Minicurso de Ciência de Dados

Aula 3: Classificação, KNN, Árvores e Florestas

Lucas Pedroso

04 de fevereiro de 2020

CiDAMO

O problema de classificação

As variáveis resposta são categóricas (qualitativas).

Exemplos:

Entrada: imagem de uma letra
Saída: identificação da letra

Entrada: sintomas de um paciente
Saída: diagnóstico

Entrada: dados de uma transação bancária
Saída: é uma fraude?

O problema de classificação

Entrada: foto de uma pessoa
Saída: identificação da pessoa

Entrada: gravação de áudio
Saída: grau de irritação do interlocutor

Entrada: texto jurídico
Saída: sentença

Observações:

- Os dados de entrada podem ou não ser categóricos.
- Os dados precisam ser matematicamente tratáveis.
- Classificação binária vs. multiclasse.

Funcionamento

Entrada → Classificador → Resposta

Como construir o classificador?

Precisamos de um (grande) conjunto de **treinamento** e de um conjunto de **testes**.

Dado	Rótulo	Caract. 1	Caract. 2	Caract. 3	...
1	+1	105	13.15	-5.1	...
2	-1	102	12.30	-7.8	...
3	-1	110	14.05	-6.1	...
4	+1	105	13.27	-8.0	...
...	...	...	...	...	...

Treinamento vs testes

Treinamento → Testes → Lançamento

Treinamento: algoritmo tem acesso às entradas e aos rótulos.

Testes: modelo tem acesso apenas às entradas e tenta descobrir os rótulos, que depois são conferidos. É observada a **capacidade de generalização**.

Lançamento: modelo tem acesso apenas às entradas, os rótulos sugeridos por ele são acatados.

Medidas de desempenho

- **Acurária:** Porcentagem de acertos. Pode ser enganosa se o conjunto de dados for muito assimétrico.
  - Nem sempre queremos acurácia de 100%, mesmo no treinamento.

- **Matriz de confusão:** cada linha representa uma classe real, cada coluna uma classe prevista.

51325	1245	Era -
2560	5903	Era +
Previu -	Previu +	

Medidas de desempenho

Precisão: VP / (VP + FP)

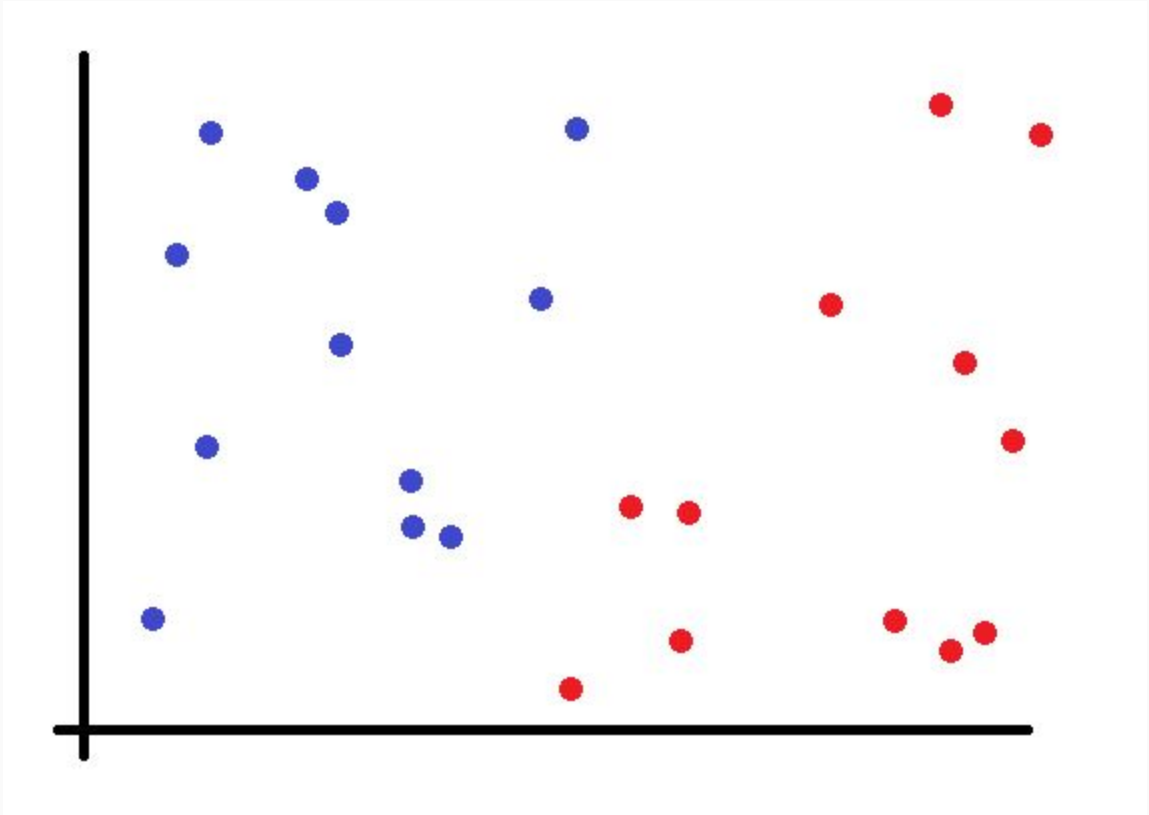
Baixa: Muitos FP (alertas falsos)

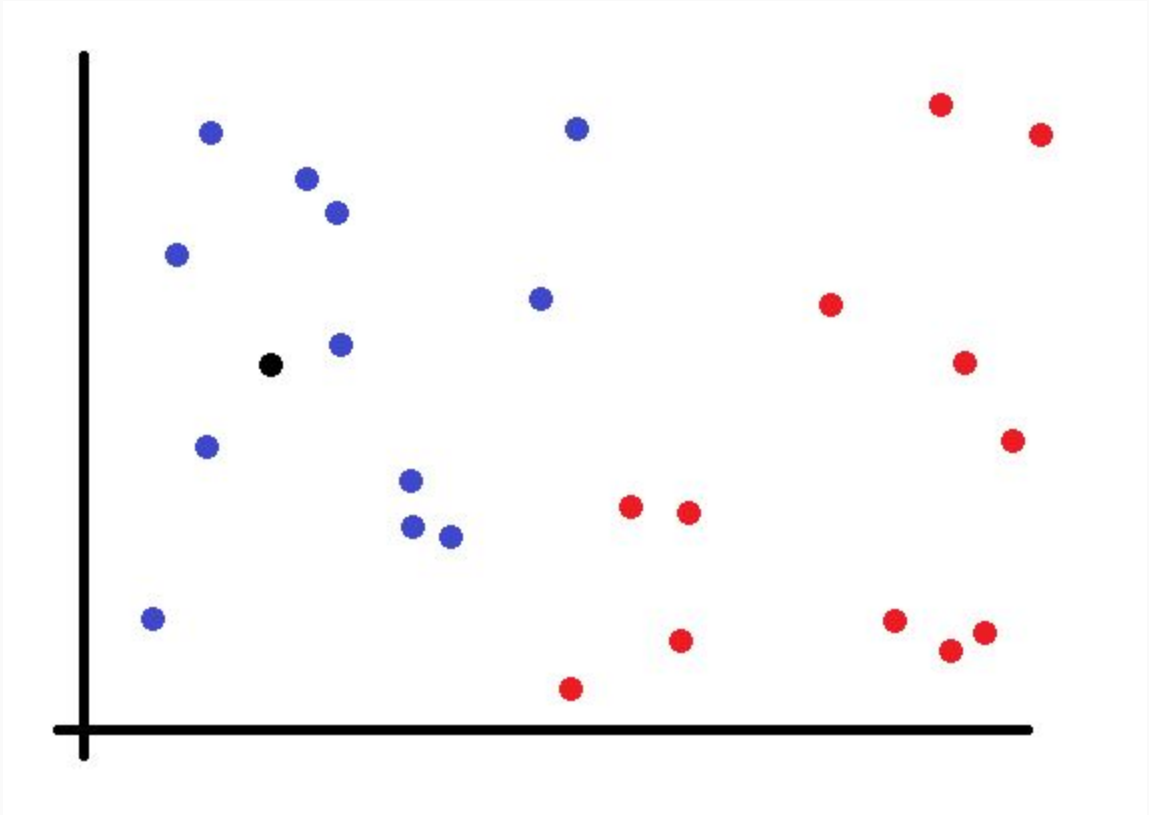
51325	**1245**	Era -
2560	**5903**	Era +
Previu -	Previu +	

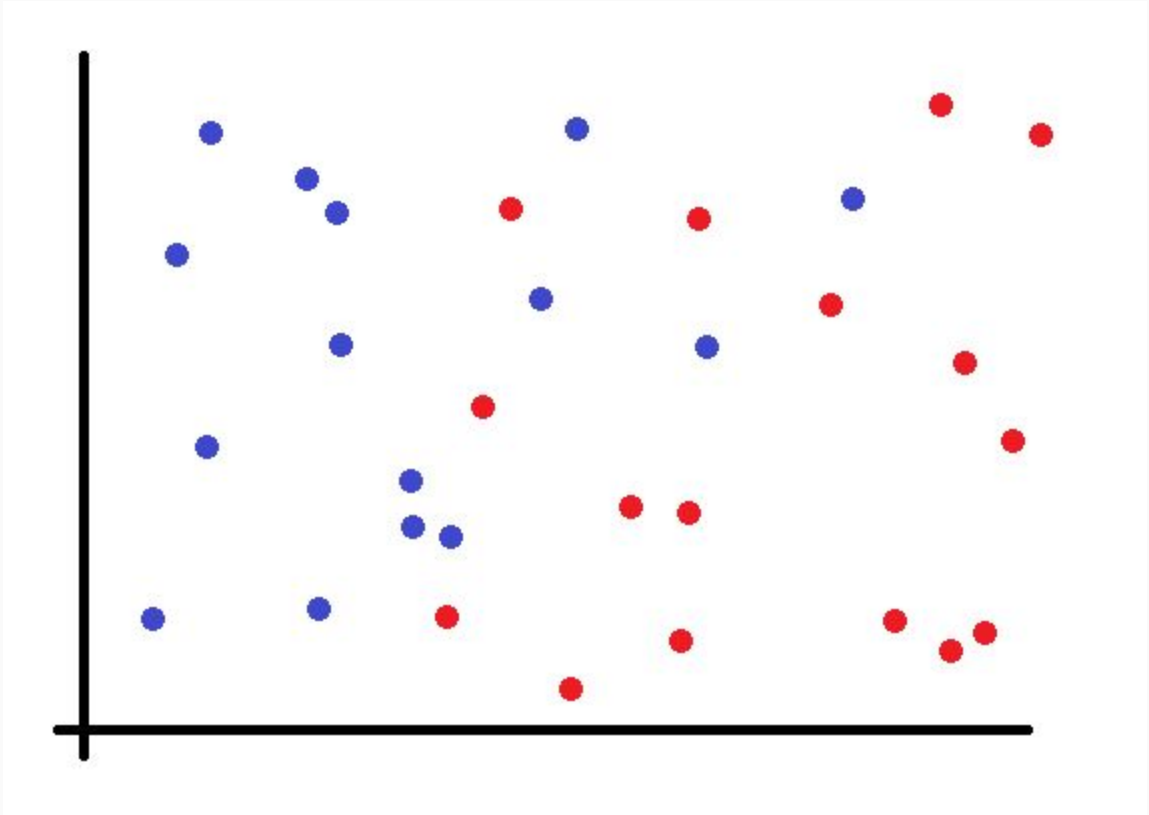
Recall: VP / (VP + FN)

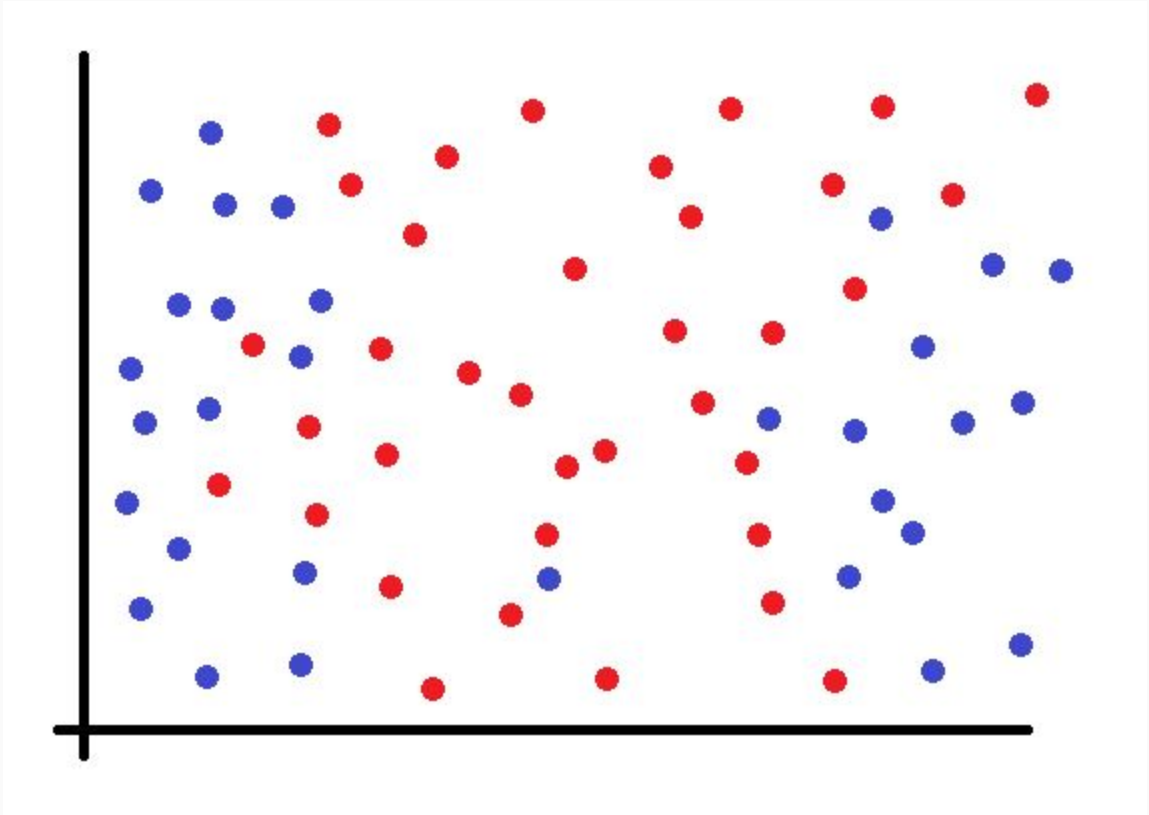
Baixa: Não enxerga bem a classe +1

51325	1245	Era -
2560	**5903**	Era +
Previu -	Previu +	







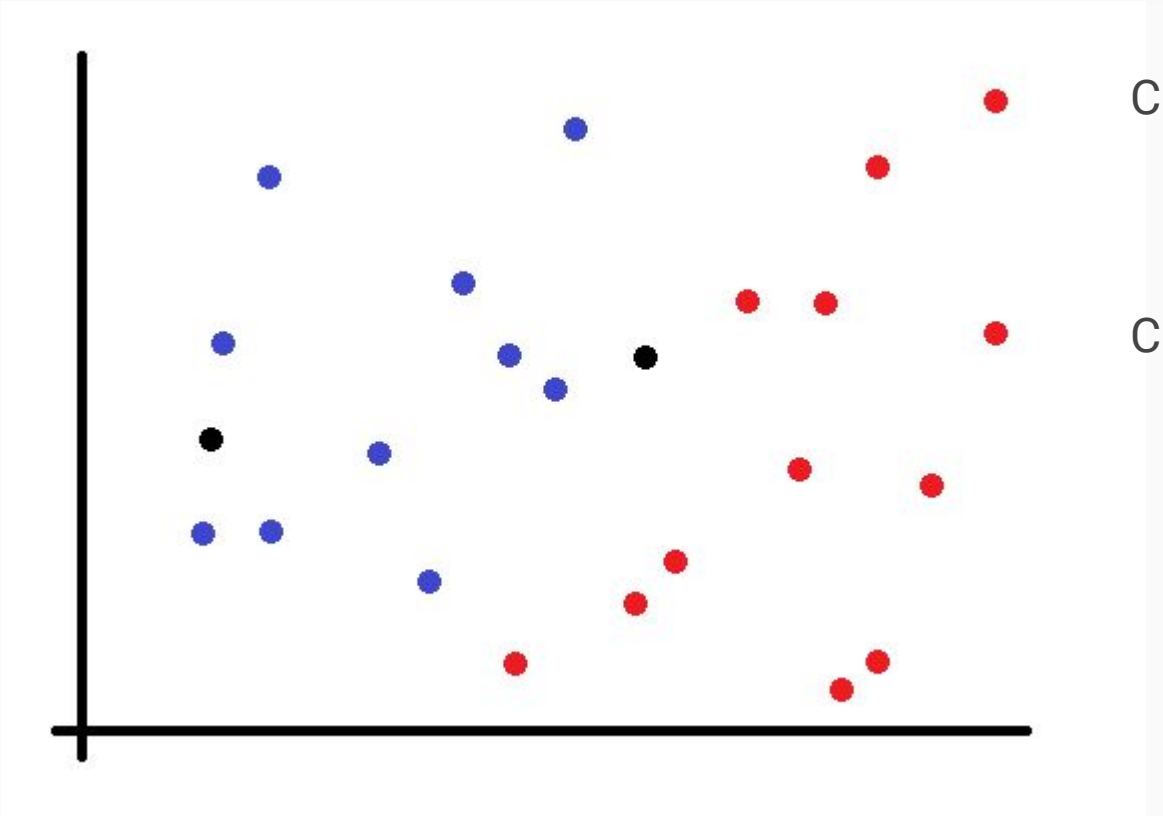


KNN

K-nearest neighbors

Escolhido K pelo usuário, dado um ponto novo, encontra seus K vizinhos mais próximos e atribui aquele à classe majoritária.

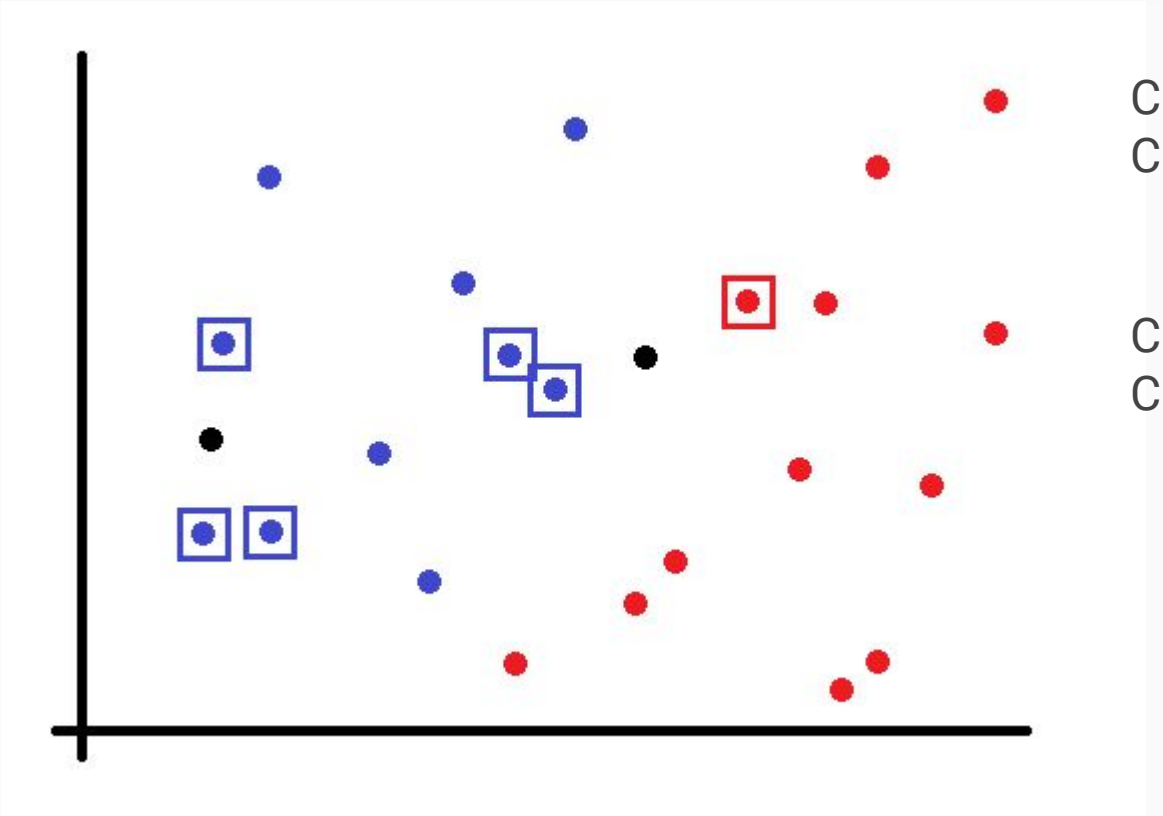
Exemplo com K = 3



Classificação do ponto 1?

Classificação do ponto 2?

Exemplo com K = 3



Classificação do ponto 1?
Classe +1 (100%)

Classificação do ponto 2?
Classe +1 (66%)

Exemplo: Íris

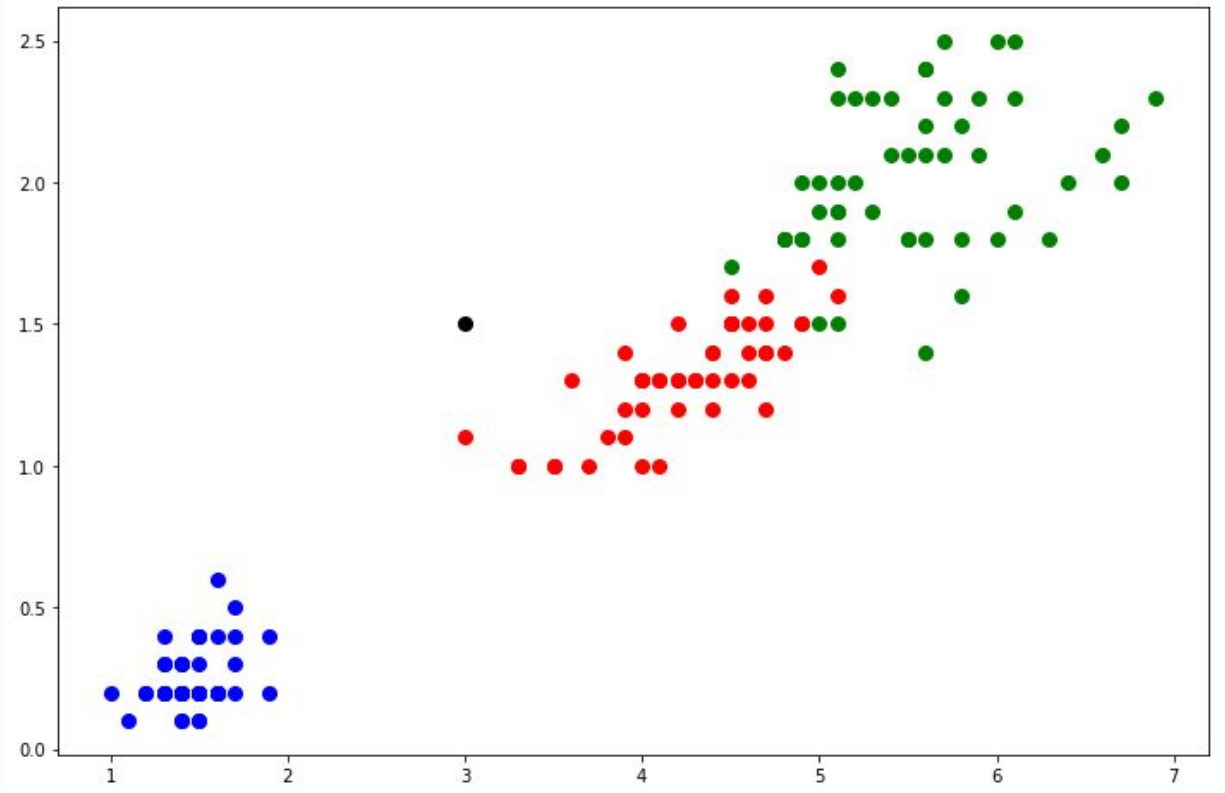
Iris Setosa

Iris Versicolor

Iris Virginica

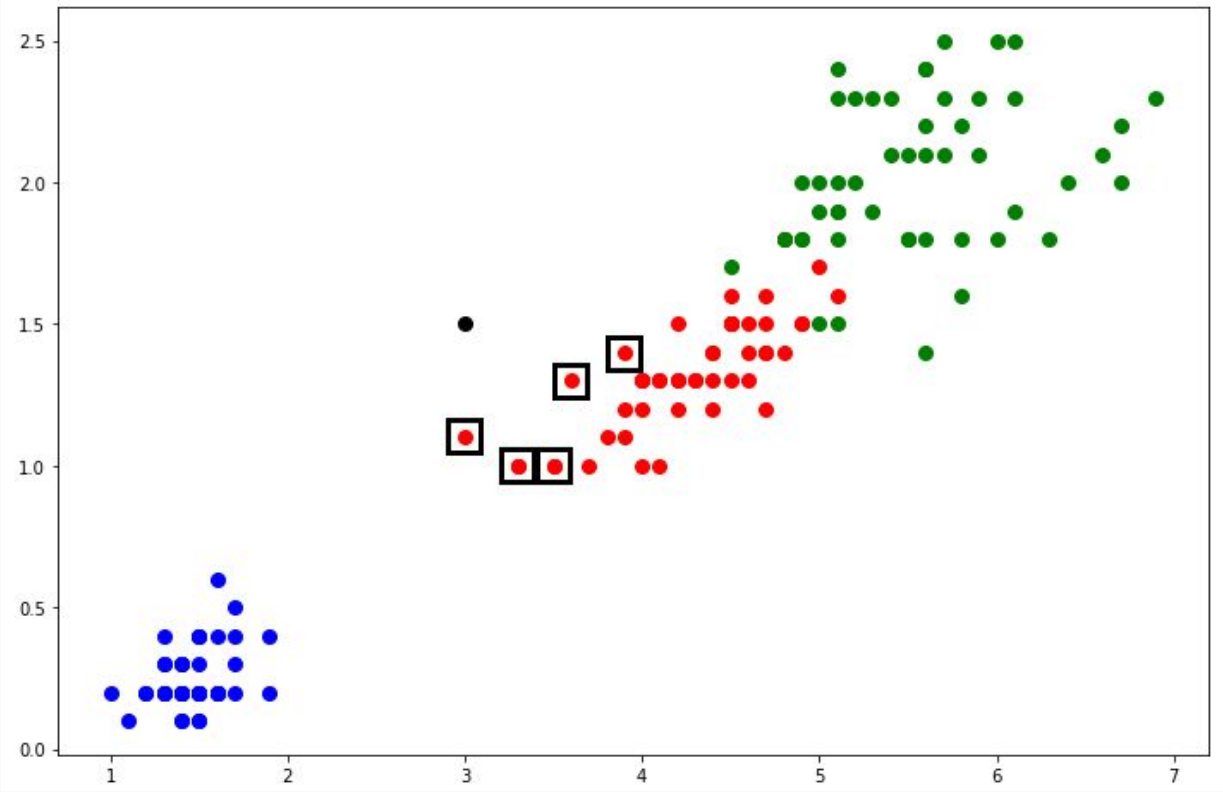
Fonte: site da O'Reilly

Exemplo: Íris (K = 5)



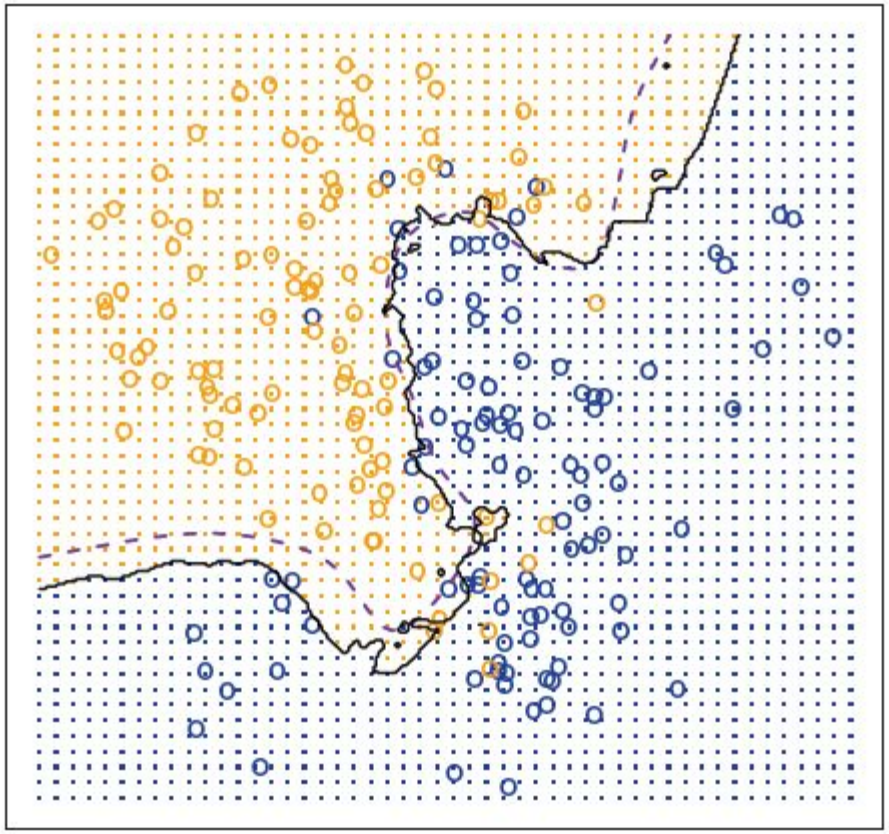
Se uma flor tem pétala com **3cm** de comprimento e **1.5cm** de largura:

Exemplo: Íris (K = 5)



Se uma flor tem pétala com **3cm** de comprimento e **1.5cm** de largura:

- **100.0%** de chance de ser Versicolor
- **0.0%** de chance de ser Virginica
- **0.0%** de chance de ser Setosa



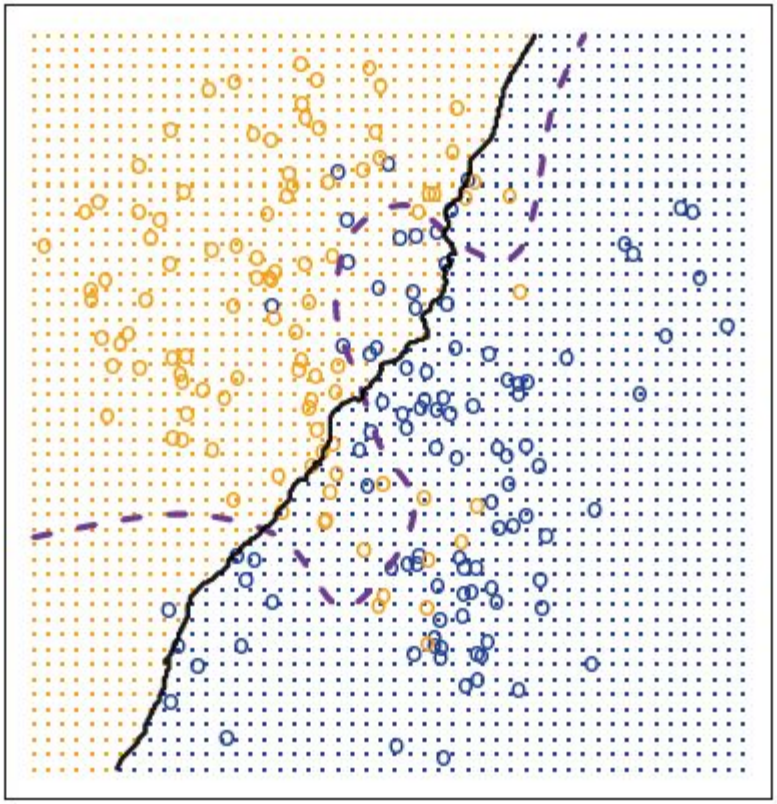
K = 10

Fonte: G. James et al, An Introduction to Statistical Learning.

K = 1

Overfitting

Fonte: G. James et al, An Introduction to Statistical Learning.

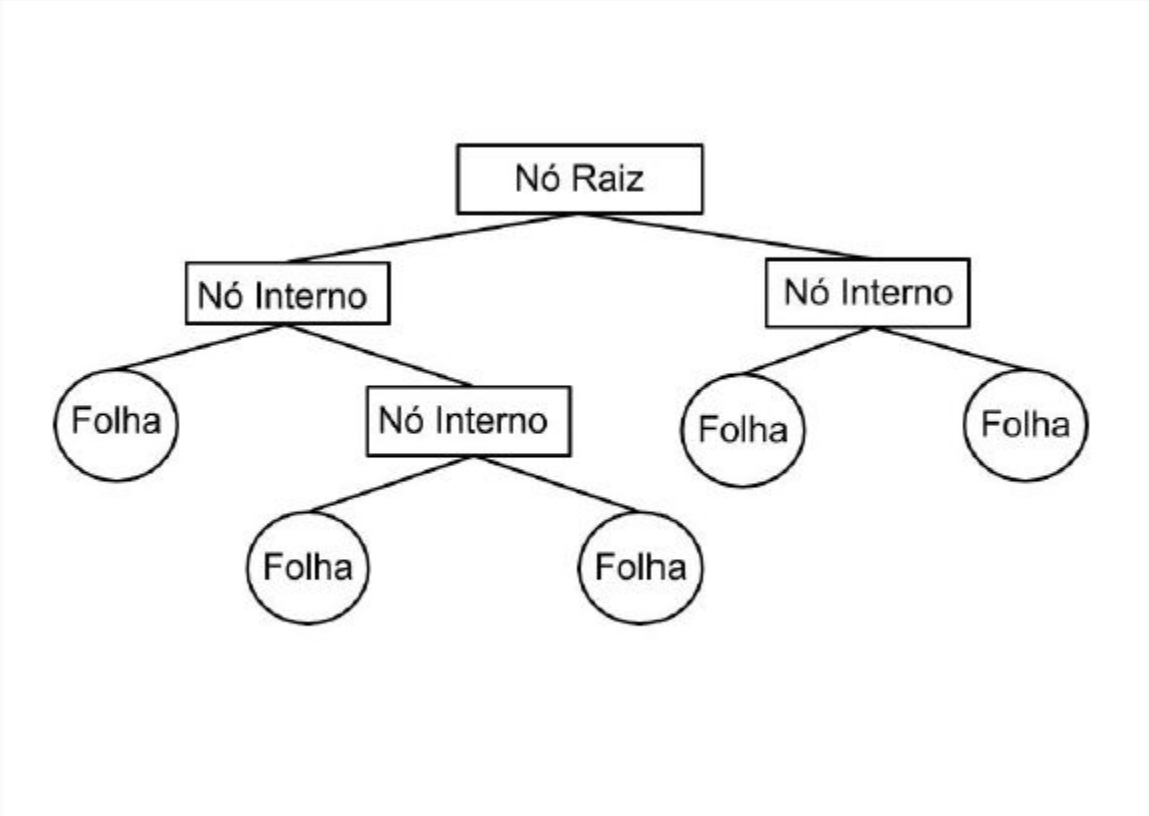


K = 100

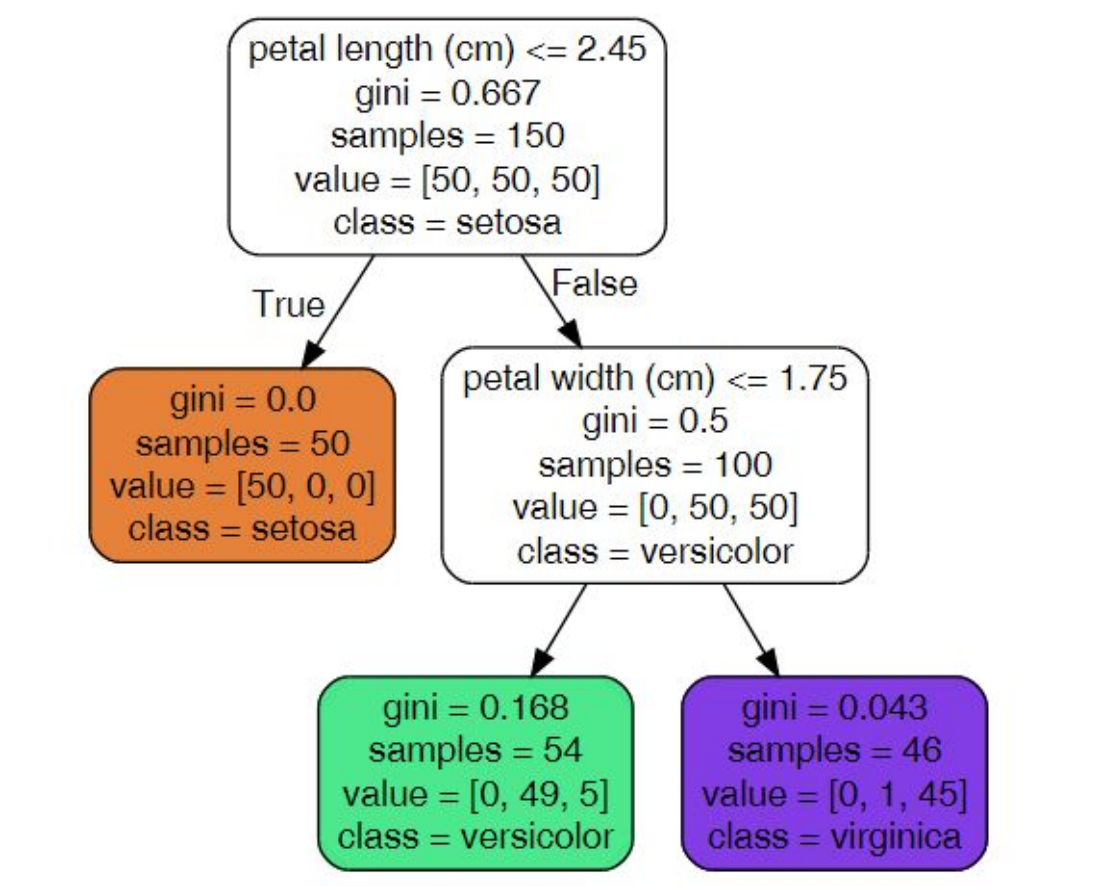
Underfitting

Fonte: G. James et al, An Introduction to Statistical Learning.

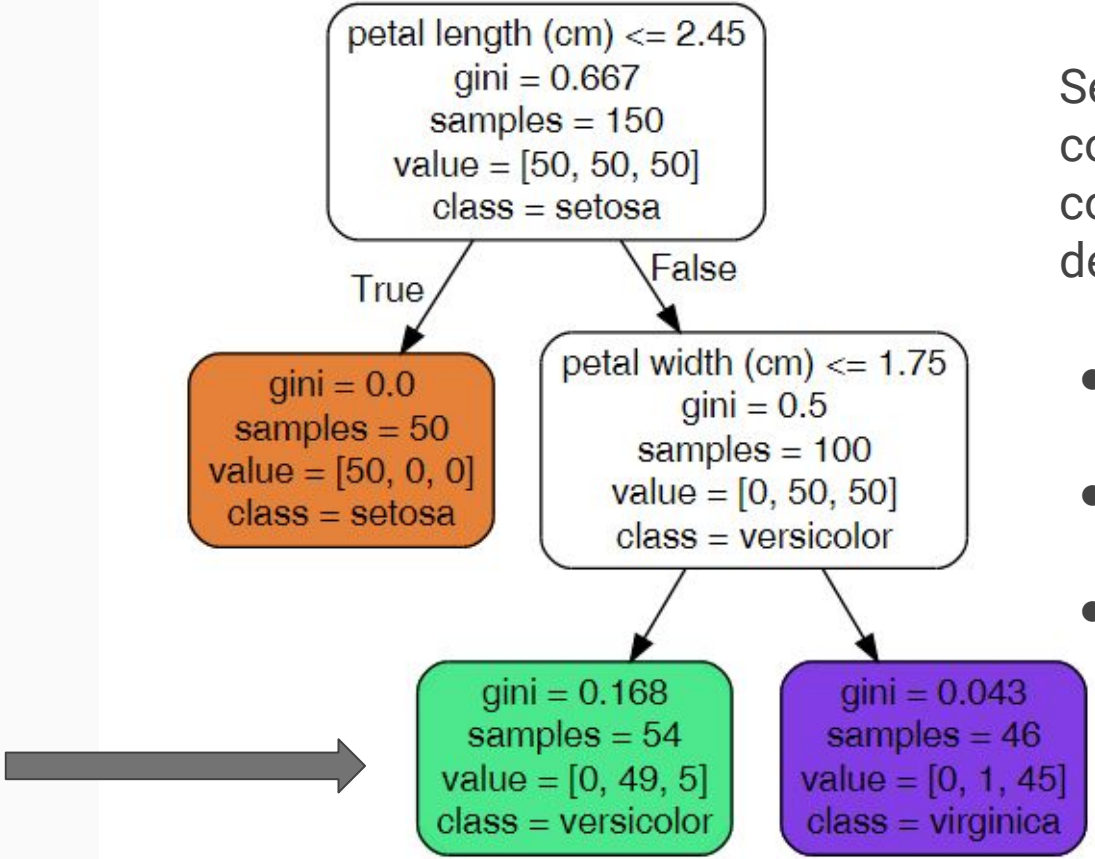
Árvores de decisão



Exemplo: Íris



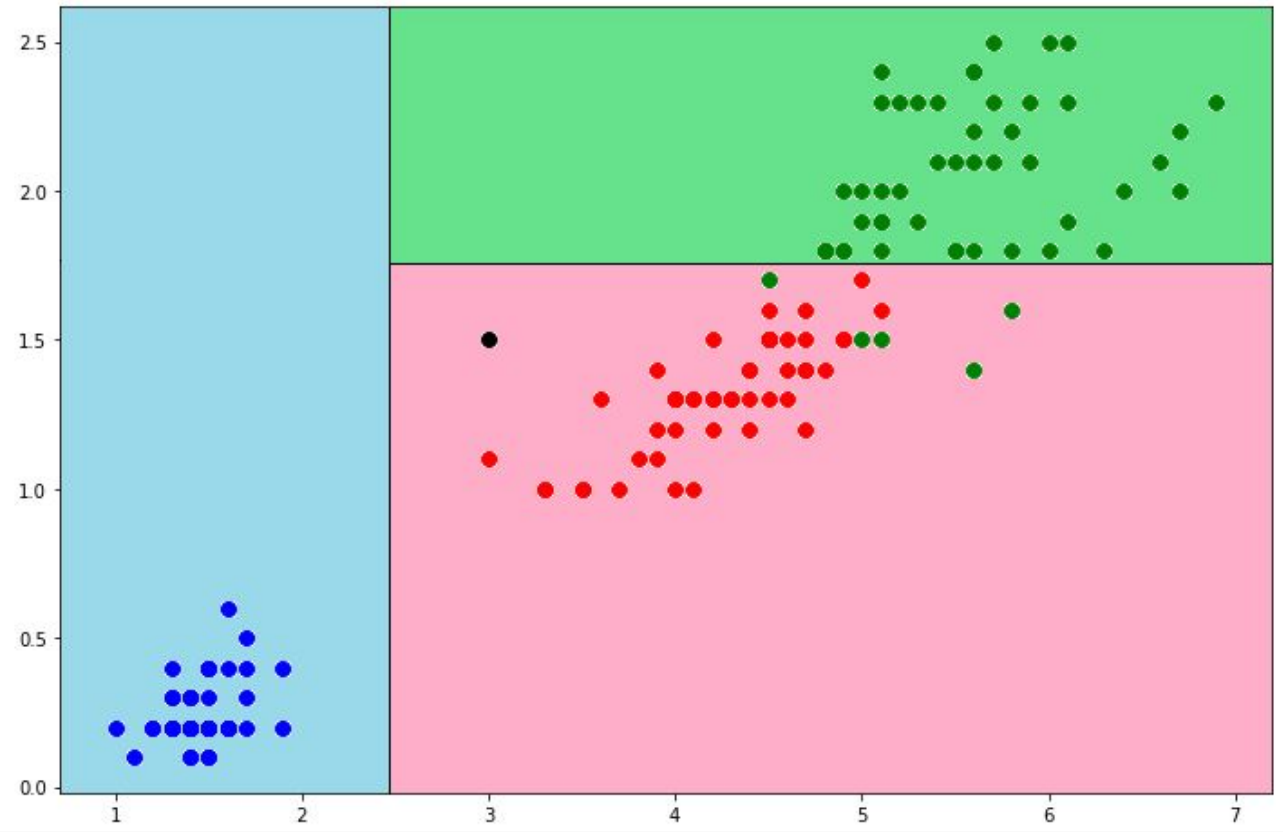
Exemplo: Íris



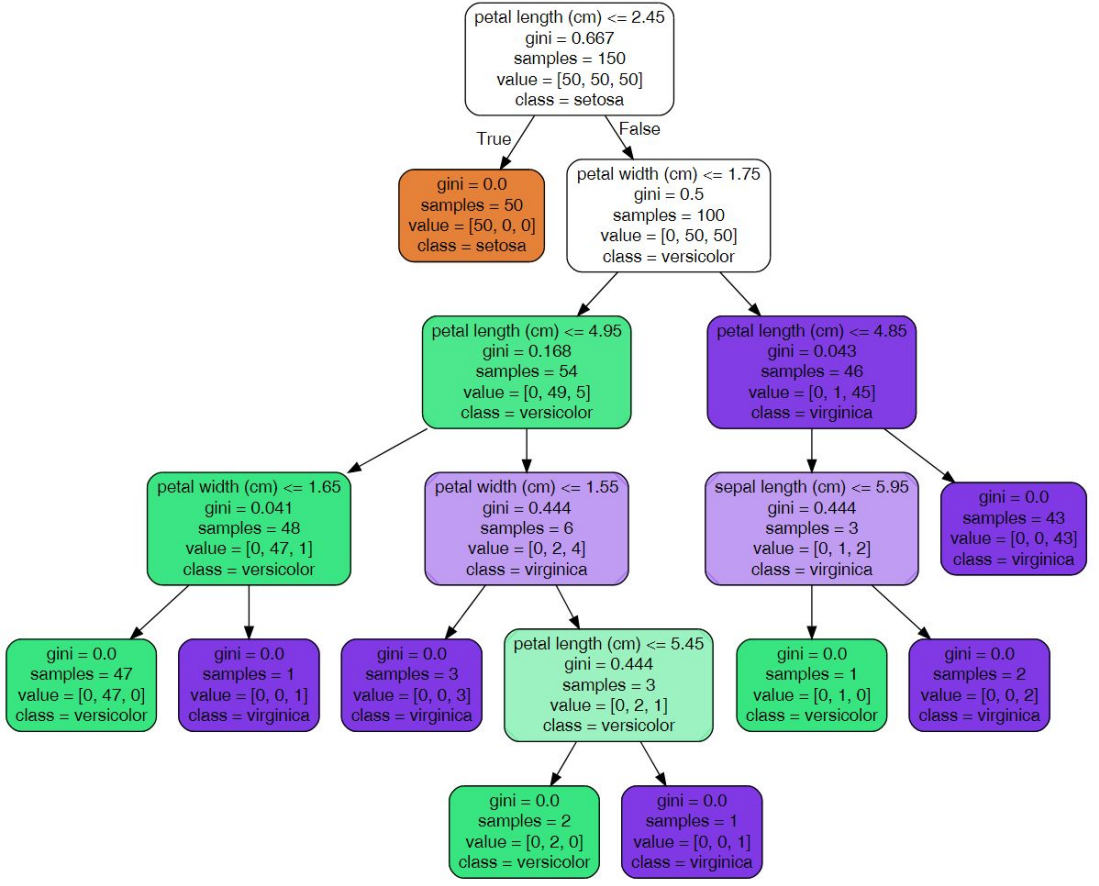
Se uma flor tem pétala com **3cm** de comprimento e **1.5cm** de largura:

- **90.7%** de chance de ser Versicolor
- **9.3%** de chance de ser Virginica
- **0.0%** de chance de ser Setosa

Exemplo: Íris



Exemplo: Íris

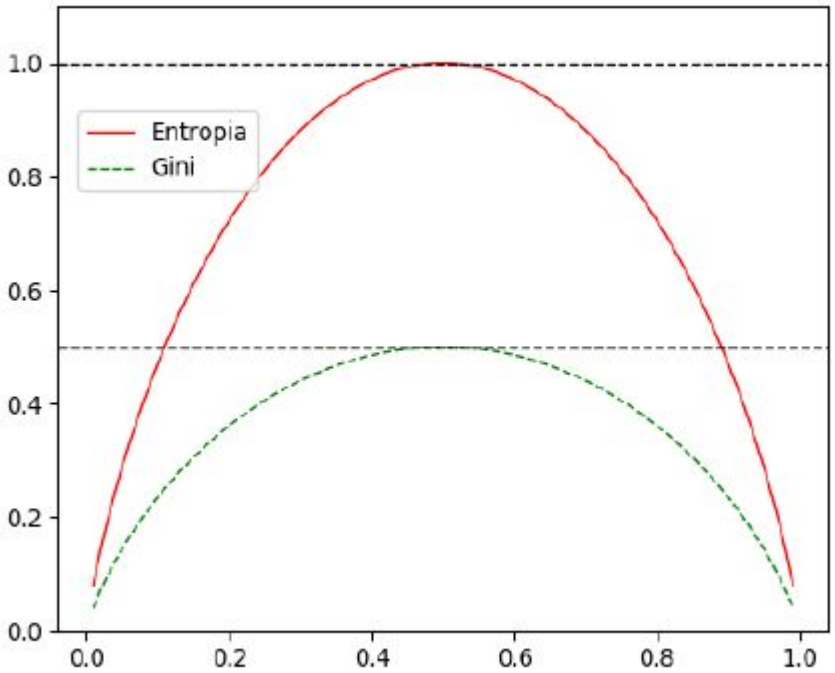


Gini e entropia

$$G = \sum_{k=1}^{K} p_k(1 - p_k)$$

$$E = -\sum_{k=1}^{K} p_k \ln(p_k)$$

em que pk é a proporção de elementos de cada uma das K classes



Poda

- Uma árvore muito grande pode resultar em overfitting.

- Limitar a geração de nós pode impedir o surgimento de um nó bom abaixo de um nó ruim

  Essa situação pode ser contornada através do processo de **poda**: uma árvore grande é construída, para que depois uma de suas sub-árvores seja escolhida.

  Porém, mesmo com a poda, a utilização de uma árvore sozinha não é uma técnica competitiva.

Ensemble

É um comitê de classificadores, que decide a classificação pela maioria dos votos.

É uma técnica promissora, especialmente se os classificadores forem suficientemente independentes.

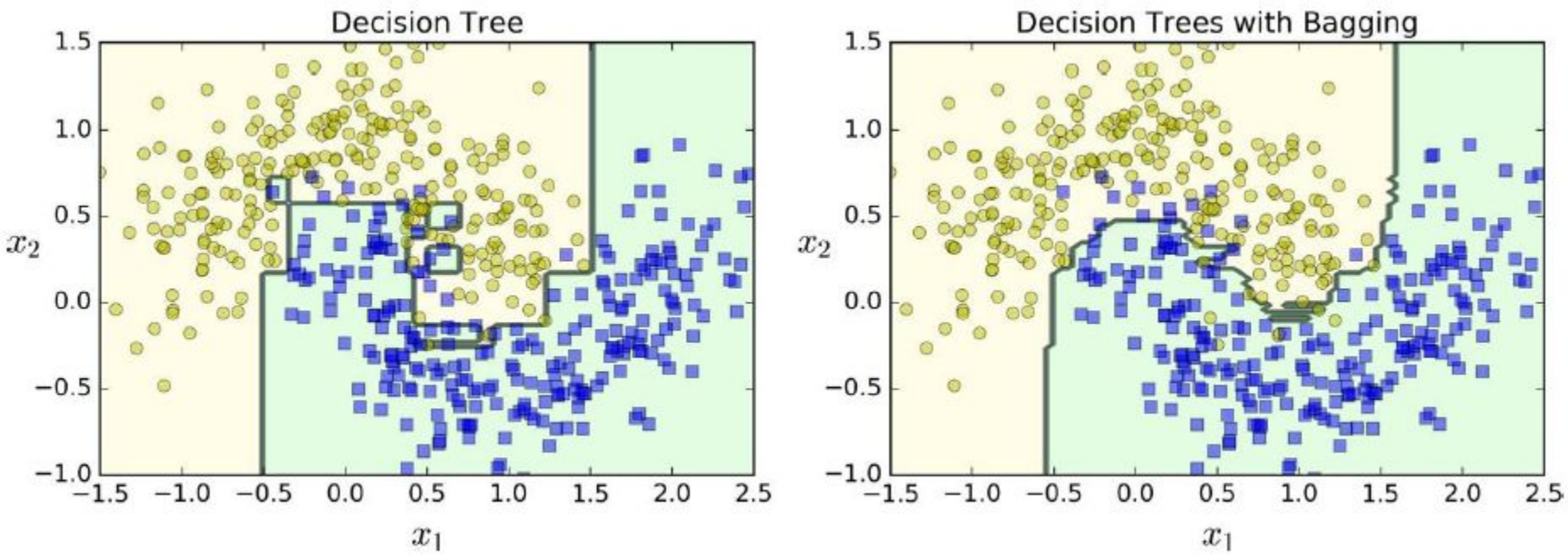
Exemplo: um bom ensemble poderia ser formado por KNN, SVM, árvore de decisão, etc.

Bagging

Bagging é uma abreviação para **Bootstrap Aggregating**.

- Pode ser utilizado em outros contextos que não o de árvores de decisão.
- São construídas K árvores profundas (baixo viés e alta variância) com reposição.
- A classe de um dado novo é determinada por votação (soft ou hard).
- A interpretação de uma árvore é perdida, porém ganha-se muito com a redução da variância.

Bagging: exemplo com K = 500



Fonte: A. Géron, Hands-on: Machine Learning with Scikit-Learn and TensorFlow.

Florestas aleatórias

É uma maneira de des-correlacionar as árvores.

- As árvores também são construídas com reposição.
- Quando uma divisão em um nó é analisada, apenas m (~ √p) preditores são considerados.
- Justificativa: se há um preditor muito forte, as árvores tendem a ser muito semelhantes.
- Árvores menores podem ser construídas.
- Florestas ajudam a medir a importância da característica avaliando a redução da impureza quando ela é utilizada.

Exemplo: Titanic

Sobreviveu	Classe	Sexo	Idade	Casal	Parentes	Passagem	Embarque
0	3	M	22	1	0	7.25	S
1	1	F	38	1	0	71.28	C
1	3	F	26	0	0	7.92	S
1	1	F	35	1	0	53.1	S
...	...	...	...	...	...	...	...
0	3	M	35	0	0	8.05	S

Dados de 891 passageiros (577 homens e 314 mulheres).

Exemplo: Titanic

Método	Acurácia
Árvore de decisão	0.7462
Bagging	0.8171
Floresta aleatória	0.8283

Conclusões

- Há outras técnicas para classificação que valem a pena ser estudadas.
- O processo de tratamento dos dados é, muitas vezes, mais complexo do que o do treinamento.
- Os algoritmos do KNN, Árvores, Florestas etc estão implementados no Scikit-Learn.
- Para cada problema, o usuário precisa escolher os parâmetros adequados.

Referências

- G. James, D. Witten, T. Hastie e R. Tibshirani, An Introduction to Statistical Learning with Applications in R, Springer.

- A. Géron, Mãos à Obra: Aprendizado de Máquina com Scikit-Learn & TensorFlow: Conceitos, Ferramentas e Técnicas para a Construção de Sistemas Inteligentes, Alta Books Editora.

Obrigado!

Lucas Pedroso

lucaspedroso@ufpr.br

CiDAMO