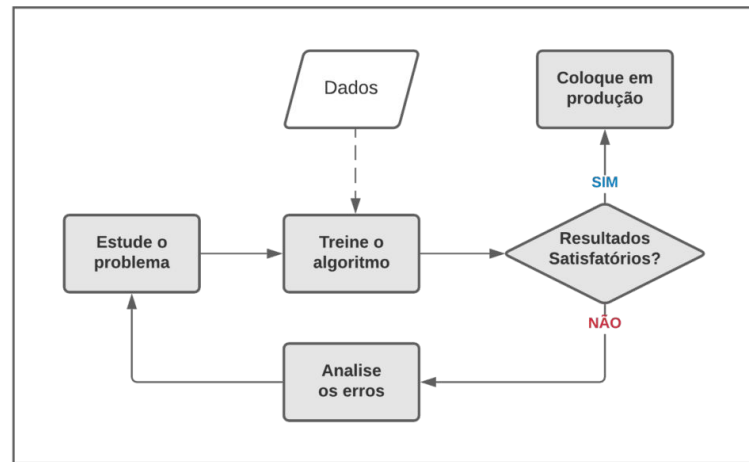


Imagem: Caíque Filipini (feito com Lucidchart)

Aprendizado de máquina

- A maioria dos problemas precisam ser retreinadas no futuro

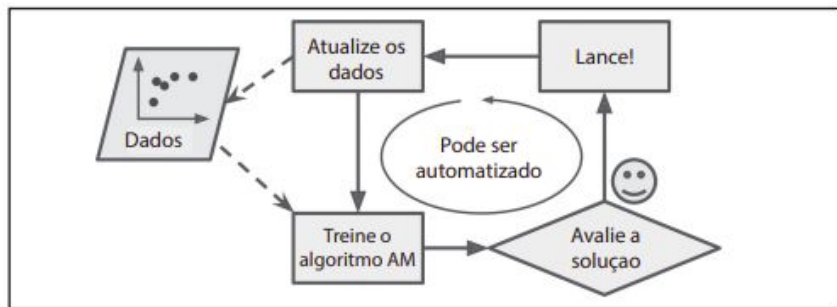


Figura 1-3. Adaptando-se automaticamente à mudança

Aprendizado de máquina

E muitas vezes aprendemos com os resultados dos modelos

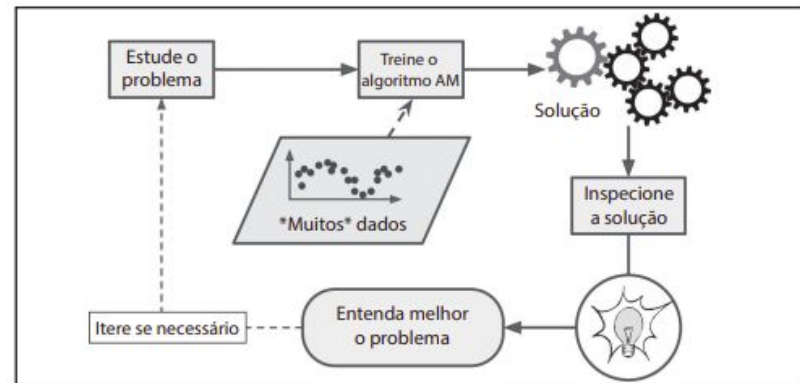


Figura 1-4. O Aprendizado de Máquina pode ajudar no ensino dos humanos

Aprendizado de máquina

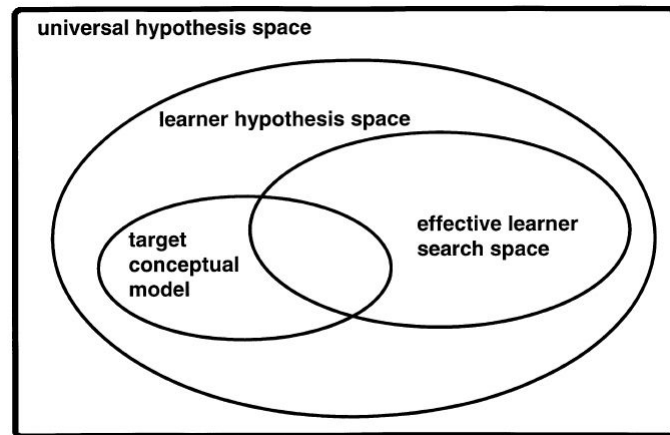
- É ótimo para:
 - Problemas para os quais as soluções atuais exigem muitos ajustes finos ou extensas listagens de regras
 - Problemas complexos para os quais não existe uma boa solução quando utilizamos uma abordagem tradicional
 - Adaptabilidade de ambientes
 - Entendimento de problemas complexos e grandes quantidades de dados

Tipos de Aprendizado

- Serem ou não treinados com supervisão humana (**aprendizado supervisionado, não supervisionado e semi supervisionado e aprendizado por reforço**)
- Se podem ou não aprender gradativamente em tempo real (**aprendizado online vs aprendizado em batch**)
- Se funcionam simplesmente comparando novos pontos de dados com pontos de dados conhecidos, ou se detectam padrões em dados de treinamento e criam um modelo preditivo, como cientistas (**aprendizado baseado em instâncias vs aprendizado baseado em modelo**)

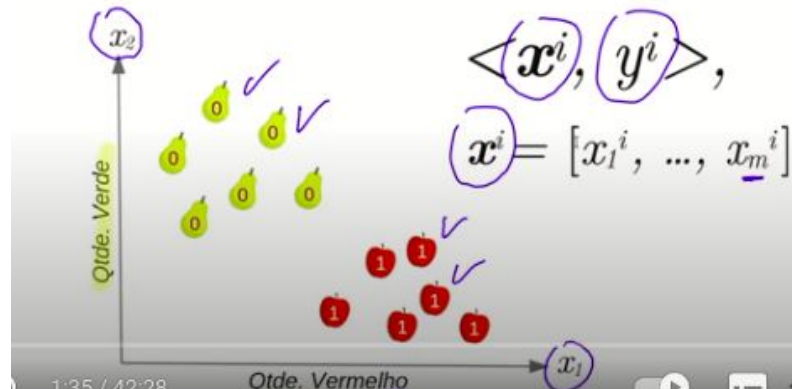
Aprendizado

- A aprendizagem é uma busca através do espaço de hipóteses possíveis por aquele que terá um bom desempenho, mesmo com novos exemplos além do conjunto de treinamento



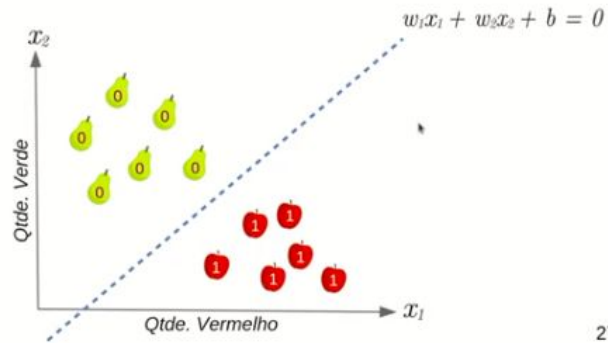
Viés indutivo - Exemplo

Espaço de *features*:



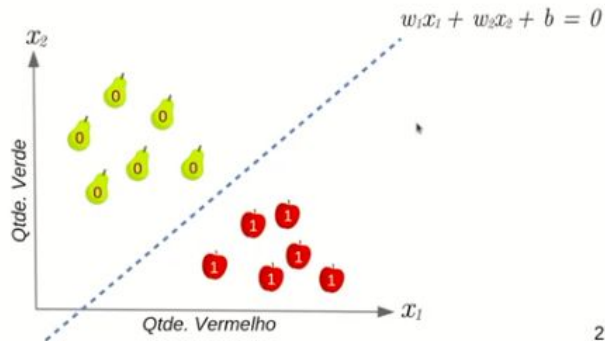
Viés indutivo - Exemplo

Hipótese como um separador linear



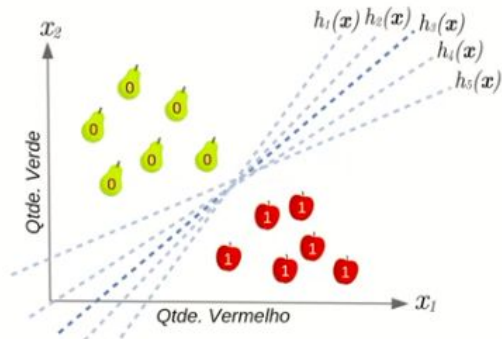
Viés indutivo - Exemplo

Hipótese como um separador linear

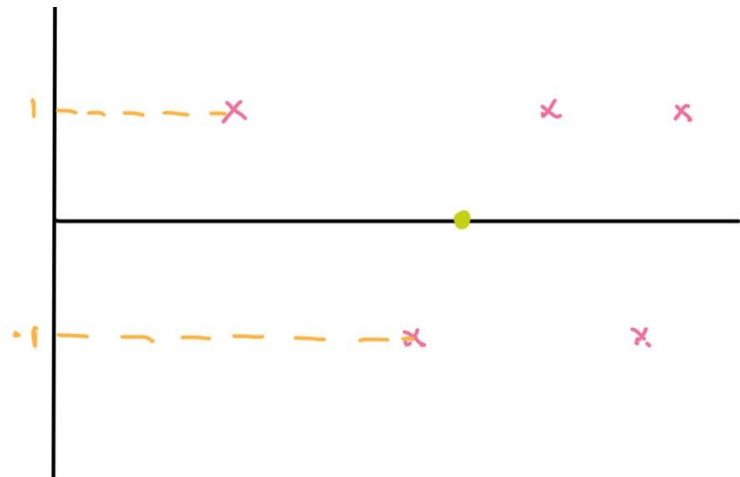


27

Gerando hipóteses com base em X



Viés indutivo - Exemplo

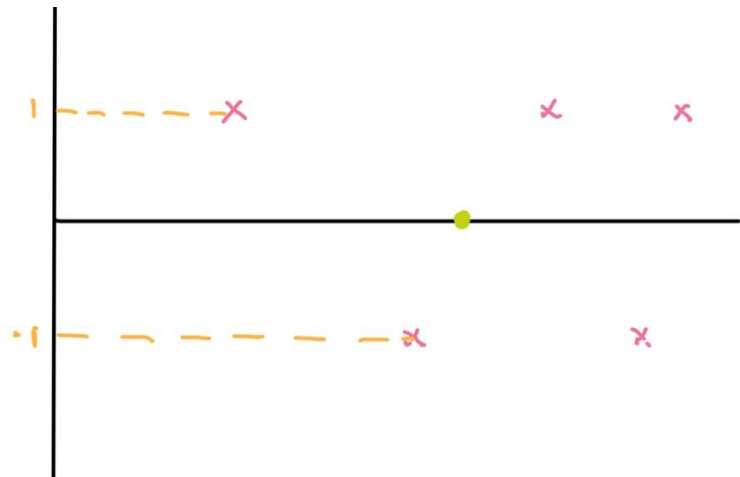


Espaço de entradas: $X = \{0.01, 0.02, \dots, 1\}$

Objetivos: Aprender uma função baseada nos exemplos

Sem bias: cada função $f: X \rightarrow y$ pode estar correta 2^{100}

Viés indutivo - Exemplo



Espaço de entradas: $X = \{1, 2, \dots, 100\}$

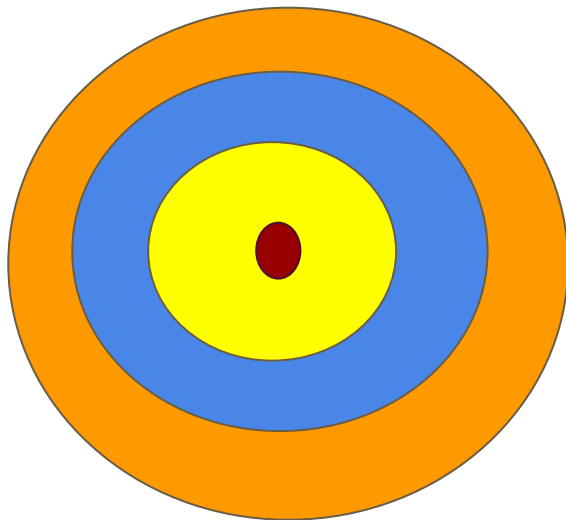
Objetivos: Aprender uma função baseada nos exemplos

Sem bias: cada função $f: X \rightarrow y$ pode estar correta: 2^{100} possibilidades

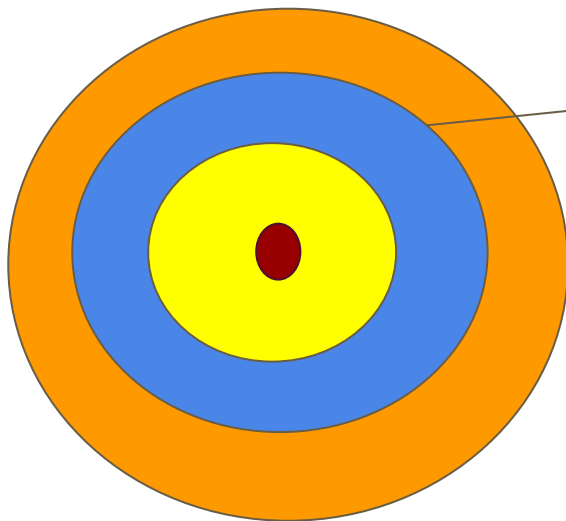
Agora nosso espaço de hipóteses é uma constante $f = \{0, 1\}$

Logo, podemos observar que a função é 1

Viés indutivo

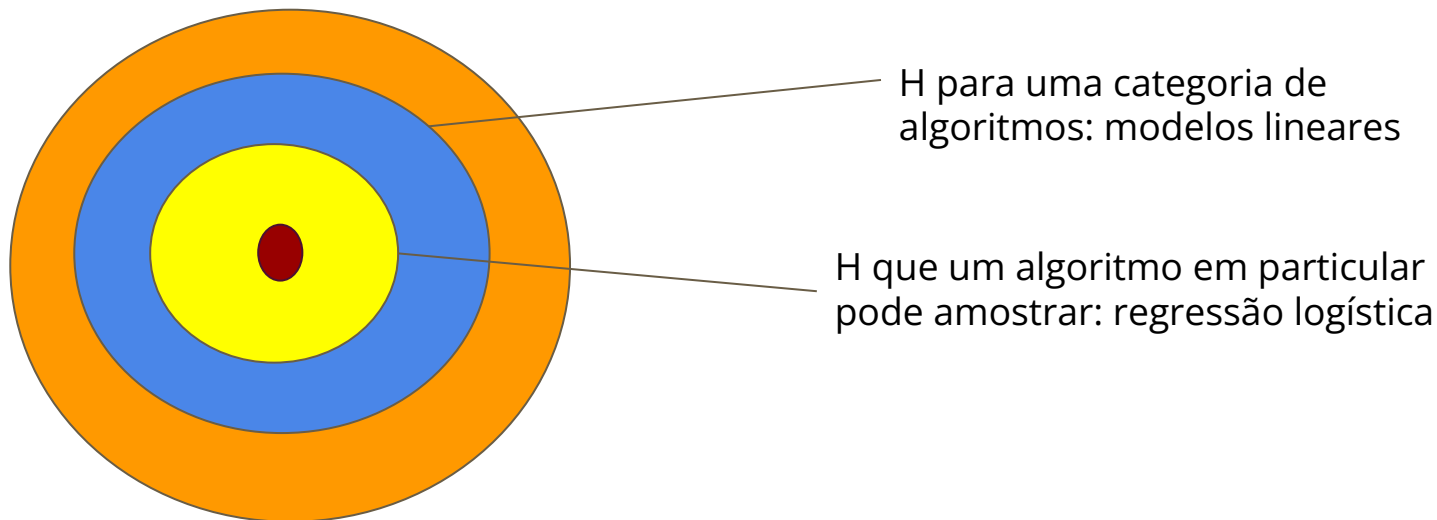


Viés indutivo

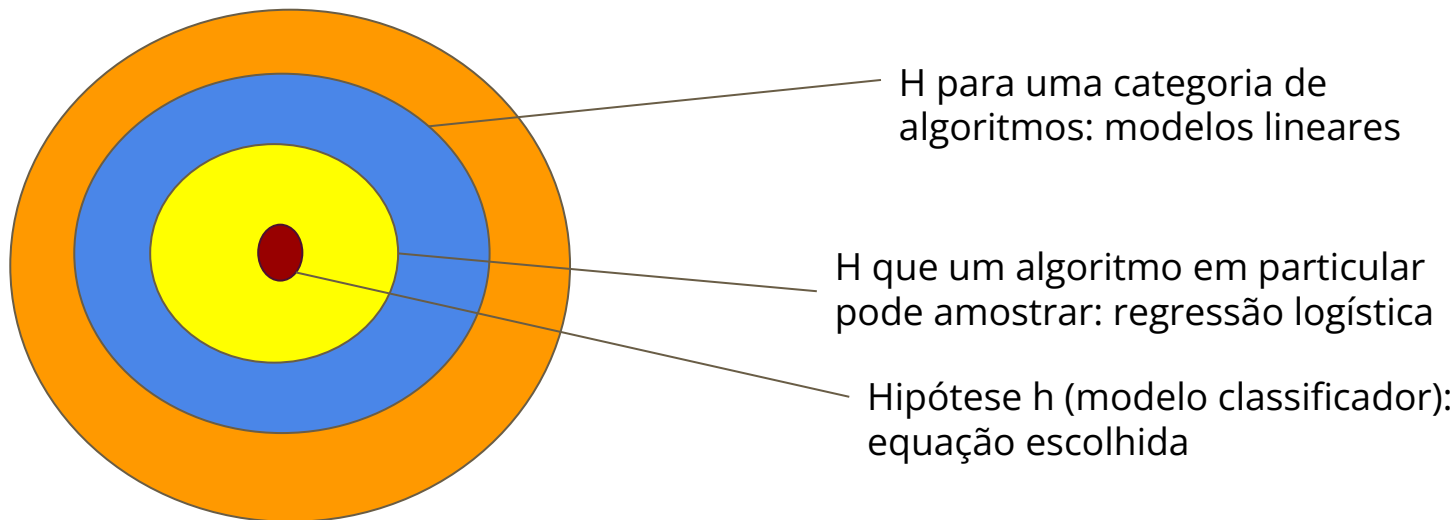


H para uma categoria de
algoritmos : modelos lineares

Viés indutivo



Viés indutivo



Aprendizado de máquina

- Ajuste de uma função de uma única variável a alguns pontos de dados
- Como escolhermos entre várias hipóteses consistentes?

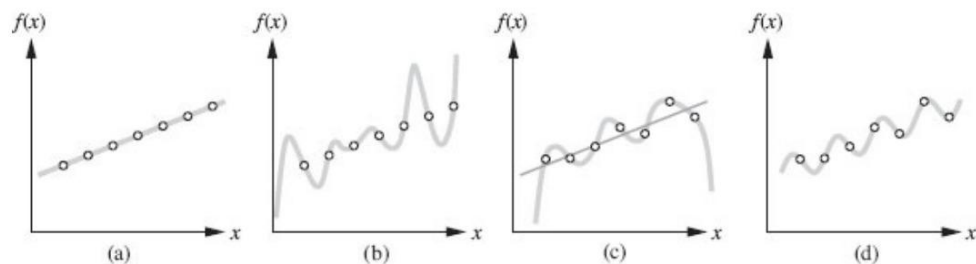


Figura 18.1 (a) Exemplo de pares $(x, f(x))$ e uma hipótese linear consistente. (b) Hipótese de polinômio de grau 7 consistente para o mesmo conjunto de dados. (c) Conjunto de dados diferente que admite um ajuste de polinômio de grau 6 exato ou um ajuste linear aproximado. (d) Um simples ajuste senoidal exato para o mesmo conjunto de dados.

Aprendizado de máquina

- Ajuste de uma função de uma única variável a alguns pontos de dados
- Como escolhermos entre várias hipóteses consistentes?
- **É preferível escolher entre a mais simples!**
- **Navalha de Ockam:** devido ao filósofo Guilherme Ockam, que a usou para argumentar contra todo tipo de complicação

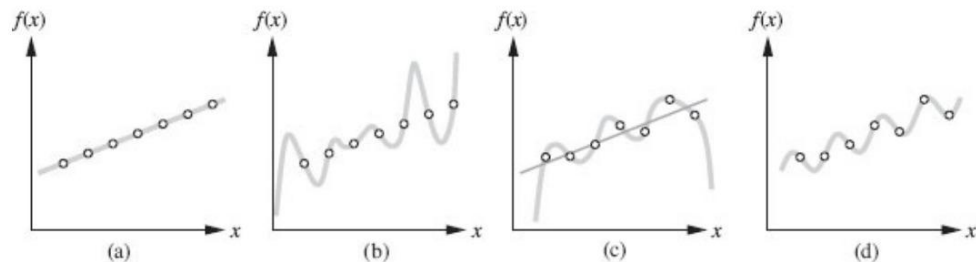


Figura 18.1 (a) Exemplo de pares $(x, f(x))$ e uma hipótese linear consistente. (b) Hipótese de polinômio de grau 7 consistente para o mesmo conjunto de dados. (c) Conjunto de dados diferente que admite um ajuste de polinômio de grau 6 exato ou um ajuste linear aproximado. (d) Um simples ajuste senoidal exato para o mesmo conjunto de dados.

Aprendizado supervisionado

- Dizemos que um problema de aprendizagem é realizável se o espaço de hipóteses contiver a função verdadeira.
- Nem sempre podemos dizer se um problema de aprendizagem é realizável, pois a função verdadeira não é conhecida

Teoria do Aprendizado

Premissas

- Distribuição qualquer $P(x,y)$
- Amostras de $P(x,y)$ de forma igualmente distribuída
- Exemplos independentes (amostrar de qualquer região do espaço, ex de amostragem com reposição)
- Rótulos/labels podem ter ruídos
- $P(x,y)$ deve ser estática (não mudar ao longo do tempo)
- $P(x,y)$ é desconhecida durante o treinamento

Aprendizado supervisionado

Dado um conjunto de treinamento de N pares de exemplos de entrada e saída

$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$

onde cada y_j , foi gerado por uma função desconhecida $y = f(x)$,

descobrir uma função h que se aproxime da função verdadeira f .

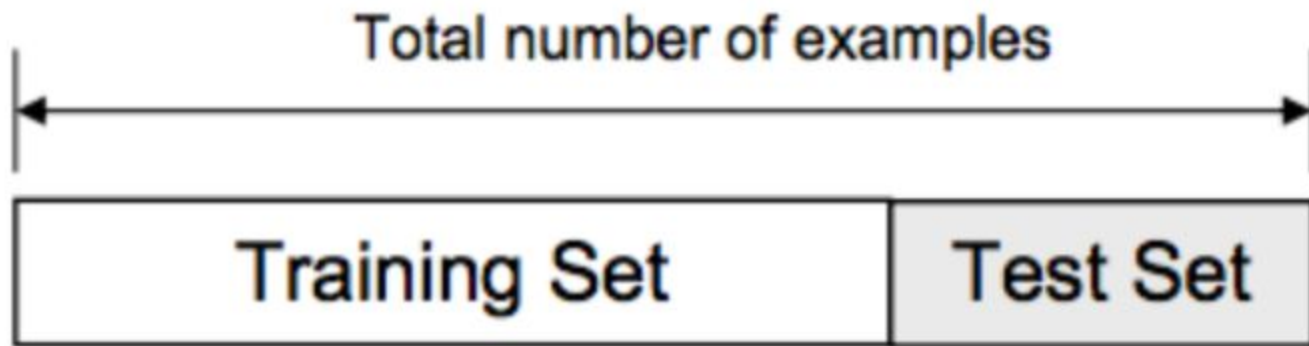
Aprendizado supervisionado

- x e y podem ter qualquer valor, não precisando ser números;
- y é o que chamamos de classe ou rótulo
- A função h é uma hipótese

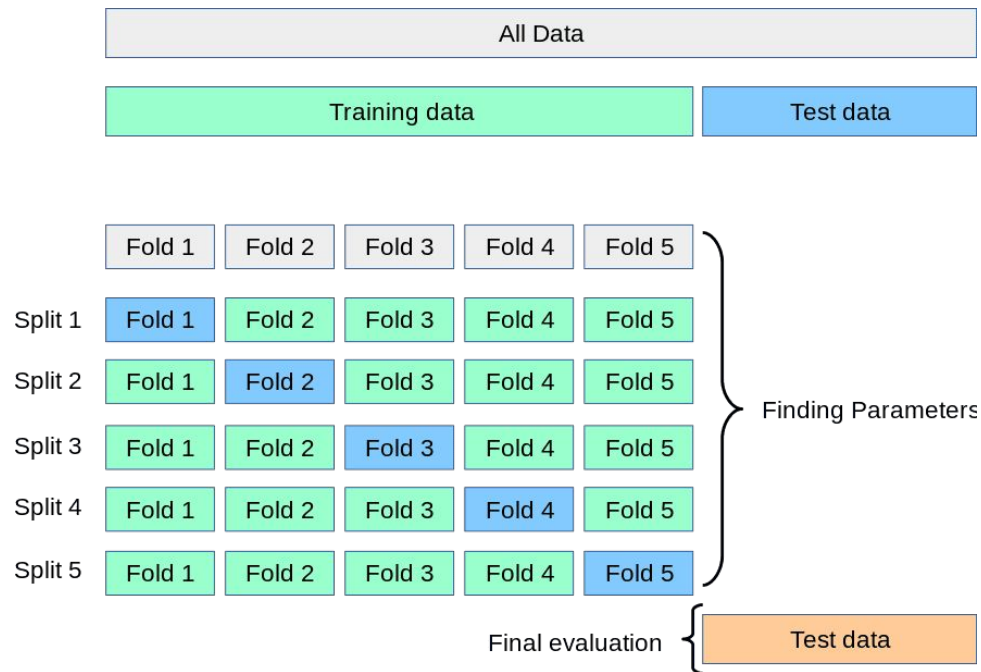
Aprendizado supervisionado

- Para medir a precisão de uma hipótese, fornecemos um conjunto de testes de exemplos que são distintos do conjunto de treinamento
- Dizemos que uma hipótese generaliza bem se prevê corretamente o valor de y para novos exemplos

(Validação)



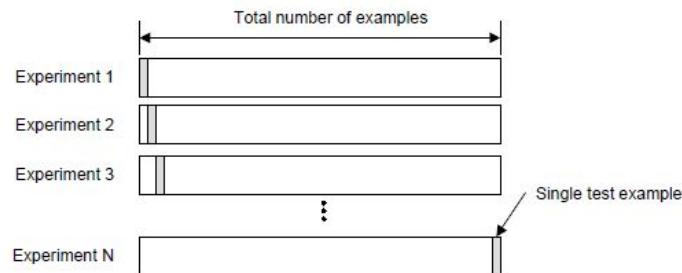
(Validação)



(Validação)

Leave-one-out Cross Validation

- Leave-one-out is the degenerate case of K-Fold Cross Validation, where K is chosen as the total number of examples
 - For a dataset with N examples, perform N experiments
 - For each experiment use N-1 examples for training and the remaining example for testing

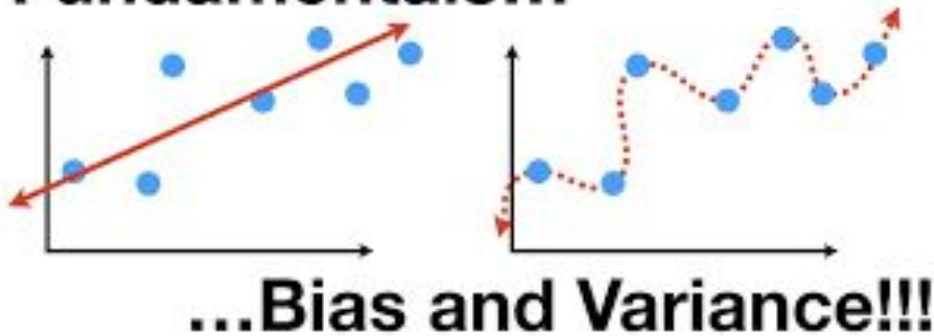


- As usual, the true error is estimated as the average error rate on test examples

$$E = \frac{1}{N} \sum_{i=1}^N E_i$$

(Viés e Variância)

Machine Learning Fundamentals...

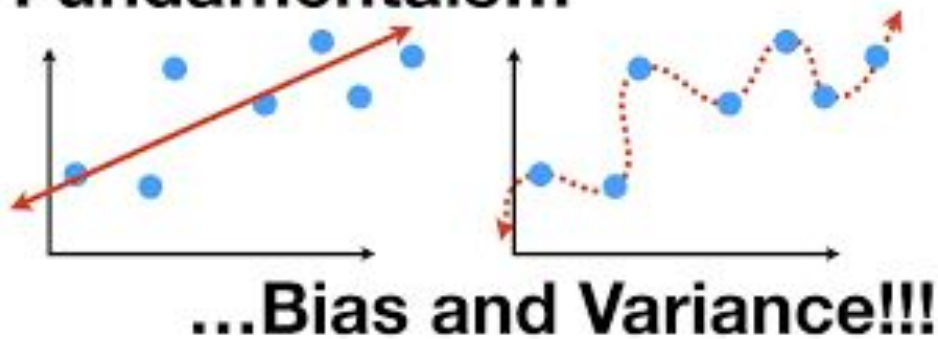


Alto viés: modelo não está aprendendo

Alta variância: modelo aprende exatamente os pontos de treinamento

(Viés e Variância)

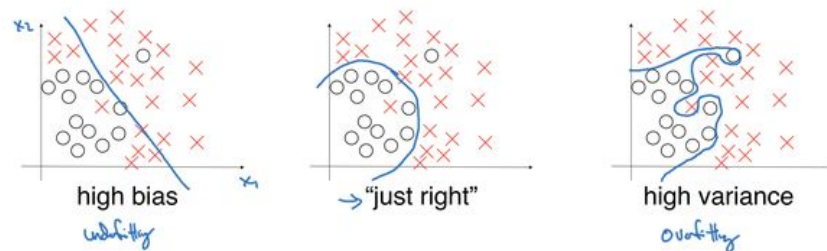
Machine Learning Fundamentals...



Alto viés: modelo não está aprendendo

Alta variância: modelo aprende exatamente os pontos de treinamento

Bias and Variance



Aprendizado supervisionado

- Quando y for de um conjunto finito de valores (ensolarado, nublado ou chuvoso), o problema de aprendizagem será chamado de classificação
- Será chamado de classificação binária se houver apenas dois valores (spam e não spam)



Figura 1-5. Um conjunto de treinamento rotulado para aprendizado supervisionado (por exemplo, classificação de spam)

Aprendizado supervisionado

- Quando y for um número (temperatura de amanhã), o problema de aprendizagem será chamado de regressão
- Tecnicamente, a solução de um problema de regressão é encontrar uma expectativa condicional ou valor médio de y porque a probabilidade de acharmos exatamente o número real certa para y é 0

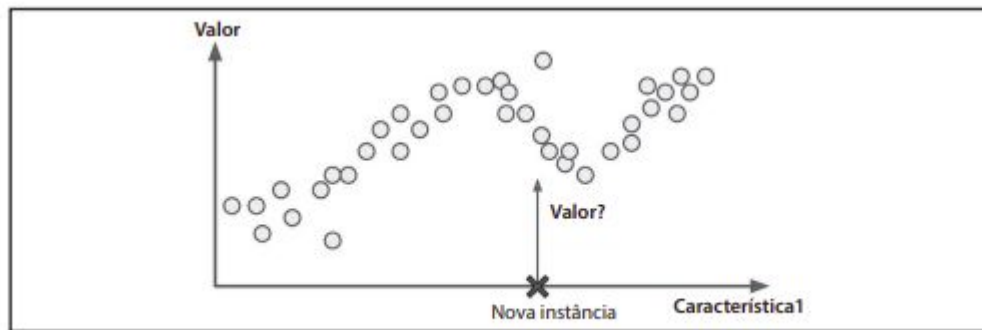


Figura 1-6. Regressão

Aprendizado supervisionado

- KNN
- Regressão linear
- Regressão logística
- SVM
- Árvore de Decisão e Random Forest
- Redes Neurais

Aprendizado não supervisionado

- Os dados de treinamento não são rotulados



Figura 1-7. Conjunto de treinamento não rotulado para aprendizado não supervisionado

Aprendizado não supervisionado

- Tarefa de detecção de grupos de visitantes que sejam semelhantes
- Resultados que podem ser encontrados:
 - 40% dos visitantes são homens que adoram histórias em quadrinhos e em geral leem a noite
 - 20% são jovens, amantes de ficção científica e o visitam durante o final de semana

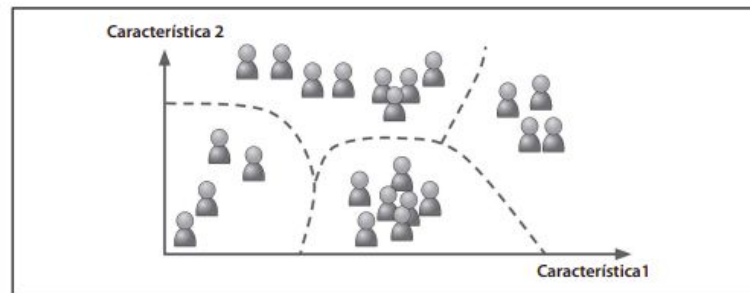


Figura 1-8. Clustering

Aprendizado não supervisionado

- Visualização: método estatístico para visualizar dados de alta dimensão, dando a cada ponto de dados uma localização em um mapa bidimensional ou tridimensional.

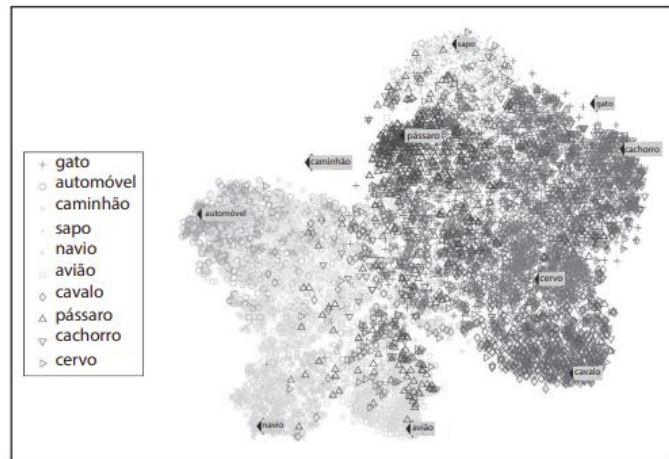


Figura 1-9. Exemplo de uma visualização t-SNE destacando grupos semânticos³

Aprendizado não supervisionado

- Detecção de anomalias
 - transações incomuns em cartões de crédito
 - defeitos de fabricação
 - outlier em um conjunto de dados
- Detecção de novidade
 - detectar instâncias novas que parecem diferentes de todas as instâncias no conjunto de treinamento

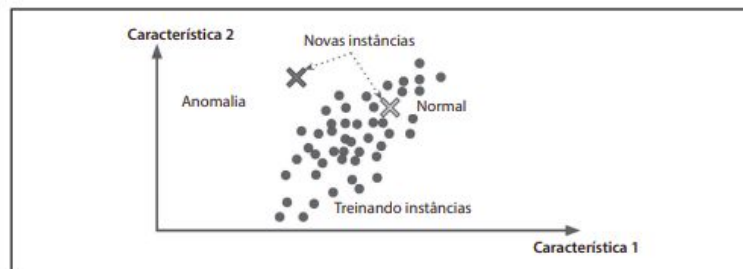


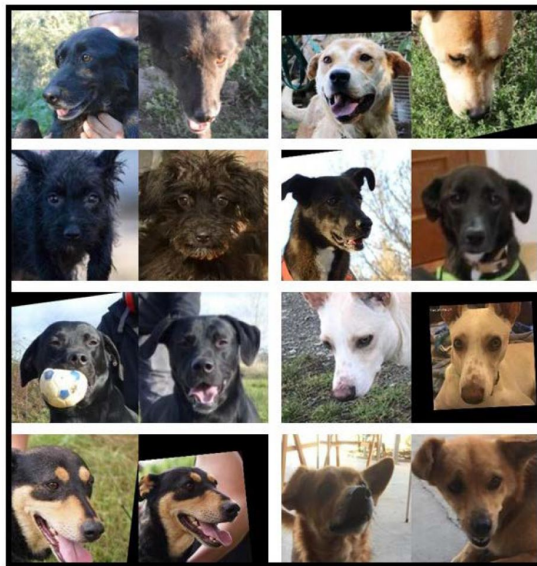
Figura 1-10. Detecção da anomalia

Aprendizado não supervisionado: novidade x anomalia

- Ex do Chihuahua



(a)



(b)

Se existe um dataset com apenas um 1% de chihuahuas, o algoritmo de detecção de novidade não podem detectar o chihuahua como novidade, pois ele já existe na base, no entanto, um algoritmo de detecção de anomalia....

Aprendizado semissupervisionado

Same or different person?



Same



Different



Not sure

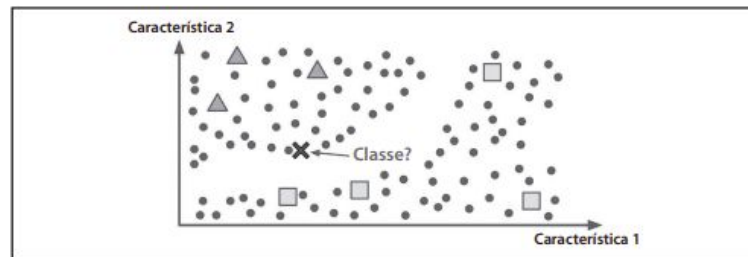


Figura 1-11. Aprendizado Semissupervisionado

[Previous](#)

[Done](#)

Aprendizado por reforço

- O sistema de aprendizado nesse caso é chamado de agente
- Pode assistir o ambiente, selecionar e executar ações e obter recompensas em troca ou penalidades

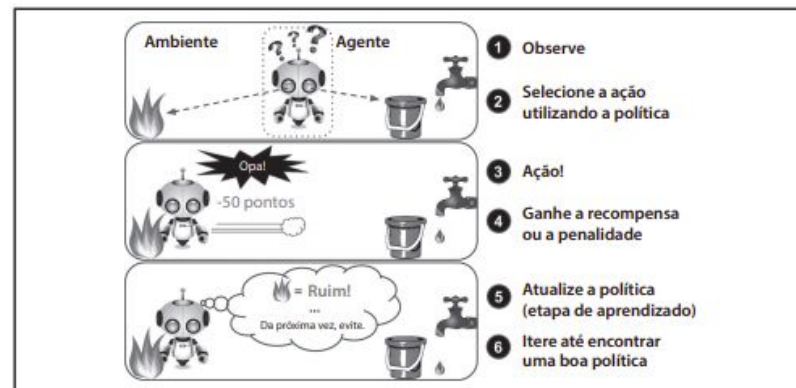


Figura 1-12. Aprendizado por reforço

<https://www.sciencedirect.com/science/article/pii/S0952197622000598>

Dica de estudos

CIENTISTA de DADOS um guia de estudos

Seja **especialista** em um assunto.

1. Conheça bem poucos algoritmos, escolha os melhores.

Estude **programação**

Estude matemática e **estatística**.

Estudo as classes de algoritmos

1. Classificação / Regressão / cluster
2. Redução de dimensão
3. Representação de dados

Pegue um algoritmo por vez:

1. Qual a função custo?
2. Quais os parâmetros e o que significam?
3. Quais as métricas de avaliação?
4. Quais os prós e contras?
5. Como você compara com outros algoritmos?

Aplique em um problema real




1. O que seu modelo agrega?
2. O que a empresa pode ganhar com o modelo?
3. Como você implantaria sua solução

 By <https://www.linkedin.com/in/jeleandro/>

CIÊNCIA de DADOS um guia de estudos

 By <https://www.linkedin.com/in/jeleandro/>

Machine learning by family - part 1

	Algorithm	Math needed
Trees 	<ol style="list-style-type: none">1. Decision Tree2. Random Forest3. Boosting algorithms (xgboost, etc)4. Local outlier factor5. Extra tree	<ol style="list-style-type: none">1. Basic statistics / percentile2. Bootstrap3. Information Theory4. Distance functions
Gradient 	<ol style="list-style-type: none">1. GLM in general<ol style="list-style-type: none">1. Linear regression. (OLS)2. Logistic regression (Logit)2. Neural Nets<ol style="list-style-type: none">1. Multi-Layer perceptron2. Recurrent neural net family	<ol style="list-style-type: none">1. Calculus<ol style="list-style-type: none">1. Gradient / Hessian2. Deep multivariate statistical analysis3. Sequence prediction problem
Distance 	<ol style="list-style-type: none">1. KNN2. DbSCAN3. K-Means / K-Medoids4. SVM5. hierarchical clustering	<ol style="list-style-type: none">1. Distance functions2. Divide and conquer3. Margin maximization

Aprendizado em batch e online

Será que o sistema pode aprender de forma incremental a partir da entrada de um fluxo de dados?

Aprendizado em batch e online

- Em batch (por ciclo)
- Sistema incapaz de aprender de forma incremental: deve ser treinado usando todos os dados disponíveis
- Normalmente realizado offline. **Porque fazer dessa forma?**
- O sistema é treinado, em seguida é implementado em produção e roda sem aprender mais nada, somente aplicando o que aprendeu
- **E como fazer para esse tipo de sistema tenha acesso a novos dados?**

Aprendizado online (incremental)

- É possível treinar o sistema incrementalmente, fornecendo as instâncias de dados de forma sequencial, individual ou em pequenos grupo, chamados mini-batches
- Cada etapa do aprendizado é rápida e tem um custo baixo, assim o sistema pode aprender instantaneamente os dados novos em tempo real

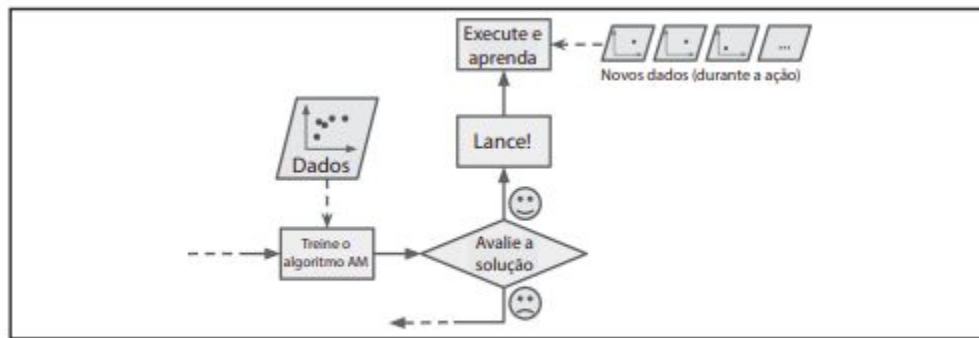


Figura 1-13. Aprendizado online

Aprendizado online (incremental)

- Excelente para sistemas que recebem dados em um fluxo contínuo (preço das ações) e precisam se adaptar a mudanças rápido ou autonomamente
- Ideal quando se tem recursos computacionais limitados
- Mas também pode ser usado para treinar sistemas em grandes conjuntos de dados

Aprendizado online (incremental)

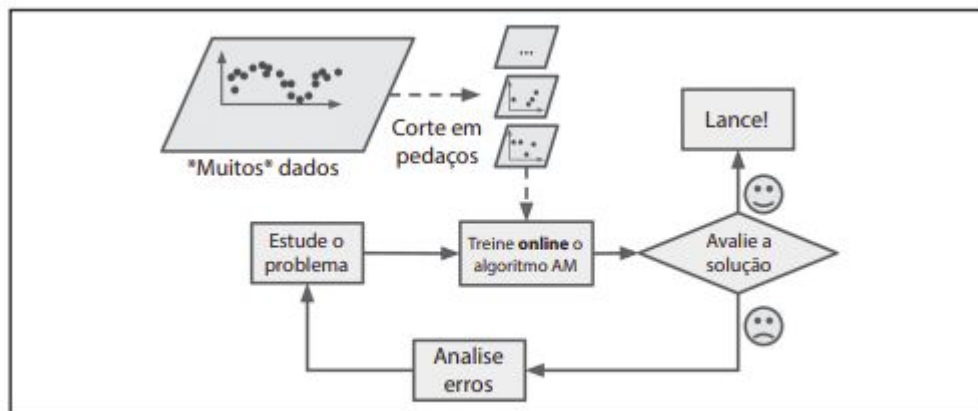


Figura 1-14. Utilizando aprendizado online para lidar com grandes conjuntos de dados

Aprendizado online (incremental)

- Taxa de aprendizagem
- Dados ruins
- Percepção do usuário

Aprendizado baseado em instâncias versus aprendizado baseado em modelo

- Outro meio de categorizar os algoritmos de aprendizado de máquina é por meio da generalização
- Predições: dada uma série de exemplos de treinamento, o sistema precisa ser capaz de fazer boas predições para (generalizar) exemplos que nunca viu
- Um bom rendimento de desempenho nos dados de treinamento é ótimo, mas o verdadeiro objetivo é ter um bom desempenho em instâncias novas.

Aprendizado baseado em instância

- Memorização
 - filtro de spam que sinalizaria todos os e-mails idênticos aos já marcados pelo usuário
- Similaridade
 - e-mails semelhantes aos e-mails conhecidos como spam

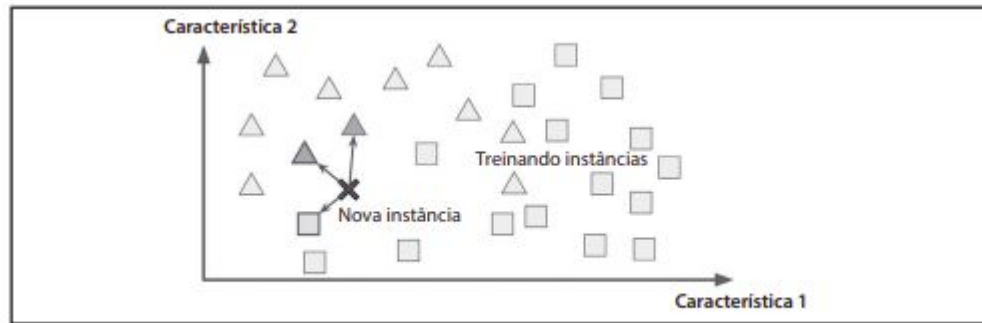


Figura 1-15. Aprendizado baseado em instância

Aprendizado baseado em modelo

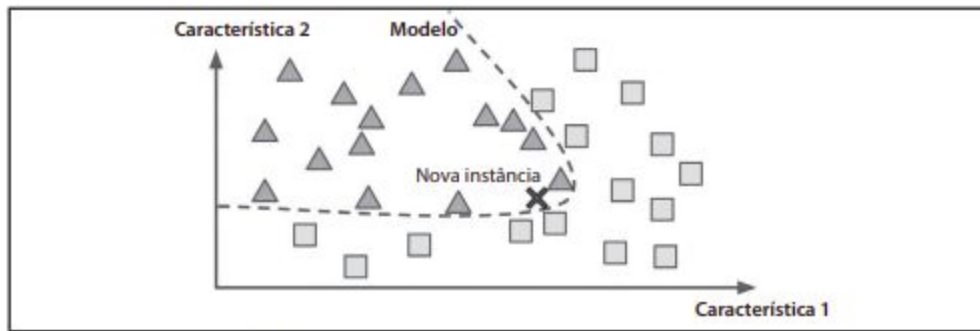


Figura 1-16. Aprendizado baseado em modelo

Aprendizado baseado em modelo

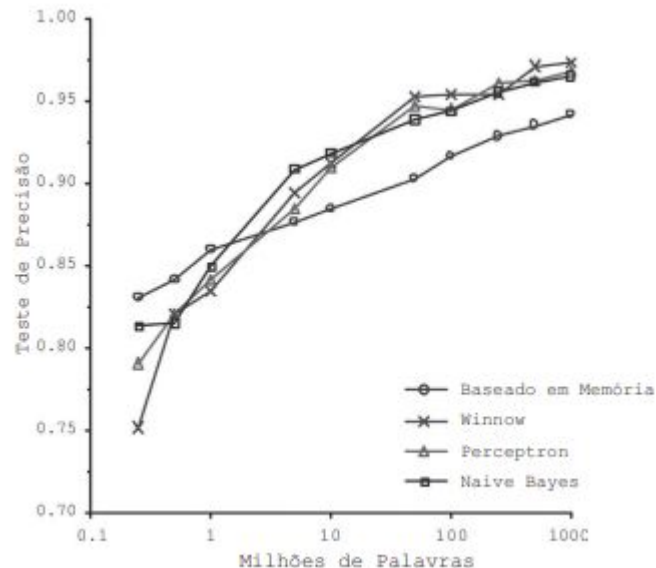
- $\text{satisfacao_vida} = \text{teta0} + \text{teta1} \times \text{PIB_per_capita}$
- função de avaliação ou função de custo

Tabela 1-1. O dinheiro torna as pessoas mais felizes?

País	PIB per capita (USD)	Satisfação de vida
Hungria	12.240	4,9
Coreia	27.195	5,8
França	37.675	6,5
Austrália	50.962	7,3
Estados Unidos	55.805	7,2

Desafios do aprendizado de máquina

- Quantidade insuficiente de dados de treinamento
- Dados de treinamento não representativos
- Dados de baixa qualidade
- Características irrelevantes



Resumo

- Aprendizado de máquina é garantir que as máquinas evoluem em algumas tarefas aprendendo com os dados, em vez de ter que programar explicitamente as regras
- Existem muitos tipos diferentes de sistemas de AM
- Em um projeto de AM, você coleta dados em um conjunto de treinamento e os fornece para um algoritmo de aprendizado. Se o algoritmo for baseado em modelo, ele ajusta alguns parâmetros para adequar o modelo ao conjunto de treinamento. Se o algoritmo for baseado em instância, ele simplesmente aprende os exemplos de cor e salteado e utiliza uma medida de similaridade para generalizar em instâncias novas
- O sistema não terá um bom desempenho se o conjunto de treinamento for muito pequeno ou se os dados não forem representativos, ruidosos ou poluídos com características irrelevantes
- O modelo não precisa ser simples demais nem muito complexo