

INTRODUÇÃO A APRENDIZADO DE MÁQUINA

Edian F. Franco De Los Santos

O que é aprendizado de máquina?

Lorena, et al. (2011): Algoritmos de AM induzem uma função ou hipótese capaz de resolver o problema a partir de exemplos (instâncias) do problema a ser resolvido .

O machine learning é uma forma de inteligência artificial que permite que um sistema aprenda a partir de dados, e não através de programação explícita. No entanto, o machine learning não é um processo simples. Como os algoritmos ingerem dados de treinamento, é possível produzir modelos mais precisos com base nesses dados.

Premissa de aprendizado de máquina

Tom Mitchell (1998) coloca muito bem o conceito de Aprendizado de Máquina:

Um programa de computador aprende a partir da Experiência **E**, em relação a uma classe de tarefas **T**, com medida de desempenho **P**, se seu desempenho em **T**, medido por **P**, melhora com **E** .



Exemplo de programa de AM

Vamos supor o problema de aprender a jogar damas. Identifique a tarefa **T**, a experiência **E** e o desempenho **P** para este problema

- Jogar damas → Tarefa
- Prática de jogo → Experiência
- Porcentagem de vitórias contra oponentes → Desempenho



Identificando Spam

Vamos supor que seu programa de e-mail favorito “observa” quais e-mails você marca ou não marca como spam. Em seguida com base em suas observações (aprendizado) ele consegue uma forma de melhorar o filtro de spam. Defina qual a tarefa T , a experiência E e o desempenho P para o cenário.

- Classificar e-mails como spam ou não spam → Tarefa
- O número de e-mails corretamente classificados como spam e não spam. → Desempenho
- Observar o conjunto de exemplos de spam e não spam Experiencias → Experiência

Aprendizado por Inferência indutiva

- Inferência " Ato de derivar conclusões a partir do conhecimento e de evidências disponíveis.
- Na inferência indutiva “Conclusões são derivadas da observação sistemática de fenômenos empíricos e de experimentos”
- Em outras palavras AM: Faz-se um raciocínio do geral para o particular

Id.	Nome	Idade	Sexo	Peso	Manchas	Temp.	# Int.	Est.	Diagnóstico
4201	João	28	M	79	Concentradas	38,0	2	SP	Doente
3217	Maria	18	F	67	Inexistentes	39,5	4	MG	Doente
4039	Luiz	49	M	92	Espalhadas	38,0	2	RS	Saudável
1920	José	18	M	43	Inexistentes	38,5	8	MG	Doente
4340	Cláudia	21	F	52	Uniformes	37,6	1	PE	Saudável
2301	Ana	22	F	72	Inexistentes	38,0	3	RJ	Doente
1322	Marta	19	F	87	Espalhadas	39,0	6	AM	Doente
3027	Paulo	34	M	67	Uniformes	38,4	2	GO	Saudável

PODER NOS DADOS

Cada coluna é uma característica (atributo, parâmetro, campo ou variável) que descreve os principais aspectos de um objeto

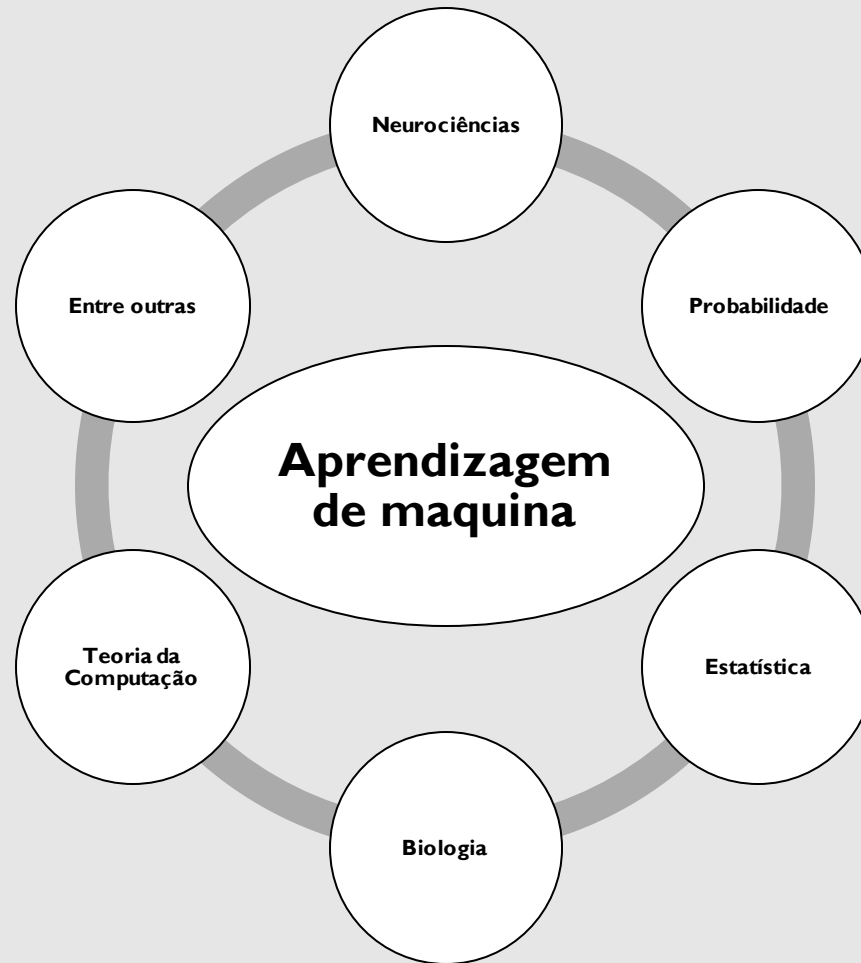


Id.	Nome	Idade	Sexo	Peso	Manchas	Temp.	# Int.	Est.	Diagnóstico
4201	João	28	M	79	Concentradas	38,0	2	SP	Doente
3217	Maria	18	F	67	Inexistentes	39,5	4	MG	Doente
4039	Luiz	49	M	92	Espalhadas	38,0	2	RS	Saudável
1920	José	18	M	43	Inexistentes	38,5	8	MG	Doente
4340	Cláudia	21	F	52	Uniformes	37,6	1	PE	Saudável
2301	Ana	22	F	72	Inexistentes	38,0	3	RJ	Doente
1322	Marta	19	F	87	Espalhadas	39,0	6	AM	Doente
3027	Paulo	34	M	67	Uniformes	38,4	2	GO	Saudável

Cada linha é um dado (objeto, instância, exemplo, padrão ou registro)

Atributo ou parâmetro de saída (alvo): presente em algumas tarefas (ex. Classificação), seus valores devem ser estimados usando outros atributos

AM é uma área multidisciplinar



Tipos de aprendizado de maquina

Supervisionado

- Preditivo
- Encontrar uma função o modelo que posso prever um rotulo o valor.
- Classe associada

Não supervisionado

- Descritivos
- Explorar ou descrever um conjunto de dados.
- Classe não associada

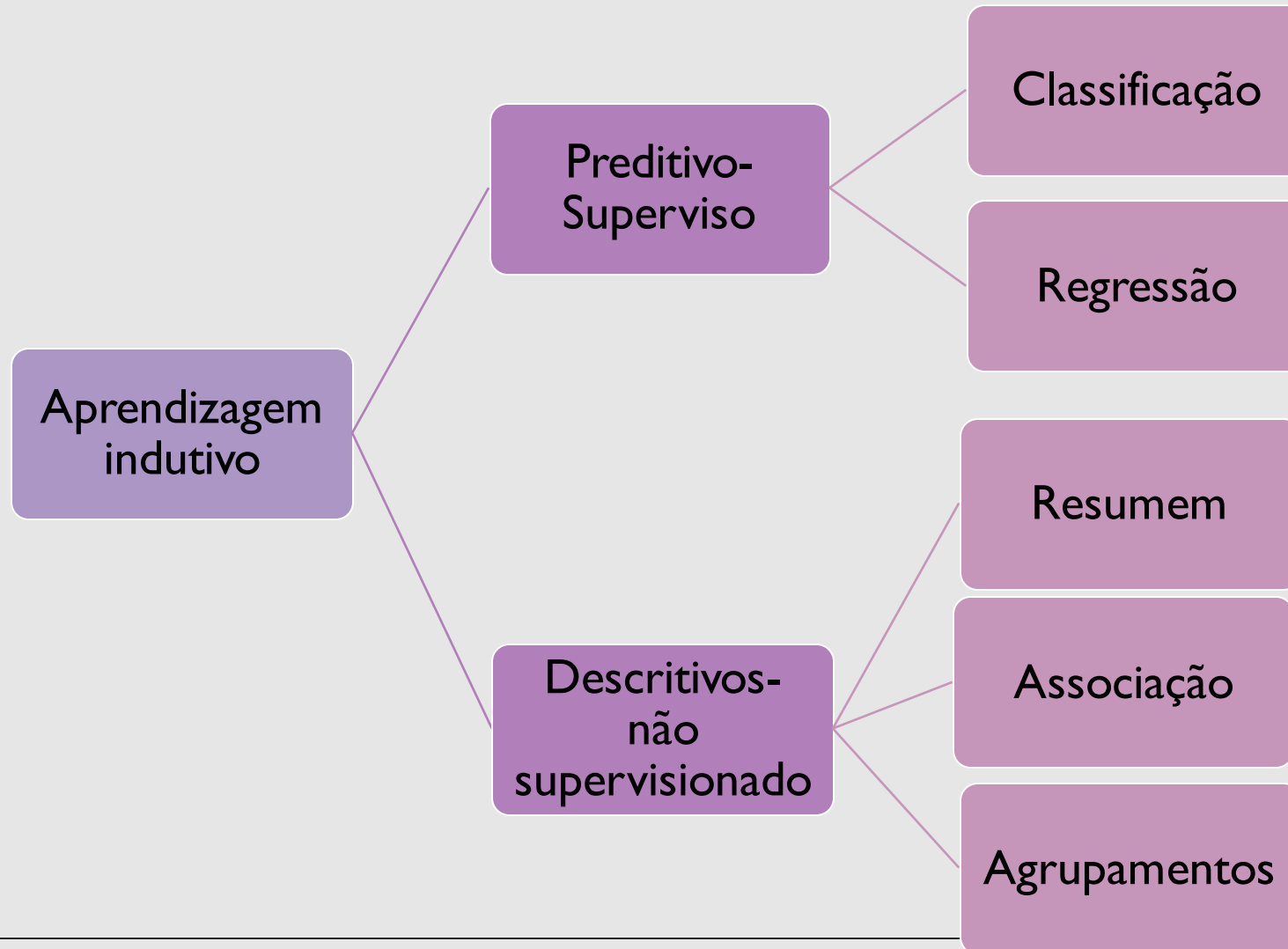
Aprendizado por reforço

- É um modelo de aprendizado comportamental
- O algoritmo recebe feedback da análise de dados, para o melhorar os resultado.

Deep Learning

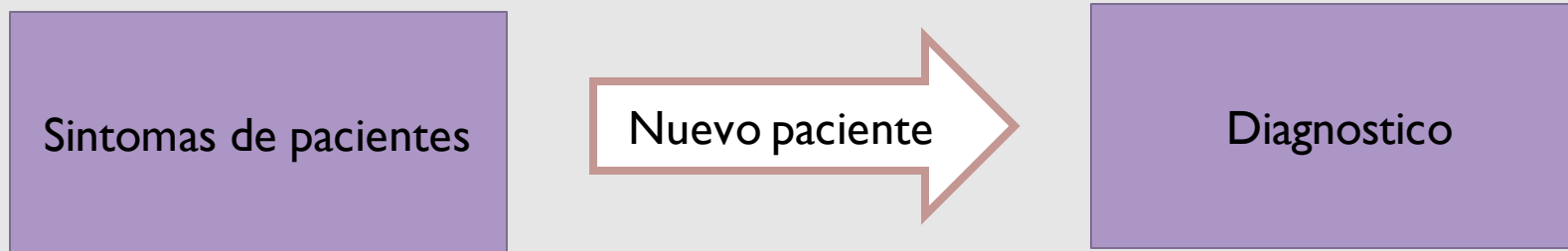
- É um método específico de aprendizado de máquina que incorpora redes neurais.
- Útil quando você está tentando aprender padrões de dados não estruturados

Hierarquia de AM



Aprendizado Supervisionado

- O algoritmo de aprendizado (indutor) recebe um conjunto de exemplos de treinamento para os quais os rótulos da classe associada são conhecidos.



- Métodos supervisionados distinguem pelo tipo dos rótulos dos dados
- Rótulos discretos (**classificação**) Ex: diagnóstico, bom/mau pagador, etc.
- Rótulos contínuos (**regressão**)

Ex: Vendas de uma loja, Previsão de séries temporais

Input data



Annotations

These are
apples



Model



Prediction

Its an
apple!

Aprendizado Não - Supervisionado

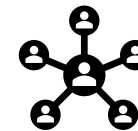
Exploram regularidades nos dados não fazendo uso de atributos de saída



Sumarização: encontrar uma descrição compacta para os dados. Ex: Media, Mediana, porcentagem

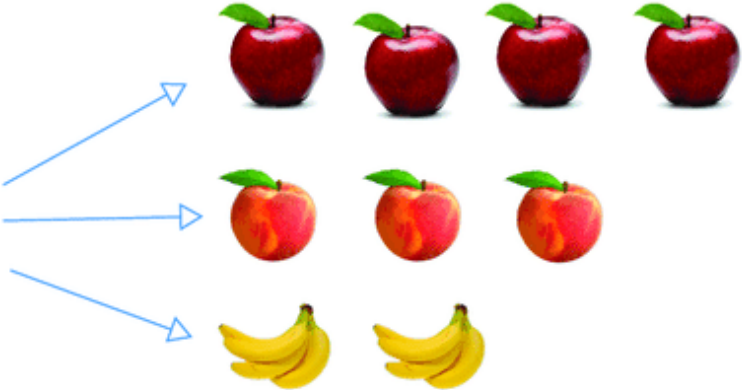
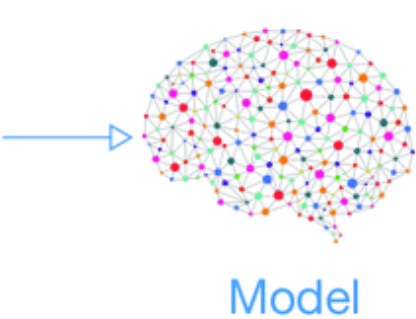


Associação: encontrar padrões frequentes de associações entre atributos



Agrupamento: Dados agrupados de acordo com sua similaridade

unsupervised learning



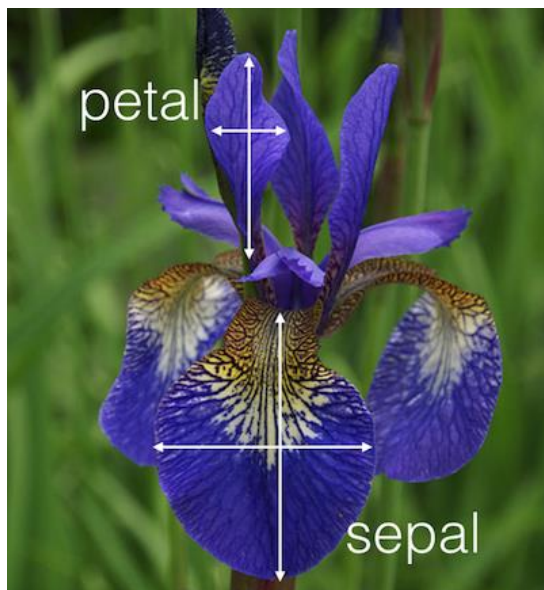


Walmart 





PRE-PROCESAMIENTO



Conjunto de dado Iris

- Um dos conjuntos de dados mais famosos em AM
- Os dados consistem de 50 unidades amostrais de três espécies (setosa, virginica, versicolor) de íris (uma espécie de planta), ou seja, temos um total de 150 unidades amostrais.
- De cada uma delas mediu-se quatro variáveis morfológicas: comprimento e largura da sépala (CS, LS) e comprimento e largura da pétala (CPLP).
- O objetivo original é quantificar a variação morfológica em relação a essas espécies com bases nas quatro variáveis de interesses



Iris Versicolor

Iris Setosa

Iris Virginica

Pre- Processamento dos dados

O pré-processamento é o processo de preparação, organização e estruturação dos nossos dados. A qualidade dos dados pode influenciar diretamente no resultado do modelo. O pre-processamento pode consistir em:

- **Tratamento de dados faltantes**

- **Deletar as colunas com dados faltantes:** Só quando a variável não exercer uma certa influência no resultado procurado.
- **Deletar os exemplos com dados faltantes:** Solução melhor, não recomendada quando temos pouco dados.
- **Preencher os dados faltantes com a média dos valores do atributo:** Uma das soluções preferidas pelos cientistas de dados.
- **Preencher os dados faltantes com o valor que você quiser:** Cuidado com os valores que são selecionados, um dos mais utilizados é zero

Pre- Processamento dos dados

- **Tratamento de variáveis categóricas:** uma variável categórica é uma variável nominal, sem escala, não numérica Ex. Cargo, sexo, país, etc.
 - **Uma solução é transformar este tipo de dados a números.**
- **Reescala dos dados:** quando nosso dataset possui atributos em diferentes escalas. Ex idade= 20-65, Salario = 2000-15000, vale transporte = 75-500. Isto causa muito problema já que um parâmetro terá maior influencia que outro.
 - **Normalizar:** reescala os dados tomando em conta os valores que apresentam os atributos dentro do data-set. O processo é feito por colunas.
 - **Max-Min:** é uma outra alternativa a reescala de dados, o cálculo da reescala é feito de forma independente entre cada coluna, de tal forma que a nova escala se dará entre 0 e 1 (ou -1 e 1 se houver valores negativos no dataset).
$$\text{valor} = (\text{valor} - \text{Coluna.min}) / (\text{Coluna.max} - \text{Coluna.min})$$
 - **Estandarização:** age sobre as colunas, este subtrai do valor de cada instancia em questão a média da coluna e divide o resultado pelo desvio padrão. Esse método trabalha melhor em dados com distribuição normal porém vale a tentativa para outros tipos de distribuições. $\text{valor} = (\text{valor} - \text{média}) / \text{desvioPadão}$

Pre- Processamento dos dados

- **RobustScaler:** atua sobre as colunas e este método nos garante um bom tratamento dos outliers. Em seu método subtrai a média do valor em questão e então divide o resultado pelo segundo quartil.
- **Transformação por quartil:** Este método transforma os valores de tal forma que a distribuição tende a se aproximar de uma distribuição normal. Uma observação importante é que essa transformação pode distorcer as correlações lineares entre as colunas. Todos os valores serão reescalados em um intervalo de 0 a 1.
- **PowerTransformer:** procura transformar os valores em uma distribuição mais normal, sendo indicado em situações onde uma distribuição normal é desejada para os dados

	Método	Quando usar	Observações
1	Normalizer	Quando a distribuição dos seus dados não é normal ou quando você não sabe qual é o tipo de distribuição dos seus dados.	Atua sobre as linhas/exemplos e não sobre as colunas/atributos
2	MinMaxScaler	Quando a distribuição dos dados não for normal e se o desvio padrão for pequeno	Não reduz de forma eficaz o impacto de outliers e também preserva a distribuição original. valor = (valor – Coluna.min) /
3	StandardScaler	Quando os dados estão com distribuição normal ou quando é necessário transformar os valores em uma distribuição mais normal	É uma boa combinação com algoritmos como Linear Regression e Logistic Regression
4	RobustScaler	Quando queremos reduzir o impacto de outliers	
5	QuantileTransformer	Quando queremos reduzir o impacto de outliers	Trata os outliers de uma forma mais agressiva do que o RobustScaler
6	PowerTransformer	Quando é necessário transformar os valores em uma distribuição mais normal	

Método		Dados em distribuição normal	Dados não estão em distribuição normal	É desejado que os dados estejam em distribuição normal	É desejado eliminar a influência dos outliers
1	Normalizer	✗	✓	✗	✗
2	MinMaxScaler	✗	✓	✗	✗
3	StandardScaler	✓	✓	✓	✗
4	RobustScaler				✓
5	QuantileTransformer				✓
6	PowerTransformer				



ALGORITMOS DE CLASSIFICAÇÃO

Modelos → Generalização

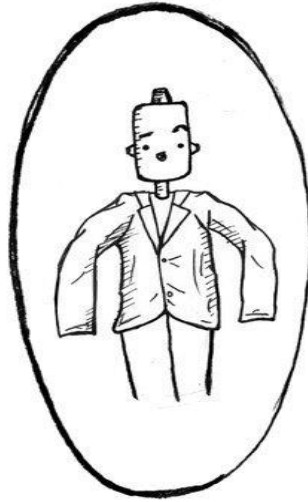
Capacidade de generalização de uma hipótese

- Propriedade de continuar válida para outros objetos que não fazem parte de seu conjunto de treinamento.
- Problemas
 - **Overfitting:** especialização nos dados de treinamento, não generaliza
 - **Underfitting:** baixa taxa de acerto mesmo nos dados de treinamento

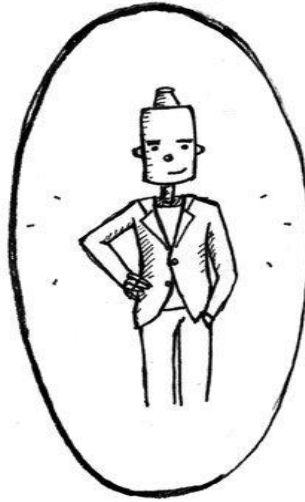
MACHINE LEARNING GENERALIZATION

FINDING THE PERFECT FIT

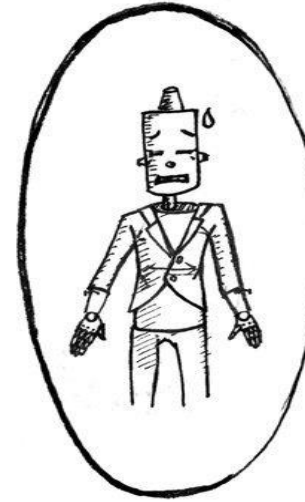
UNDERFIT



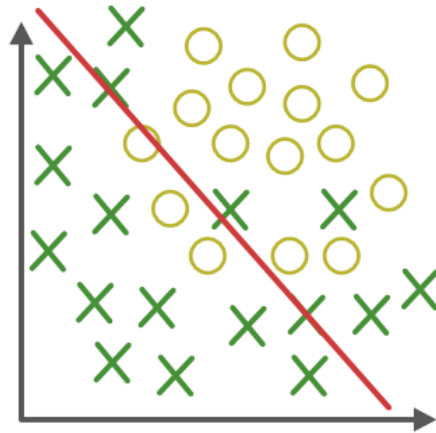
GOLDBLOCKS ZONE



OVERFIT

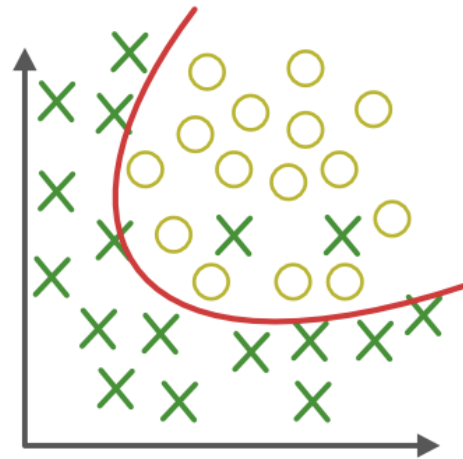


EUCLEDEAN TECHNOLOGIES MANAGEMENT ©

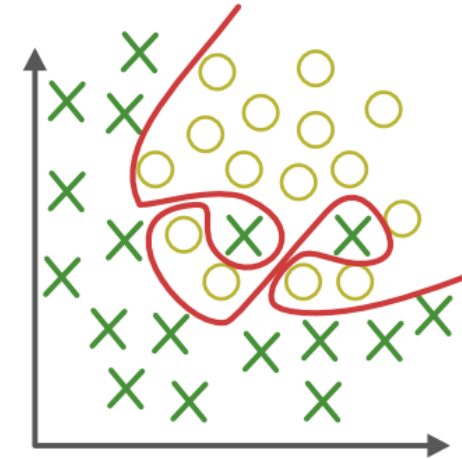


Under-fitting

(too simple to explain the variance)



Appropriate-fitting



Over-fitting

(forcefitting--too good to be true)





K-VIZINHO MAIS PRÓXIMO

K-Vizinho mais próximo

KNN

- É um dos modelos preditivos mais simples
- Não possui premissa matemática
- Não requer computadores de alto desempenho

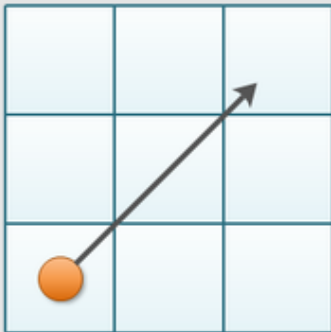
Premissas do modelo

- Utiliza uma noção de distancia
- Os pontos que estão mais perto um do outro são similares.

Metricas de distancias

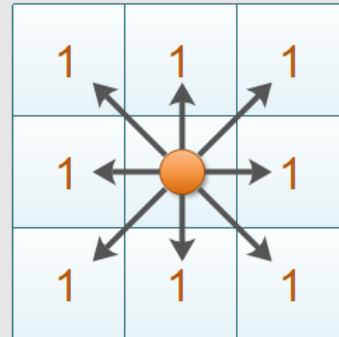
- Distancia Euclidean (popular)
- Distancia Manhattan
- Distancia Chebyshev

Euclidean Distance



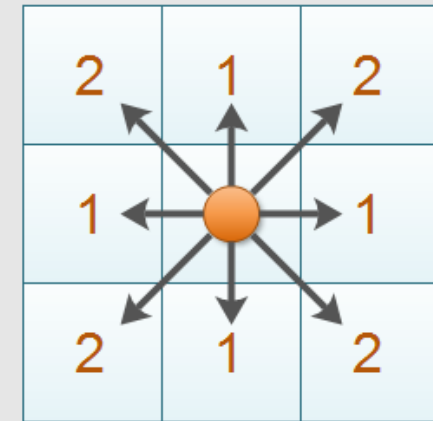
$$\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

Chebyshev Distance

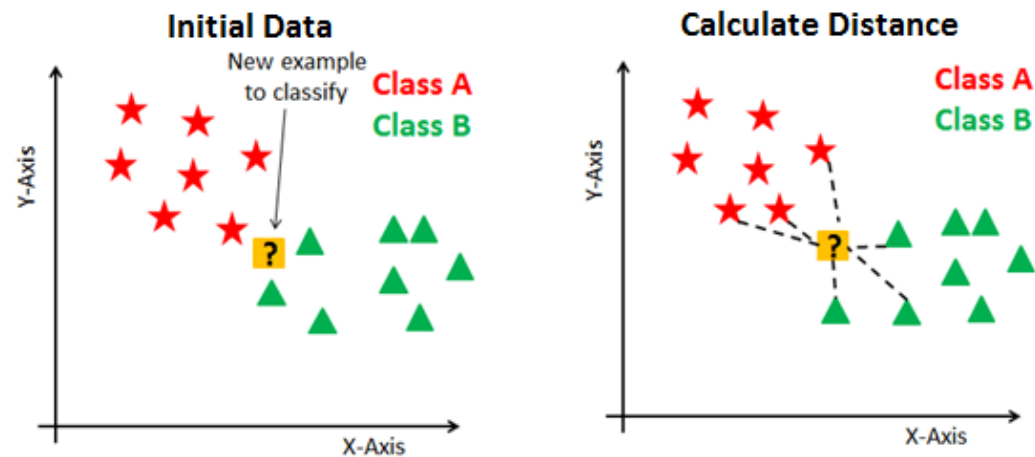


$$\max(|x_1 - x_2|, |y_1 - y_2|)$$

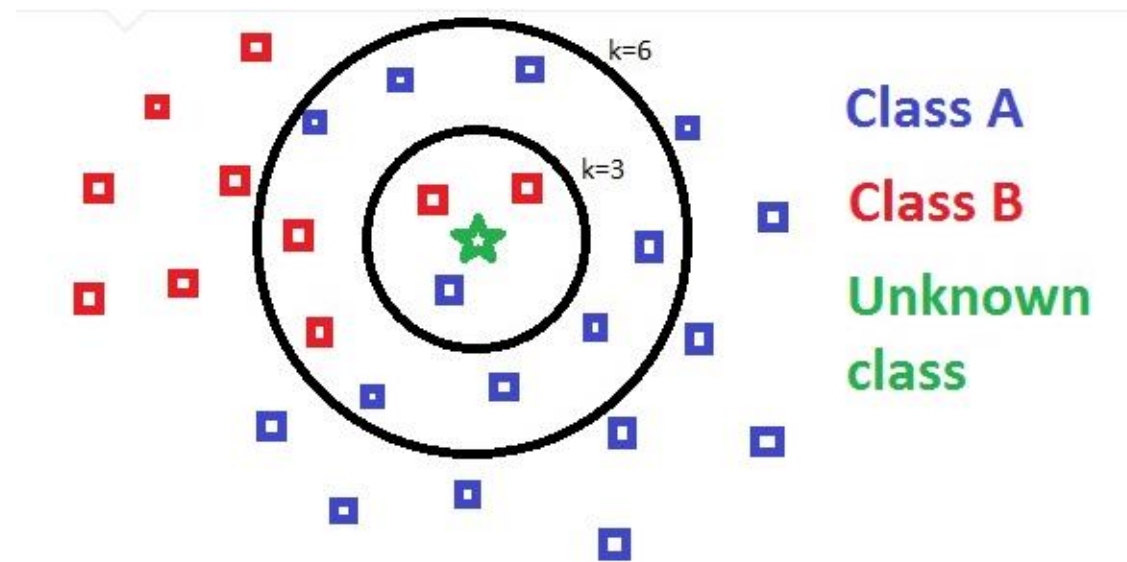
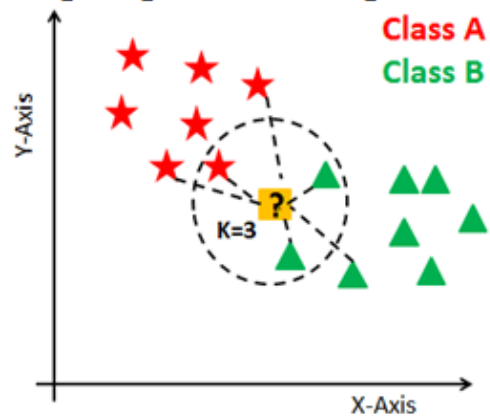
Manhattan Distance



$$|x_1 - x_2| + |y_1 - y_2|$$



Finding Neighbors & Voting for Labels



CÓDIGO KNN-CLASIFICACÃO

```
1 <html http-equiv="Content-Type" content="text/html; charset=UTF-8" >
2 <head>
3 <meta http-equiv="Content-Type" content="text/html; charset=UTF-8" >
4 <title>html</title>
5 <style type="text/css">
6 #xsnazzy h1, #xsnazzy h2, #xsnazzy p {margin:0 10px; letter-spacing:0.1em;}
7 #xsnazzy h1 {font-size:2.5em; color:#A95051;}
8 #xsnazzy h2 {font-size:2em; color:#7B793D; border:0;}
9 #xsnazzy p {padding-bottom:0.5em;}
10 #xsnazzy h2 {padding-top:0.5em;}
11 #xsnazzy {background:transparent; margin:25px 1em 100px 1em;}
12 .xtop, .xbottom {display:block; background:transparent; font-size:1.2em;}
13 .xb1, .xb2, .xb3, .xb4 {display:block; overflow:hidden;}
14 .xb1, .xb2, .xb3 {height:1px;}
15 .xb2, .xb3, .xb4 {background:#FFF; border-left:1px solid #A93298;}
16 .xb1 {margin:0 5px; background:#A93298;}
17 .xb2 {margin:0 3px; border-width:0 2px;}
18 .xb3 {margin:0 2px;}
19 .xb4 {height:2px; margin:0 2px;}
```

Pros e contras

Prós

- A fase de treinamento da classificação K-vizinho é muito mais rápida em comparação com outros algoritmos de classificação.
- Não é necessário treinar um modelo para generalização. É por isso que o KNN é conhecido como algoritmo de aprendizado simples e baseado em instância.
- KNN pode ser útil no caso de dados não lineares.
- Pode ser usado com o problema de regressão.

Contras

- A fase de teste da classificação K-vizinho é mais lenta e mais cara em termos de tempo e memória.
- Requer muita memória para armazenar todo o conjunto de dados de treinamento para previsão.
- O KNN requer dimensionamento de dados porque o KNN usa a distância euclidiana entre dois pontos de dados para encontrar vizinhos mais próximos.
- A distância euclidiana é sensível às magnitudes. Os recursos com altas magnitudes pesam mais do que os recursos com baixas magnitudes.
- KNN também não é adequado para grandes dados dimensionais.



NAIVES BAYES

Classificador Naive-Bayes

- O algoritmo “**Naive Bayes**” é um classificador probabilístico baseado no “**Teorema de Bayes**”,
- Naive Bayes assume que a presença de uma característica particular em uma classe não está relacionada com a presença de qualquer outro recurso.
- Todas estas propriedades contribuem de forma **independente** (relação entre eles) para a probabilidade de que este fruto é uma maçã e é por isso que é conhecido como ‘Naive’ (ingênuo).



- Vermelha
- Redondo
- Cerca de 3 polegadas de diâmetro.

Diagram illustrating Bayes' Theorem with arrows pointing from terms to their descriptions:

- $P(c | x)$ is labeled "Probabilidade posterior" (Posterior Probability).
- $P(x | c)$ is labeled "Probabilidade" (Probability).
- $P(c)$ is labeled "Probabilidade original da Classe" (Original Class Probability).
- $P(x)$ is labeled "Preditor da probabilidade posterior" (Predictor of posterior probability).

$$P(c | x) = \frac{P(x | c)P(c)}{P(x)}$$

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

Exemplo

Imaginemos que estamos trabalhando no diagnóstico de uma nova doença, e que fizemos testes em 12 pessoas distintas.

No	Resultado teste	Doente
1	Positivo	Sim
2	Negativo	Não
3	Positivo	Não
4	Negativo	Não
5	Positivo	Sim
6	Positivo	Não
7	Positivo	Sim
8	Negativo	Não
9	Positivo	Sim
10	Negativo	Sim
11	Negativo	Não
12	Positivo	Sim

Teste	Correto	Incorreto
Positivo	5	2
Negativo	4	1
total	9	3

- 71% pacientes que o teste foi positivo estão doentes (verdadeiros positivos)
- 28 % dos pacientes que foram positivos não estão doentes (Falsos positivos)
- 80% dos resultados negativos não estão doentes. (Verdadeiros negativos)
- 25% dos pacientes negativos estão doentes. (Falsos negativos)

Se uma nova pessoa realizar o teste e receber um resultado positivo, qual a probabilidade de ela possuir a doença?

- **$P(\text{Correto}|\text{positivo}) = P(\text{positivo}|\text{correto}) * P(\text{correto}) / P(\text{Positivo})$**
 - $P(\text{positivo}|\text{correto}) = 5 / 9 = 0.555$ * $P(\text{correto}) = 9 / 12 = 0.75$ / $P(\text{positivo}) = 7 / 12 = 0.583$
 $0.555 * 0.75 = 0.416$ $0.416 / 0.583 = 0.71$
- **$P(\text{Incorreto}|\text{positivo}) = P(\text{positivo}|\text{incorreto}) * P(\text{incorreto}) / P(\text{Positivo})$**
 - $P(\text{positivo}|\text{incorreto}) = 2 / 3 = 0.666$ * $P(\text{incorreto}) = 3 / 12 = 0.25$ / $P(\text{positivo}) = 7 / 12 = 0.583$
 $0.666 * 0.25 = 0.166$ $0.166 / 0.58 = 0.285$
- **Positivo e estar doente = 71%**
- **Positivo e não estar doente = 29%**

Foram coletados dados sobre a quantidade de homens e mulheres foram admitidos nos cursos?

Que probabilidade existes que um esdudante seja homem e seja admitido?

Dados	Admitido	Rejetado
Female	6	4
Male	4	6
Totales	10	10

No	Admit	Gender
1	Admitted	Female
2	Rejected	Male
3	Admitted	Female
4	Rejected	Male
5	Admitted	Female
6	Rejected	Male
7	Admitted	Female
8	Rejected	Male
9	Admitted	Female
10	Rejected	Male
11	Admitted	Female
12	Rejected	Male
13	Admitted	Male
14	Rejected	Female
15	Admitted	Male
16	Rejected	Female
17	Admitted	Male
18	Rejected	Female
19	Admitted	Male
20	Rejected	Female


- **$P(\text{Homem}|\text{admitido}) = P(\text{admitido}|\text{homem}) * P(\text{admitido}) / P(\text{homem})$**
 - $P(\text{admitido}|\text{homem}) = 4 / 10 = 0.4 * P(\text{admitido}) = 10/20 = 0.5 / P(\text{homem}) = 10/20 = 0.5$
 - $0.4 * 0.5 = 0.2 \qquad 0.2/0.5 = 0.4$

- **$P(\text{Homem}|\text{rejeitado}) = P(\text{rejeitado}|\text{homem}) * P(\text{rejeitado}) / P(\text{homem})$**
 - $P(\text{rejeitado}|\text{homem}) = 6/10 = 0.6 * P(\text{rejeitado}) = 10/20 = 0.5 / P(\text{homem}) = 10/20 = 0.5$
 - $0.6 * 0.5 = 0.3 \qquad 0.3/0.5 = 0.6$

- **Homem e ser admitido = 40%**
- **Homen e ser rejeitado = 60%**



CROSS VALIDATION



CODIGO NAIVES BAYES

Pro e contra de Naïve Bayes

- Pro

- É fácil e rápido para prever o conjunto de dados da classe de teste. Também tem um bom desempenho na previsão de classes.
- Quando a suposição de independência prevalece, um classificador Naive Bayes tem melhor desempenho em comparação com outros.
- O desempenho é bom em caso de variáveis categóricas de entrada comparada com a variáveis numéricas.

- Contra

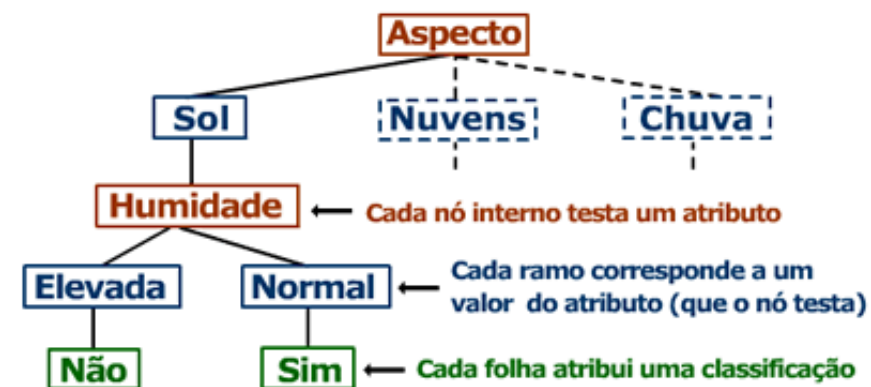
- Se a variável categórica tem uma categoria que não foi observada no conjunto de dados de treinamento, então o modelo irá atribuir uma probabilidade de 0 (zero) e não será capaz de fazer uma previsão.
- Conhecido como um mau estimador, por isso, as probabilidades calculadas não devem ser levadas muito a sério.
- É a suposição de preditores independentes. Na vida real, é quase impossível que ter um conjunto de indicadores que sejam completamente independentes



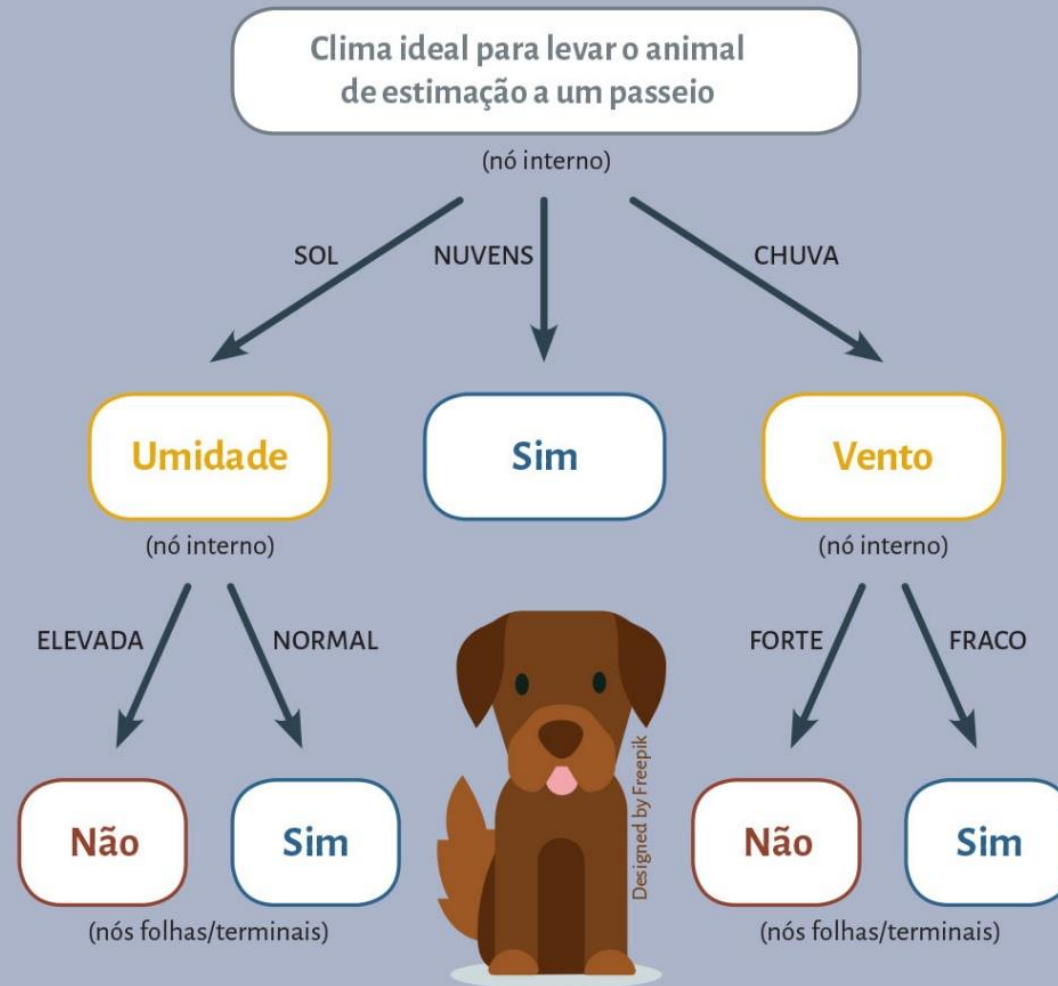
ARVORES DE DECISÃO

Árvores de decisão

- Árvores de decisão são métodos de aprendizado de máquinas supervisionado não-paramétricos, muito utilizados em tarefas de classificação e regressão.
- São estruturas de dados formadas por um conjunto de elementos que armazenam informações chamadas **nó**.
- Toda árvore possui um nó chamado **raiz**, que possui o maior nível hierárquico (o ponto de partida) e ligações para outros elementos, denominados filho.
- O nó que não possui filho é conhecido como **nó folha** ou termina.
- Em uma árvore de decisão, uma decisão é tomada através do caminhamento a partir do nó raiz até o nó folha



Exemplo de árvore de decisão



Por que árvores de decisão são tão populares?

- Fácil explicabilidade e interpretação, já que podemos facilmente visualizá-las (quando não são muito profundas).
- Requerem pouco esforço na preparação dos dados, métodos baseados em árvores normalmente não requerem normalização dos dados.
- Conseguem lidar com valores faltantes, categóricos e numéricos (não é o caso da CART que implementamos).
- Complexidade logarítmica na etapa de predição.
- São capazes de lidar com problemas com múltiplos rótulos.

CODIGO

Contra

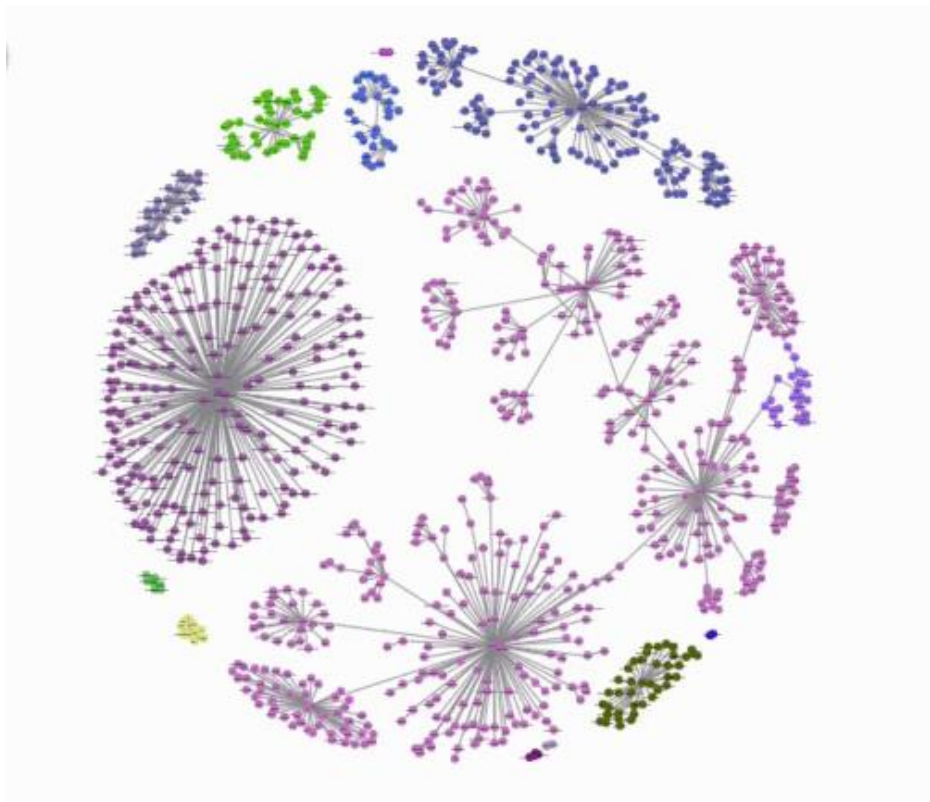
- Árvore crescida até sua profundidade máxima pode decorar o conjunto de treino (o temido overfitting), o que pode degradar seu poder preditivo quando aplicado a novos dados. Isso pode ser mitigado "podando" a árvore de decisão ao atribuir uma profundidade máxima ou uma quantidade máxima de folhas.
- São modelos instáveis (alta variância), pequenas variações nos dados de treino podem resultar em árvores completamente distintas. Isso pode ser evitado ao treinarmos várias árvores.
- O algoritmo de construção da árvore de decisão é guloso, ou seja, não garante a construção da melhor estrutura para o dados de treino em questão.



ALGORITMOS DE AGRUPAMENTOS (CLUSTERING)

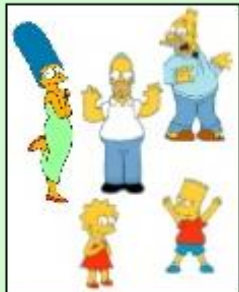
Agrupamentos

Clustering



- Baseado na informação intrínseca dos dados e suas relações
- Compactos, menor distancia intracluster
- Separados, maior distancia extracusters

What is a natural grouping?



Simpson's Family



School Employees



Females



Males

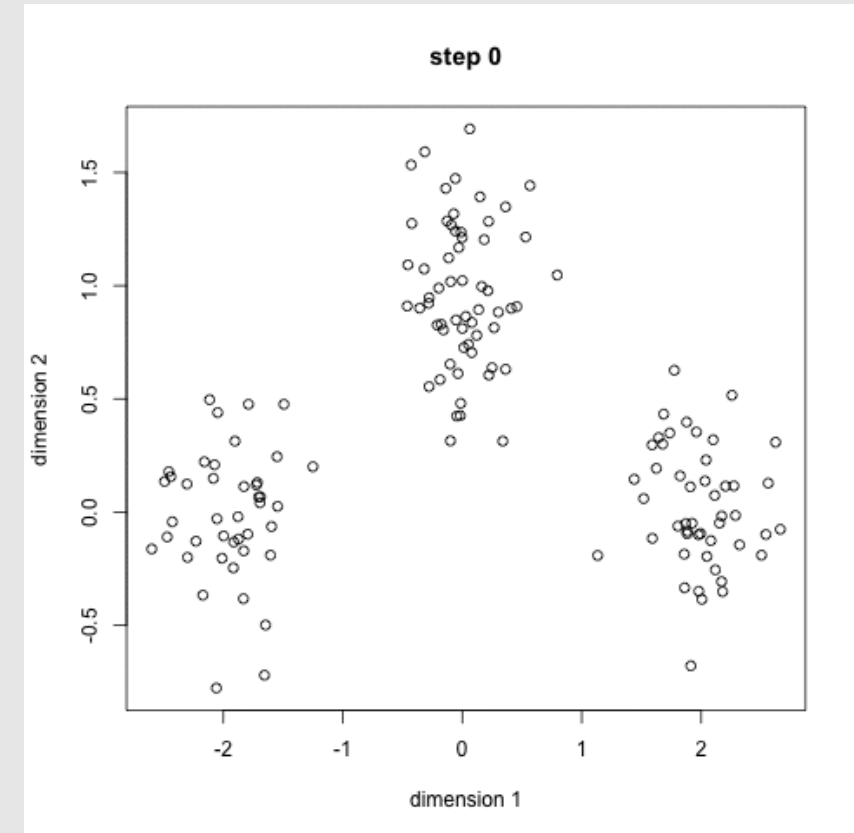
Os agrupamentos são subjetivos

K-means

- O agrupamento K-means é um dos algoritmos de aprendizado de máquina não supervisionados mais simples e populares.
- O objetivo do K-means é simples: agrupar **pontos de dados semelhantes** e descobrir **padrões subjacentes**.
- O K-means procura um número fixo (k) de clusters em um conjunto de dados.
- O usuário define um número de k , que se refere ao número de centróides (clusters) necessários no conjunto de dados.
- Um centróide é o local imaginário ou real que representa o centro do cluster.

Como funciona K-means

- Para começar, primeiro selecionamos um número de classes / grupos para usar e inicializamos aleatoriamente seus respectivos pontos centrais (parte mais difícil).
- Cada ponto de dados é classificado calculando a distância entre esse ponto e cada centro de grupo e depois classificando o ponto no grupo cujo centro está mais próximo.
- Com base nesses pontos classificados, recalculamos o centro do grupo, calculando a média de todos os vetores do grupo.
- Repita essas etapas para um número definido de iterações ou até que as centrais do grupo não alterem muito entre as iterações.



CODIGO

Pro e contra

- **Pro**

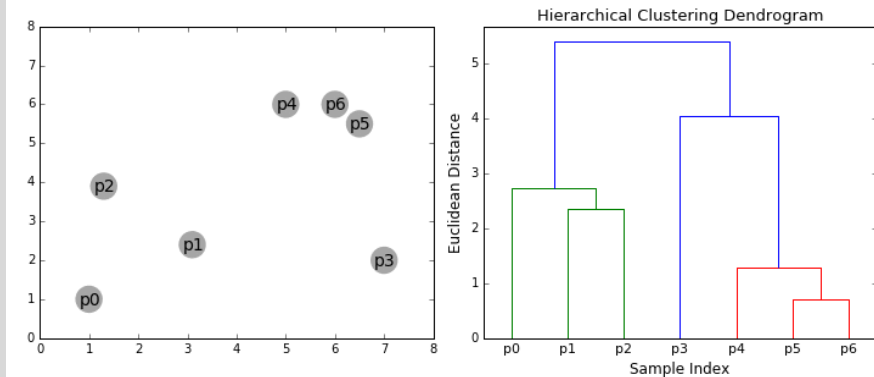
- Um algoritmo fácil de implementar e de entender.
- Usa diferentes tipos de distâncias o que permite que possa ser adaptado a diferentes tipos de dados.

- **Contra**

- Ele é extremamente sensível à seleção inicial dos clusters
- Precisa saber de antemão o número **k** de clusters nos dados.
- Ele assume que os clusters são “esféricos”.

Hierárquico Clustering

- Os algoritmos de cluster hierárquicos se enquadram em 2 categorias: de cima para baixo ou de baixo para cima.
- Os algoritmos **bottom-up** tratam cada ponto de dados como um único cluster no início e, em seguida, mesclam sucessivamente (ou aglomeram) até formar os clusters.
- Os algoritmos up-bottom tratam todos os pontos como agrupadas em único cluster, baseados em isto começa o processo de divisão, até formas os agrupamentos finais.
- A hierarquia dos clusters é representada como uma árvore (ou dendrograma).
- A raiz da árvore é o cluster único que reúne todas as amostras, sendo as folhas os agrupamentos com apenas uma amostra.



Como funciona o algoritmo Hierárquico

1. Começa tratando cada ponto de dados como um único cluster, ou seja, se houver X pontos de dados em nosso conjunto de dados, teremos X clusters.
2. Seleciona uma métrica de distância que mede a distância entre dois clusters. Como exemplo, usaremos a *average linkage* que define a distância entre dois clusters como a distância média entre os pontos de dados no primeiro cluster e os pontos de dados no segundo cluster.
3. Em cada iteração, combina dois clusters em um. Os dois clusters a serem combinados são selecionados como aqueles com o menor vínculo médio, de acordo com nossa métrica de distância selecionada. esses dois clusters
4. A etapa 2 é repetida até chegarmos à raiz da árvore, ou seja, temos apenas um cluster que contém todos os pontos de dados.

CODIGO

Características de Hierárquico

- O armazenamento em cluster hierárquico não exige que especifiquemos o número de clusters e podemos até selecionar qual número de clusters fica melhor, pois estamos construindo uma árvore.
- O algoritmo não é sensível à escolha da métrica de distância; todos eles tendem a funcionar igualmente bem, enquanto com outros algoritmos de agrupamento, a escolha da métrica de distância é crítica.
- Um caso de uso particularmente bom dos métodos de cluster hierárquico é quando os dados subjacentes têm uma estrutura hierárquica e você deseja recuperar a hierarquia; outros algoritmos de cluster não podem fazer isso.
- Essas vantagens do agrupamento hierárquico têm um custo de menor eficiência.

Site recomendados

- medium.com
- Towardsdatascience.com
- Datacamp.com
- www.coursera.org/