## 5. Árvores de Decisão, Seleção de Atributos e outros Classificadores

Nesta trilha você vai aprender:

* Sobre Classificador de Árvores de Decisão e Random Forests
* O que é Ganho de Informação ou Informação Mútua, e como empregar essa e outras medidas para a seleção de atributos
* Como empregar todos esses conceitos para a seleção de melhores modelos

Vamos partir aqui de um exemplo bastante simples de aplicação de árvores de decisão para, em seguida, explorarmos diversos conceitos que têm aplicação geral em modelos de aprendizado supervisionado, como entropia, informação mútua e ganho de informação. Você verá como esses conceitos são empregados para a seleção de atributos de aprendizado (*feature selection*).

No final você vai aprender o que são os *ensemble models* através de um dos mais importantes modelo *ensemble*, as Random Forests, e vamos adaptar o código da trilha anterior para de seleção de melhores hiperparâmetros para a seleção mais ampla de melhores modelos para classificação.

import numpy as np  
import pandas as pd  
import matplotlib.pyplot as plt  
%matplotlib inline  
from matplotlib.lines import Line2D  
import seaborn as sns

# Árvores de Decisão e o Método Partitivo

O classificador de k-vizinhos mais próximos é bastante poderoso. Ele pode lidar de várias classes de problemas e fazer classificações com limites de decisão não lineares (diferentemente da regressão logística). O Knn entretanto precisa armazenar todos os dados de treinamento e, quanto mais dados de treinamento, maior o custo do processamento e mais lento ele se torna pois é preciso calcular a distância para todas as entradas de treinamento, mesmo existindo estruturas de dados eficientes para pesquisa de vizinhos mais próximos. O uso de uma métrica de distância e a necessidade de atributos numéricos (ou transformados em numéricos) pode, em alguns casos, ser também uma limitação.

Você deve lembrar que os algoritmos de aprendizado de máquina buscam por padrões nos dados, uma estrutura inerente que é inerente a eles. O Knn, por exemplo, parte da ideia de que semelhança entre itens vizinhos e que pontos de dados de várias classes não são espalhados aleatoriamente. Pelo contrário, as classes aparecem em grupos mais ou menos homogêneos. Assim se você sabe que um ponto de teste cai em um grupo com vários pontos de uma mesma classe, você sabe a classe de seus vizinhos mesmo sem calcular as distâncias para cada deles. É, neste caso, suficiente saber que o ponto de teste está em uma área onde todos os vizinhos são de uma classe. As árvores de decisão exploram exatamente isso.

As árvores de decisão estão explorando exatamente isso. Aqui, não são armazenados os dados de treinamento mas uma estrutura de árvore *'modela'* os dados e divide recursivamente o espaço em regiões com classes semelhantes. Essa estrutura faz particionamentos sucessivos dos dados em que as partições são cada vez mais *puras*, isto é, são cada vez mais homogêneas no sentido das classes que contêm.

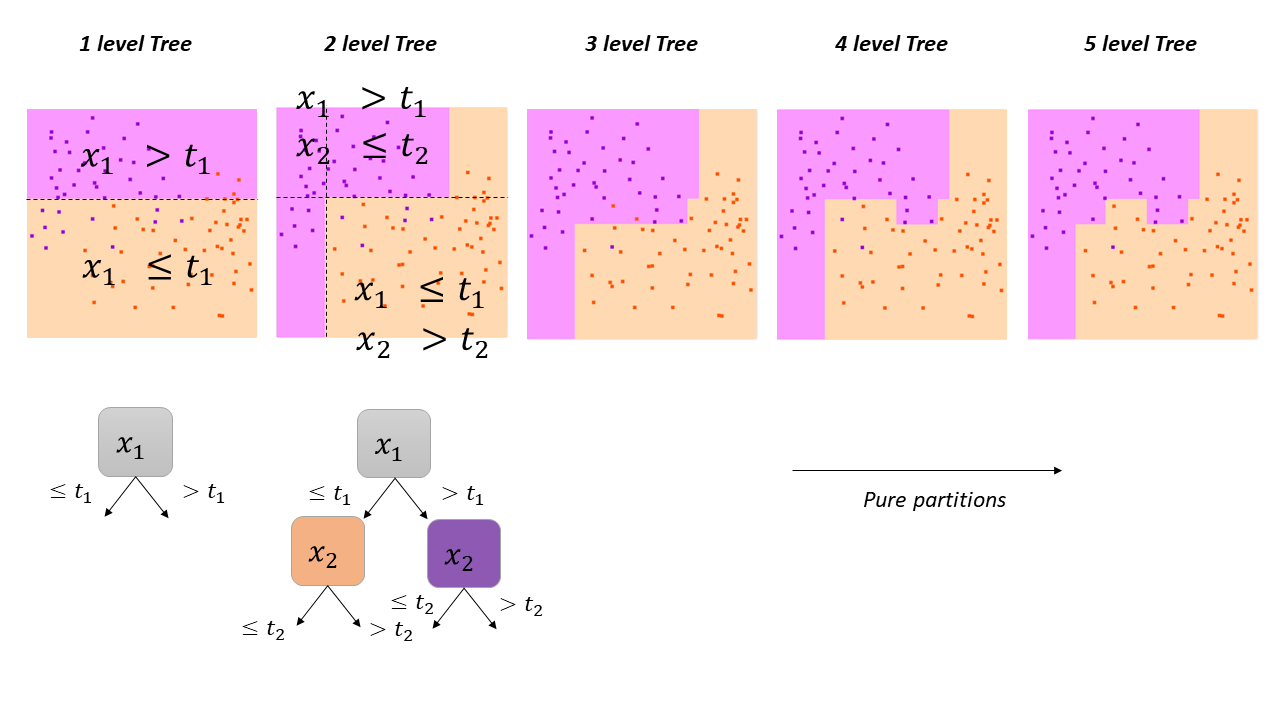


Figura 1. Particionamentos Sucessivos das Árvores de Decisão.

O nó raiz da árvore representa todo o conjunto de dados. Este conjunto é então dividido aproximadamente ao meio ao longo de uma dimensão (um atributo) por um limite simples. Todos os pontos que possuem um valor caem no nó filho direito e todos os demais no nó filho esquerdo. O limiar a dimensão é escolhida de forma que os nós filhos resultantes sejam mais puros em termos de associação de classe. O ideal é que todos os pontos positivos (por exemplo, pontos laranja na figura acima) caiam em um nó filho e todos os pontos negativos (roxos) no outro. Se for esse o caso, a árvore está pronta. Caso contrário, os nós folha são novamente divididos até que eventualmente todas as folhas, ou nós terminais, sejam puras com todos seus pontos com o mesmo rótulo ou não podem ser mais divididos (dois pontos idênticos com rótulos diferentes, o que em casos reais podemos de fato ter).

Uma vez que a árvore é construída os dados de treinamento não precisam ser mais armazenados (como no Knn) pois a árvore *captura* todo o padrão dos dados, e você pode pensar em como uma única fórmula y = a0 + a1 x modela um conjunto de dados em uma regressão linear. A árvore passa a ser nossa '*fórmula*' de predição, as entradas de teste simplesmente precisam descer da árvore até uma folha que contendo a predição da classe, e são muito eficientes. Além disso as árvores de decisão não requerem métricas porque as divisões são baseadas em limites dos valores dos atributos (ou ainda suas proporções ou probabilidades) e não em distâncias.

# Construindo uma Árvore de Decisão: Algoritmo de Hunt

Vamos começar com um exemplo e entender como construir uma árvore de decisão a partir dos dados.

# Exemplo livremente adaptado de:   
# Weinberger, Kilian. \*\*Machine Learning for Intelligent Systems\*\*, Lecture 28: Decision / Regression Trees   
#  
  
df = pd.read\_csv('http://meusite.mackenzie.br/rogerio/TIC/comics.csv',sep=';')  
display(df)  
  
cases = pd.read\_csv('http://meusite.mackenzie.br/rogerio/TIC/comics\_cases.csv',sep=';')  
display(cases)

Name sex smokes tie mask cape ears class  
0 Batman male no no yes yes yes good  
1 Robin male no no yes yes yes good  
2 Catwoman female no no yes no yes bad  
3 Joker male no no no no no bad  
4 Alfred male no yes no no no good  
5 Penguin male yes yes no no no bad

Name sex smokes tie mask cape ears class  
0 Batgirl female yes yes no yes no ?  
1 Riddler male yes no no no no ?

O algoritmo de Hunt cria uma árvore de decisão de maneira recursiva, particionando os registros de treinamento em subconjuntos sucessivamente mais puros. Seja o conjunto de registros de treinamento que atingem um nó . O procedimento recursivo é, então, o seguinte:

1. Se contém registros que pertencem à mesma classe , então é um nó folha rotulado como
2. Se for um conjunto vazio, então é um nó folha rotulado pela classe padrão,
3. Se contiver registros que pertencem a mais de uma classe, use um teste de atributo para dividir os dados em subconjuntos menores.

Ele aplica recursivamente o procedimento a cada subconjunto até que todos os registros no subconjunto pertençam à mesma classe. O algoritmo de Hunt assume que cada combinação de conjuntos de atributos possui um rótulo de classe exclusivo durante o procedimento. Se todos os registros associados a tiverem valores de atributo idênticos, exceto para o rótulo da classe, não será possível dividir esses registros no futuro. Nesse caso, o nó é classificado como um nó folha com o mesmo rótulo de classe que a classe principal de registros de treinamento associados a este nó.

O procedimento talvez seja mais claro de *ver* do que *ler*...

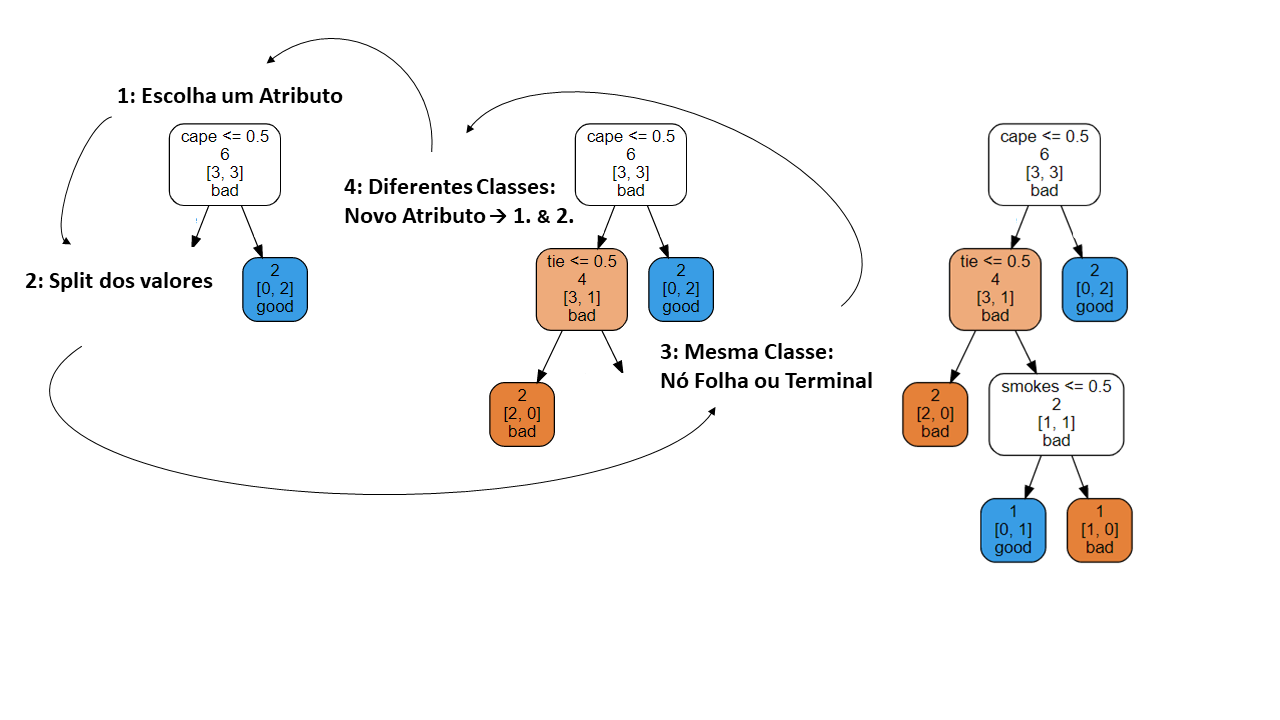


Figura 2. Construção de uma Árvore de Decisão pelo algoritmo de Hunt.

No exemplo acima escolhemos arbitrariamente iniciar o particionamento dos dados pelo atributo cape, como para todos os casos cape = yes temos somente alunos good esse ramo se torna um nó folha ou terminal, e dados que seguem nesse ramo já estão classificados e não serão mais considerados. Para os casos cape = no há alunos good e bad e, portanto, não podemos decidir e é necessário adicionarmos um novo nó com outro atributo. Note que nesse ramo só teremos agora dados com cape = no. Escolhemos novamente de modo arbitrário o atributo tie de realizamos novamente o *split* dos dados. Agora, os casos com tie = no (e cape = no!, pois é o ramo que estamos) são todos da mesma classe e são, portanto, um nó terminal. Para tie = yes (e ainda cape = no!) precisamos adicionar ainda um novo particionamento com o atributo smokes.

Note que o que a estrutura de árvore faz é simplesmente definir sucessivos particionamentos dos dados buscando partições cada vez mais *puras* (somente classes *good* ou *bad*).

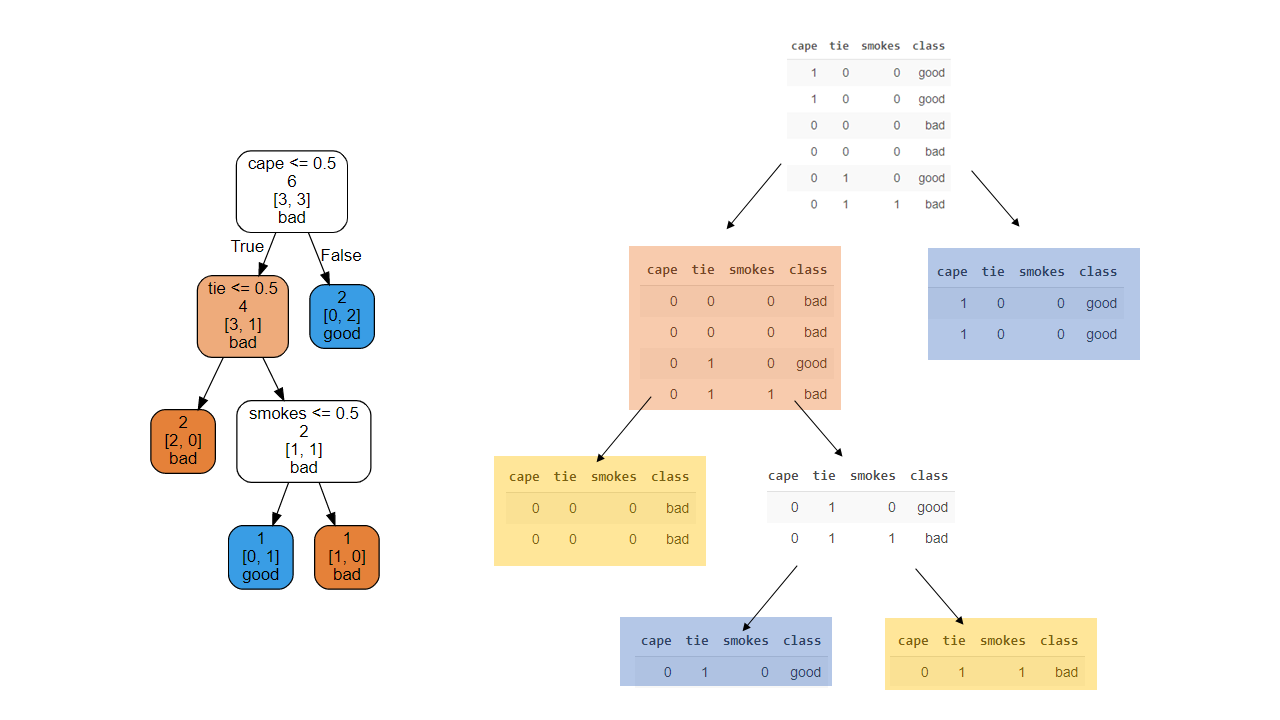


Figura 3. Árvore de Decisão e o particionamento dos dados.

No exemplo acima fizemos escolhas arbitrárias dos atributos para a construção da árvore. Se escolhêssemos outros atributos ou em uma ordem diferentes chegaríamos a outras árvores de decisão que poderíamos igualmente empregar para a classificação dos dados.

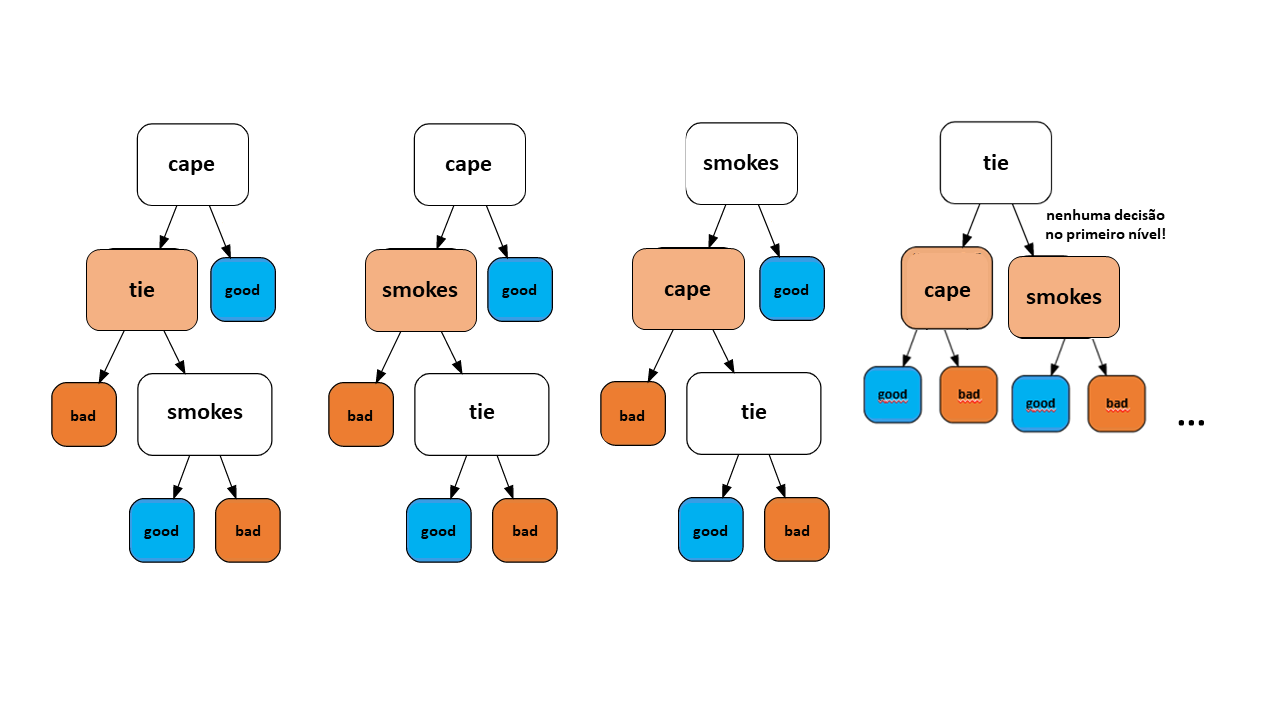


Figura 4. Diferentes possíveis Árvores de Decisão para um mesmo conjunto de dados.

Mas, como você pode ver, existem árvores melhores e piores. O atributo tie na raiz, por exemplo, não permite particionarmos dados de forma a eliminar quaisquer dos dados e, portanto, *ele não fornece qualquer informação útil para determinar a classe dos dados*. A ideia é de árvores mais eficientes possam *otimizar* a cada nível a *pureza* das partições. Você entenderá como podemos fazer isso mais adiante em que empregaremos o *ganho de informação* como uma medida para definir melhores partições dos dados a cada nível.

# Outros Algoritmos

Os algoritmos de Árvore de Decisão constroem as Árvores do mesmo modo recursivo que construímos aqui. Todos eles buscam particionamentos que maximizem a decisão, tornando a Árvores e os caminhos que temos que percorrer menores. Eles buscam a cada nível da Árvore (partição) uma maior informação para “pureza” das partições dos dados em termos de sua classe.

Diferentes algoritmos de Árvores de Decisão e seus critérios para seleção dos nós:

* ID3 algorithm (Claude Shannon Entropy): Ganho do de Informação
* C4.5 algorithm: Gain ratio
* CART algorithm: Gini index

Seja qual for o método empregado a ideia é a mesma. Existem particionamentos que alteram pouco a distribuição dos dados entre as classes e trazem, portanto, pouca informação adicional para a classificação, e outros que proporcionam informação maior para a decisão.

O pseudo código abaixo é o algoritmo clássico apresentado por Mitchell, Tom Michael (1997) em seu clássico livro de Aprendizado de Máquina.

%%script false  
# Adaptado de, Mitchell, Tom Michael (1997). Machine Learning.   
  
def ID3 (Examples, Target\_Attribute, Attributes):   
   
 Create a root node for the tree  
 if all examples are positive, return the single-node tree Root, with label = +.  
 if all examples are negative, return the single-node tree Root, with label = -.  
   
 if number of predicting attributes is empty:  
 return the single node tree Root, with label = most common value of the target attribute in the examples.  
 else:  
 A ← The Attribute that best classifies examples.  
 Decision Tree attribute for Root = A.  
 for each possible value, vi, of A,  
 Add a new tree branch below Root, corresponding to the test A = vi.  
 Let Examples(vi) be the subset of examples that have the value vi for A  
 if Examples(vi) is empty  
 below this new branch add a leaf node with label = most common target value in the examples  
 else below this new branch add the subtree ID3 (Examples(vi), Target\_Attribute, Attributes – {A})  
 End

# Decision Tree, scikit-learn

Antes de prosseguirmos vamos explorar o classificador de árvore de decisão do scikit-learn no nosso exemplo de brinquedo.

Os modelos de árvore de decisão podem lidar com variáveis categóricas sem codificá-las de uma só vez. Veja que nosso exemplo os valores atributos tinham valores 'yes' ou 'no' e, em nenhum momento precisamos supor ou transformá-los em numéricos. Entretanto as implementações mais populares de árvores de decisão, incluindo o scikit-learn não honram esse fato e requerem o encode dos dados categóricos, apesar da degradação do desempenho do modelo de árvore que isso traz. Desse modo precisamos proceder o encode dos dados para empregarmos o classificador de árvore de decisão do scikit-learn. Mas lembre-se, que isso é uma limitação apenas da implementação.

from sklearn.preprocessing import OneHotEncoder, LabelEncoder  
  
df\_labels = df.drop(columns=['Name','class']).apply(LabelEncoder().fit\_transform)  
cases\_labels = cases.drop(columns=['Name','class']).apply(LabelEncoder().fit\_transform)  
  
display( pd.concat([df.Name, df\_labels, df['class']], axis=1) )  
display( pd.concat([cases.Name, cases\_labels, cases['class']], axis=1) )

Name sex smokes tie mask cape ears class  
0 Batman 1 0 0 1 1 1 good  
1 Robin 1 0 0 1 1 1 good  
2 Catwoman 0 0 0 1 0 1 bad  
3 Joker 1 0 0 0 0 0 bad  
4 Alfred 1 0 1 0 0 0 good  
5 Penguin 1 1 1 0 0 0 bad

Name sex smokes tie mask cape ears class  
0 Batgirl 0 0 1 0 1 0 ?  
1 Riddler 1 0 0 0 0 0 ?

Neste exemplo simples, havendo para todos atributos somente dois valores, podemos fazer o LabelEncoder dos dados, tendo o mesmo resultado que teríamos ao empregar o OneHotEncoder(drop='if-binary').

Como estamos interessados apenas em *explorar* o modelo, por hora podemos deixar de lado a separação de dados de treinamento e teste. O estimador DecisionTreeClassifier() segue então a mesma forma de uso que empregamos até agora.

from sklearn.tree import DecisionTreeClassifier  
  
X = df\_labels  
y = df['class']  
  
clf = DecisionTreeClassifier(criterion='entropy')   
  
clf.fit(X,y)  
y\_pred = clf.predict(X)  
print( clf.score(X,y) )

1.0

E podemos então fazer predições a partir do modelo construído.

Uma imagem contendo Interface gráfica do usuário

Descrição gerada automaticamente

Diagrama

Descrição gerada automaticamente

Figura 5. Personagens para predição. (Fonte: imagens de divulgação, <https://universoheroico.com.br> e <https://images6.fanpop.com>)

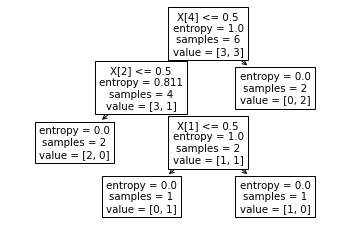
y\_pred = clf.predict(cases\_labels)  
cases['class'] = y\_pred  
display( cases )

Name sex smokes tie mask cape ears class  
0 Batgirl female yes yes no yes no good  
1 Riddler male yes no no no no bad

Uma das grandes vantagens dos modelos de Árvore de Decisão é sua *interpretabilidade*. Um tema cada vez mais exigido dos modelos de aprendizado de máquina à medida que eles são mais e mais empregados em decisões que influenciam o dia a dia e o destino das pessoas (por exemplo, aplicações na no processo seletivo das empresas ou a previsibilidade de crimes).

Essa interpretabilidade é em grande parte pela possibilidade de podermos *enxergar* como é tomada a decisão do classificador observando a estrutura da árvore ou, ao menos, a cadeia de particionamentos criada. O scikit-learn fornece algumas formas de visualização da árvore criada sendo a mais simples, o plot\_tree().

from sklearn import tree  
  
tree.plot\_tree(clf)  
plt.show()



Mas resultados muito melhores podem ser obtidos com o pacote graphviz.

import graphviz   
dot\_data = tree.export\_graphviz(clf, out\_file=None,   
 feature\_names=list(df\_labels.columns.values),   
 class\_names=list(sorted(df['class'].unique())),   
 filled=True, rounded=True,   
 special\_characters=False,  
 proportion=False, impurity=False, node\_ids=False,label=None)   
graph = graphviz.Source(dot\_data)   
graph.render('graph\_cartoons') # para gravação em .pdf  
graph

<graphviz.files.Source at 0x7f05b7c6b250>

Revisando a árvore acima podemos reproduzir os particionamentos criados e notar que o 'aprendizado' do modelo pode ser facilmente interpretado por uma série de condições (*ifs*) encadeados.

df\_tree = pd.concat([df\_labels, df['class']], axis=1)  
  
print('Root level')  
print(60\*'-')  
display(df\_tree)  
print('Continue Next Level...')  
display(df\_tree[df\_tree.cape<=0.5])  
print('Terminal')  
display(df\_tree[df\_tree.cape>0.5])  
  
print('\n\n\nSecond level')  
print(60\*'-')  
display(df\_tree[df\_tree.cape<=0.5])  
print('Continue Next Level...')  
display(df\_tree[(df\_tree.cape<=0.5) & (df\_tree.tie>0.5)])  
print('Terminal')  
display(df\_tree[(df\_tree.cape<=0.5) & (df\_tree.tie<=0.5)])  
  
print('\n\n\nThird level')  
print(60\*'-')  
display(df\_tree[(df\_tree.cape<=0.5) & (df\_tree.tie>0.5)])  
print('Terminal')  
display(df\_tree[(df\_tree.cape<=0.5) & (df\_tree.tie>0.5) & (df\_tree.smokes<=0.5) ])  
print('Terminal')  
display(df\_tree[(df\_tree.cape<=0.5) & (df\_tree.tie>0.5) & (df\_tree.smokes>0.5)])

Root level  
------------------------------------------------------------

sex smokes tie mask cape ears class  
0 1 0 0 1 1 1 good  
1 1 0 0 1 1 1 good  
2 0 0 0 1 0 1 bad  
3 1 0 0 0 0 0 bad  
4 1 0 1 0 0 0 good  
5 1 1 1 0 0 0 bad

Continue Next Level...

sex smokes tie mask cape ears class  
2 0 0 0 1 0 1 bad  
3 1 0 0 0 0 0 bad  
4 1 0 1 0 0 0 good  
5 1 1 1 0 0 0 bad

Terminal

sex smokes tie mask cape ears class  
0 1 0 0 1 1 1 good  
1 1 0 0 1 1 1 good

Second level  
------------------------------------------------------------

sex smokes tie mask cape ears class  
2 0 0 0 1 0 1 bad  
3 1 0 0 0 0 0 bad  
4 1 0 1 0 0 0 good  
5 1 1 1 0 0 0 bad

Continue Next Level...

sex smokes tie mask cape ears class  
4 1 0 1 0 0 0 good  
5 1 1 1 0 0 0 bad

Terminal

sex smokes tie mask cape ears class  
2 0 0 0 1 0 1 bad  
3 1 0 0 0 0 0 bad

Third level  
------------------------------------------------------------

sex smokes tie mask cape ears class  
4 1 0 1 0 0 0 good  
5 1 1 1 0 0 0 bad

Terminal

sex smokes tie mask cape ears class  
4 1 0 1 0 0 0 good

Terminal

sex smokes tie mask cape ears class  
5 1 1 1 0 0 0 bad

# Entropia e Ganho de Informação

Voltando ao nosso exemplo a árvore construída foi a mesma que construímos antes (lembre-se, poderíamos ter outras árvores construídas) mas a escolha não foi aqui arbitrária. A escolha dos atributos para a definição dos nós da árvore seguiu o critério empregar os atributos com maior *ganho de informação* para maximizar a *pureza* dos particionamentos.

Você pode entender o *ganho de informação* como uma medida de quanto a informação de um atributo contribui para diminuir a incerteza sobre outro, assim, o ganho de informação de cape para determinar class, , deve ser maior que o ganho de informação dos demais atributos.

Mas para entendermos o ganho de informação precisamos antes entender o conceito de entropia.

## Entropia

Na teoria da informação, a entropia é uma medida de quantidade de informação introduzido por Claude Shannon em seu artigo de 1948 *A Mathematical Theory of Communication*. A entropia de um atributo é o nível médio de informação, surpresa, ou ainda incerteza, inerente aos resultados possíveis desse atributo. Ela é definida como:

onde é a probabilidade dos valores de .

Um *bit*, por exemplo, é a unidade de informação exatamente por que carrega *1* de entropia:

onde e são respectivamente as probabilidades do bit e .

Assim, podemos ver também que a entropia de uma constante é zero, pois não há qualquer incerteza sobre seu valor:

e é exatamente isso que queremos buscar em um particionamento ao construírmos uma árvore de decisão.

Podemos empregar o pacote scipy para determinar a entropia de dados com base nas probabilidades dos diferentes valores e, abaixo, você pode ver a entropia do *bit* e de uma constante.

from scipy.stats import entropy  
E = entropy( [1/2, 1/2] , base=2 )  
E

1.0

E = entropy( 1 , base=2 )  
E

0.0

## Ganho de Informação

O ganho de informação de um atributo com relação a uma variável classe objetivo (, Target) é, portanto, uma medida de quanto a informação desse atributo diminui a incerteza sobre a classe dos dados, ou seu valor no caso de uma regressão. O ganho de informação pode ser então expresso em termos da entropia como:

Onde é a entropia do atributo objetivo e a somatória é a média ponderada da entropia do atributo target para os diferentes conjuntos de dados com valor .

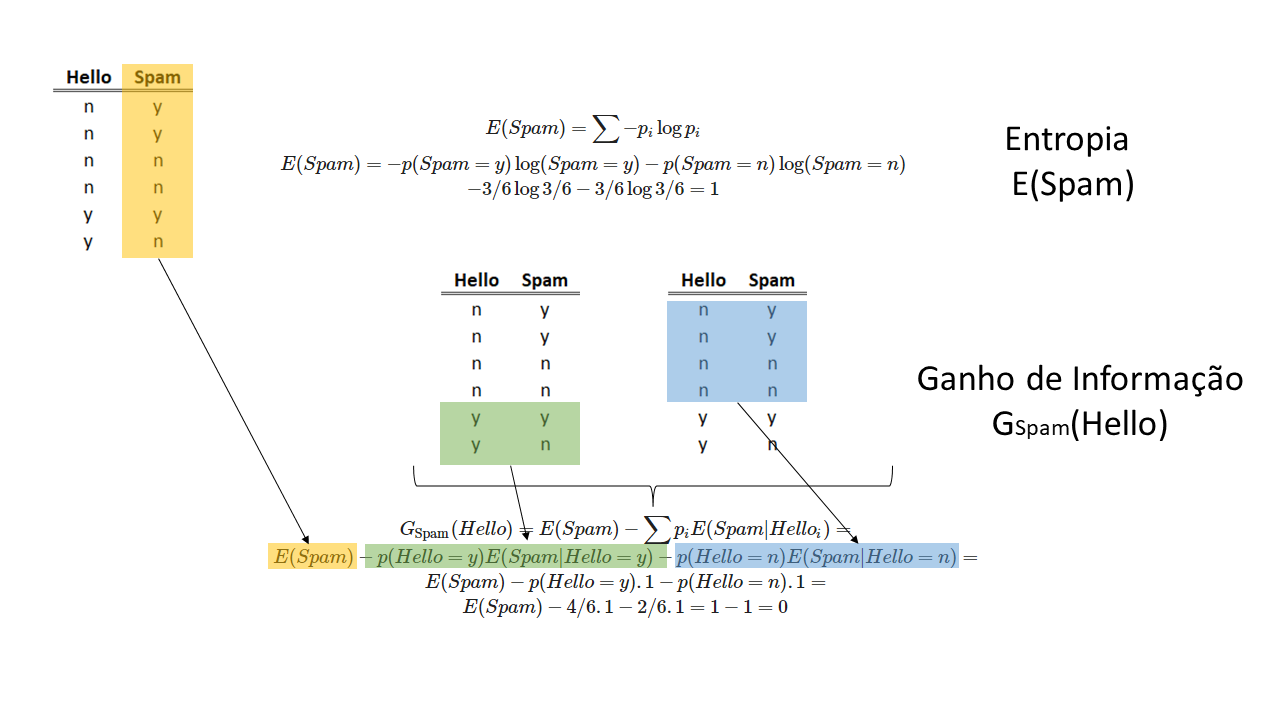


Figura 6. Exemplo do Cálculo de Entropia e Ganho de Informação.

O exemplo acima mostra o cálculo do ganho de informação do atributo *Hello*, indicando emails iniciando ou não por *Hello*, para a determinação da classe desses emails, *Spam* ou *Not Spam*.

# Ganho de Informação, Índice Gini e Classification Error

Você deve ter notado que ao empregarmos o classificador DecisionTreeClassifier() empregamos um parâmetro criterion, DecisionTreeClassifier(criterion='entropy'). Isso por que existem outras métricas de cálculo do ganho de informação como o Índice Gini e o Classification Error sendo o padrão do scikit-learn o emprego do Índice Gini, DecisionTreeClassifier(criterion='gini). Essas medidas são definidas como:

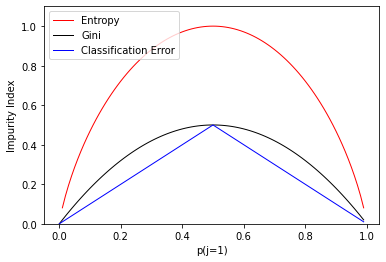
e

E são empregadas dos mesmo modo que a entropia para o cálculo do ganho de informação:

e

Tanto a entropia como o índice gini e o classification error são igualmente medidas de impuridade (0, para nenhuma incerteza dos dados) sendo apenas funções com diferentes medidas mas que preservam a mesma ordem de medida, isto é, um maior valor de entropia sempre leva a maiores valores do índice gini e do classification error, como você pode observar abaixo.

# you can skip this code!  
  
# Adaptado de: https://www.bogotobogo.com/python/scikit-learn/scikt\_machine\_learning\_Decision\_Tree\_Learning\_Informatioin\_Gain\_IG\_Impurity\_Entropy\_Gini\_Classification\_Error.php  
import matplotlib.pyplot as plt  
import numpy as np  
  
def gini(p):  
 return (p)\*(1 - (p)) + (1 - p)\*(1 - (1-p))  
  
def entropy(p):  
 return - p\*np.log2(p) - (1 - p)\*np.log2((1 - p))  
  
def classification\_error(p):  
 return 1 - np.max([p, 1 - p])  
  
x = np.arange(0.0, 1.0, 0.01)  
ent = [entropy(p) if p != 0 else None for p in x]  
c\_err = [classification\_error(i) for i in x]  
  
fig = plt.figure()  
ax = plt.subplot(111)  
  
for j, lab, c, in zip(  
 [ent, gini(x), c\_err],  
 ['Entropy', 'Gini', 'Classification Error'],  
 ['red', 'black', 'blue']):  
 line = ax.plot(x, j, label=lab, lw=1, color=c)  
  
ax.legend(loc='upper left', ncol=1, fancybox=True, shadow=False)  
  
plt.ylim([0, 1.1])  
plt.xlabel('p(j=1)')  
plt.ylabel('Impurity Index')  
plt.show()



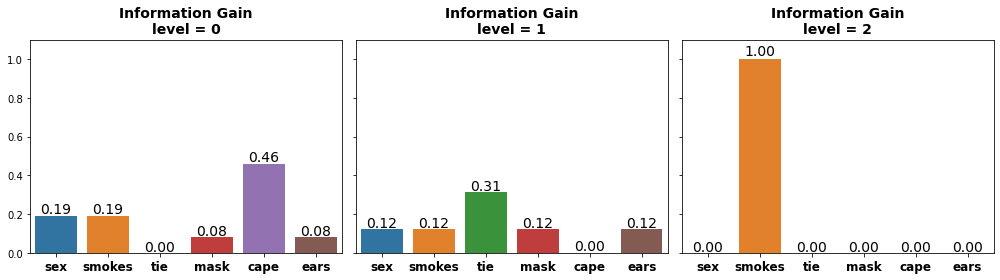
# Entropia e Ganho de Informação com Python

Você pode agora empregar as definições acima para verificar o ganho de informação de cada atributo do nosso exemplo de brinquedo.

def show\_values(axs, orient="v", space=.01):  
 def \_single(ax):  
 if orient == "v":  
 for p in ax.patches:  
 \_x = p.get\_x() + p.get\_width() / 2  
 \_y = p.get\_y() + p.get\_height() + (p.get\_height()\*0.01)  
 value = '{:.2f}'.format(p.get\_height())  
 ax.text(\_x, \_y+0.01, value, ha="center", fontsize=14)   
  
 if isinstance(axs, np.ndarray):  
 for idx, ax in np.ndenumerate(axs):  
 \_single(ax)  
 else:  
 \_single(axs)

# you can skip this code!  
  
from scipy.stats import entropy  
  
def informationgain(df,target,base=2):  
 E = entropy( df[target].value\_counts() / len(df) , base=base )  
  
 E\_x = {}  
  
 for c in df.drop(columns=[target]):  
 E\_x[c] = []  
 for v in df[c].unique():  
 e = len(df[df[c]==v])/len(df) \* entropy( df[df[c]==v]['class'].value\_counts() / len(df[df[c]==v]), base=base )  
 E\_x[c].append(np.round(e,2))  
  
 ig = {}  
  
 for c, p in E\_x.items():  
 ig[c] = E - sum(p)  
  
 return ig

# you can skip this code!  
  
ig\_root = informationgain(df=df.drop(columns=['Name']),target='class')  
ig\_level\_1 = informationgain(df=df.drop(columns=['Name'])[df.cape == 'no'],target='class')  
ig\_level\_2 = informationgain(df=df.drop(columns=['Name'])[ (df.cape == 'no') & (df.tie == 'yes') ],target='class')  
  
f , ax = plt.subplots(1,3,figsize=(14,4),sharey=True)  
  
for i, ig\_ in enumerate([ ig\_root, ig\_level\_1, ig\_level\_2 ]):  
 sns.barplot(x=list(ig\_.keys()), y=list(ig\_.values()), ax=ax[i])  
 ax[i].set\_title('Information Gain\nlevel = ' + str(i), fontsize=14, weight='bold')  
 ax[i].set\_ylim([0,1.1])  
  
 ax[i].set\_xticklabels(ax[i].get\_xticklabels(), fontsize=12, weight='bold')  
 show\_values(ax[i])  
  
plt.tight\_layout()  
plt.show()



Como você pode ver as escolhas dos atributos em cada nível segue exatamente o maior ganho de informação recalculado a cada nível da árvore.

# Informação Mútua e Seleção de Atributos

Toda essa discussão sobre o ganho de informação talvez não fosse tão importante se isso estivesse limitado ao uso em modelos de Árvore de Decisão. O ganho de informação, entretanto, é útil para a construção de muitos outros modelos pois nos dá um meio de avaliarmos atributos que são potencialmente mais preditores que outros, algo que ainda não havíamos explorado. Dentro da seleção de atributos para modelos preditivos, ou *feature selection*, o ganho de informação é normalmente denominado informação mútua (*mutual information*), correspondendo ao mesmo cálculo que você viu do ganho de informação empregando a Entropia, para entropia com base ( ou ).

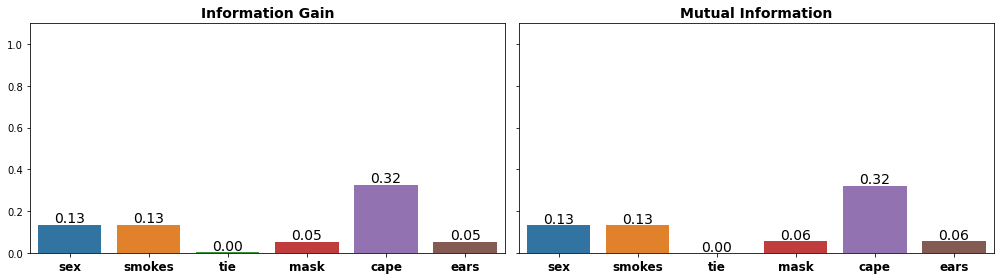
df\_tree

sex smokes tie mask cape ears class  
0 1 0 0 1 1 1 good  
1 1 0 0 1 1 1 good  
2 0 0 0 1 0 1 bad  
3 1 0 0 0 0 0 bad  
4 1 0 1 0 0 0 good  
5 1 1 1 0 0 0 bad

from sklearn.feature\_selection import mutual\_info\_classif  
  
print(mutual\_info\_classif(df\_tree.drop(columns=['class']), df\_tree['class'], discrete\_features=True))

[0.13230412 0.13230412 0. 0.05663301 0.31825708 0.05663301]

f , ax = plt.subplots(1,2,figsize=(14,4),sharey=True)  
  
ig\_root = informationgain(df=df.drop(columns=['Name']),target='class',base=None)  
  
i = 0  
sns.barplot(x=list(ig\_root.keys()), y=list(ig\_root.values()), ax=ax[i])  
ax[i].set\_title('Information Gain', fontsize=14, weight='bold')  
ax[i].set\_ylim([0,1.1])  
  
ax[i].set\_xticklabels(ax[i].get\_xticklabels(), fontsize=12, weight='bold')  
show\_values(ax[i])  
  
mutualinfo = mutual\_info\_classif(df\_tree.drop(columns=['class']), df\_tree['class'], discrete\_features=True)  
  
i = 1  
sns.barplot(x=df\_tree.drop(columns=['class']).columns, y=mutualinfo , ax=ax[i])  
ax[i].set\_title('Mutual Information', fontsize=14, weight='bold')  
ax[i].set\_ylim([0,1.1])  
  
ax[i].set\_xticklabels(ax[i].get\_xticklabels(), fontsize=12, weight='bold')  
show\_values(ax[i])  
  
plt.tight\_layout()  
plt.show()



# Seleção de Atributos

A seleção de atributos para construção de um modelo tem muitos aspectos, incluindo muitos aspectos no domínio do problema (e não da solução!), e não seria possível abordarmos todos aqui. Mas vale a pena vermos algumas dessas técnicas e aprender a empregar o scikit-learn nessa tarefa. Depois de abordagens que envolvem basicamente o entendimento dos dados, como excluir atributo único para cada instância ou um atributo com muitos dados ausentes, mas antes de empregar a informação mútua para a seleção de atributos, vale a pena removermos atributos de pouca variância: eles

As classes do sklearn.feature\_selection podem ser usadas para seleção de atributos e redução de dimensionalidade em conjuntos de amostras, seja para melhorar o *score* dos estimadores ou para impulsionar seu desempenho em conjuntos de dados de dimensões muito altas.

## Removendo dados com pouca variância

O VarianceThreshold permite remover todos os atributos com variância abaixo de um limite e, por padrão, remove todos atributos de variância zero (mesmo valor em todas as amostras), são os dados que não carregam nenhuma informação ou, como você viu entropia . O

df\_tree.var()

sex 0.166667  
smokes 0.166667  
tie 0.266667  
mask 0.300000  
cape 0.266667  
ears 0.300000  
dtype: float64

from sklearn.feature\_selection import VarianceThreshold  
  
X = df\_tree.drop(columns=['class'])  
sel = VarianceThreshold(threshold=0.1)  
sel.fit\_transform(X)

array([[1, 0, 0, 1, 1, 1],  
 [1, 0, 0, 1, 1, 1],  
 [0, 0, 0, 1, 0, 1],  
 [1, 0, 0, 0, 0, 0],  
 [1, 0, 1, 0, 0, 0],  
 [1, 1, 1, 0, 0, 0]])

No nosso exemplo há poucos atributos e as variâncias próximas não permitem selecionar atributos para remoção. Mas o exemplo vale para mostrar como empregar o estimador.

## Selecionando atributos com o SelectKBest

A seleção univariada de recursos funciona selecionando os melhores atributos com base em testes estatísticos univariados, o que inclui a informação mútua. São 3 métodos, mas que tem o mesmo tipo de princípio:

* SelectKBest remove tudo, exceto os atributos de maior pontuação
* SelectPercentile remove todos os atributos, exceto a maior porcentagem de pontuação especificada
* GenericUnivariateSelectpermite realizar a seleção univariada de recursos com uma estratégia configurável

Aplicando o SelectKBest ao nosso exemplo teríamos:

df\_tree

sex smokes tie mask cape ears class  
0 1 0 0 1 1 1 good  
1 1 0 0 1 1 1 good  
2 0 0 0 1 0 1 bad  
3 1 0 0 0 0 0 bad  
4 1 0 1 0 0 0 good  
5 1 1 1 0 0 0 bad

from sklearn.feature\_selection import SelectKBest  
from sklearn.feature\_selection import chi2  
from sklearn.feature\_selection import mutual\_info\_classif  
  
X = df\_tree.drop(columns=['class'])  
y = df\_tree['class']  
  
select\_features = SelectKBest(mutual\_info\_classif, k=3).fit(X, y)  
print( select\_features.get\_support() )  
print( X.columns[select\_features.get\_support()] )  
  
X\_new = select\_features.transform(X)  
  
pd.DataFrame( X\_new, columns= X.columns[select\_features.get\_support()] )

[ True False False False True True]  
Index(['sex', 'cape', 'ears'], dtype='object')

sex cape ears  
0 1 1 1  
1 1 1 1  
2 0 0 1  
3 1 0 0  
4 1 0 0  
5 1 0 0

Esses métodos podem ser empregados tanto para modelos de classificação como de regressão com diferentes métricas de pontuação:

* Para a regressão: f\_regression, mutual\_info\_regression
* Para a classificação: chi2, f\_classif, mutual\_info\_classif

E você deve ter cuidado para não empregar métricas de classificação em modelos de regressão e vice-versa.

## Limitações

Note que se fizermos a seleção de features como acima, teríamos eliminado o atributo tie, assim como o atributo smokes pelo critério de variância mínima, dois atributos que foram empregados no modelo de árvore. Isso por que esses critérios somente levam em consideração a variância e a informação mútua unicamente do atributo, sem levar em conta possíveis relações dos atributos para a predição. Assim, é importante você ter bastante cuidado ao remover dados de modelo a partir de critérios univariados.

## Eliminação recursiva de atributos

Uma última técnica que trataremos aqui é a eliminação recursiva de atributos. A maior parte dos estimadores fornece, após o treinamento, a importância de cada atributo (como estimador.coef\_ ou estimador.feature\_importances\_). A eliminação recursiva de recursos (RFE) consiste em selecionar atributos considerando recursivamente conjuntos cada vez menores na quantidade de atributos. Primeiro, o estimador é treinado no conjunto inicial, em seguida, os atributos menos importantes são removidos do conjunto atual repetindo-se o procedimento recursivamente até um número desejado.

RFECV executa RFE em um loop de validação cruzada para encontrar o número ideal de recursos e o exemplo poderá ser útil para aplicar em conjuntos de dados mais complexos que o nosso conjunto de dados de brinquedo.

from sklearn.tree import DecisionTreeClassifier  
from sklearn.feature\_selection import RFECV  
  
X = df\_labels  
y = df['class']  
  
clf = DecisionTreeClassifier(criterion='entropy')   
clf.fit(X,y)  
  
min\_features\_to\_select = 1 # Minimum number of features to consider  
rfecv = RFECV(  
 estimator=clf,  
 step=1,  
 cv=2,  
 scoring="accuracy",  
 min\_features\_to\_select=min\_features\_to\_select,  
)  
rfecv.fit(X, y)  
  
print("Optimal number of features : %d" % rfecv.n\_features\_)  
  
# Plot number of features VS. cross-validation scores  
plt.figure()  
plt.xlabel("Number of features selected")  
plt.ylabel("Cross validation score (accuracy)")  
plt.plot(  
 range(min\_features\_to\_select, len(rfecv.grid\_scores\_) + min\_features\_to\_select),  
 rfecv.grid\_scores\_,  
)  
plt.show()

Optimal number of features : 1

/usr/local/lib/python3.7/dist-packages/sklearn/base.py:446: UserWarning: X does not have valid feature names, but RFECV was fitted with feature names  
 "X does not have valid feature names, but"  
/usr/local/lib/python3.7/dist-packages/sklearn/utils/deprecation.py:103: FutureWarning: The `grid\_scores\_` attribute is deprecated in version 1.0 in favor of `cv\_results\_` and will be removed in version 1.2.  
 warnings.warn(msg, category=FutureWarning)



pd.DataFrame({'atributtes': df\_labels.columns, 'feature\_importances': clf.feature\_importances\_}).sort\_values('feature\_importances',ascending=False)

atributtes feature\_importances  
4 cape 0.459148  
1 smokes 0.333333  
2 tie 0.207519  
0 sex 0.000000  
3 mask 0.000000  
5 ears 0.000000

# Ensemble Models e Random Forests

*Ensemble models* são combinações de modelos. Eles são as mais eficazes abordagens em aprendizado de máquina, geralmente alcançando melhor desempenho que um único modelo. Trazem um custo algorítmico alto, uma maior complexidade e muitas vezes dificultam a interpretabilidade dos modelos. Existem duas motivações principais por trás da aprendizagem com *Ensemble models*: calcular a média das medidas pode fornecer uma estimativa mais confiável e estável e, portanto, se pudermos construir um modelo de conjunto a partir dos mesmos dados de treinamento, podemos reduzir o efeito de variações aleatórias em modelos únicos.

## Treinando uma Floresta Aleatória

Uma Floresta Aleatória consiste basicamente em empregarmos várias diferentes Árvores de Decisão sobre os mesmos dados e obtermos a média dessas árvores para a predição. Na construção de um modelo de conjunto, *ensemble model*, em que todos modelos do conjunto são árvores de decisão é importante garantir que cada árvore individual não esteja muito correlacionada a qualquer uma das outras árvores no modelo. Para garantir isso a construção de florestas aleatórias baseia-se em dois princípios: a *agregação bootstrap*, ou *bagging*; e a seleção aleatória de atributos.

### Bagging, *Bootstrap Aggregation*

As árvores de decisão são muito sensíveis aos dados de treinamento com pequenas mudanças dos dados resultando em árvores significativamente diferentes. A floresta aleatória tira vantagem disso, permitindo que cada árvore individual faça uma amostra aleatória do conjunto de dados com substituição, resultando em árvores bastante diferentes. Este processo é conhecido como *bagging*. Note que não estamos subdividindo os dados de treinamento em partes menores e treinando cada árvore em uma parte diferente. O número de amostras de treinamento em cada árvore será o mesmo, mas em vez de empregarem os dados de treinamento originais, cada árvore emprega uma amostra aleatória do mesmo tamanho dos dados originais com substituição. No exemplo abaixo, para os dados originais são empregadas as amostras , e , no que se assemelha muito ao CV, embora seja diferente.

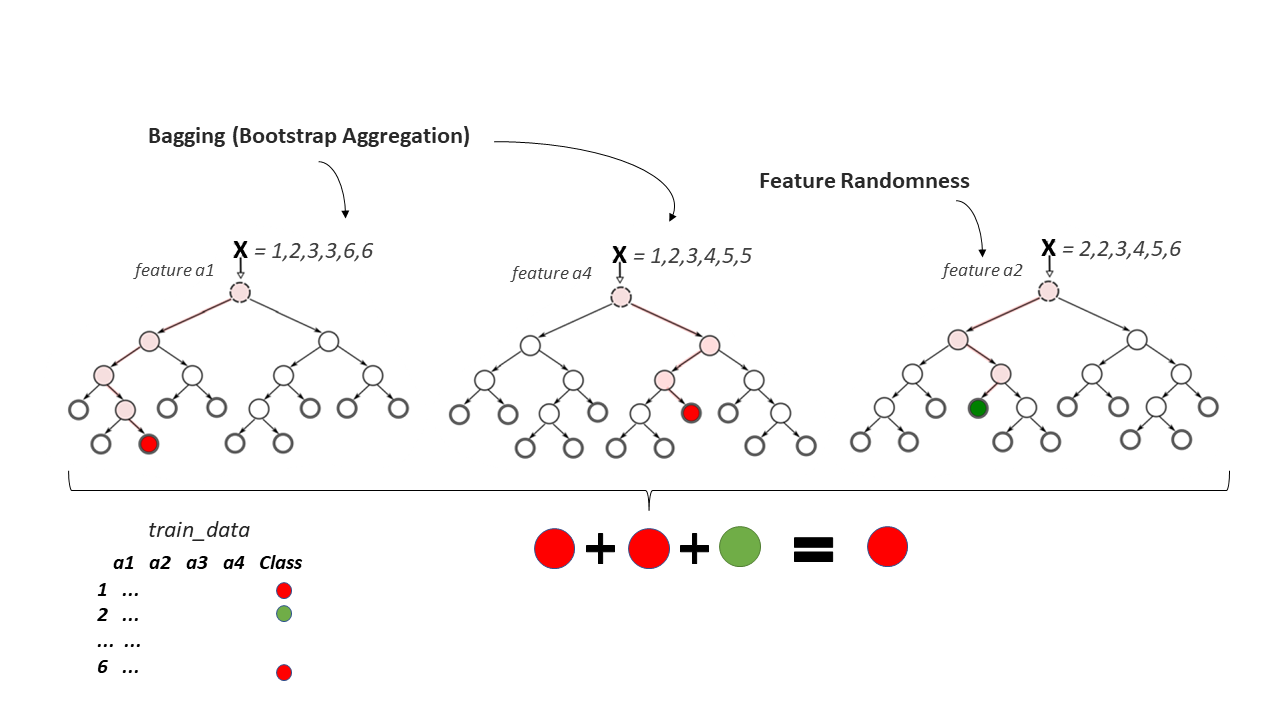


Figura 7. Esquema de Funcionamento de Random Forests.

### Seleção Aleatória de Atributos

Como você viu na construção de uma árvore de decisão normal escolhemos a cada nível os atributos que produzem a maior separação entre as amostras. Se adotássemos isso em uma Floresta de Árvores, certamente iríamos obter Árvores bastante correlacionadas. Uma Floresta Aleatória se utiliza do fato de que podemos empregar outros atributos e outra ordem dos atributos para construir a árvore, assim cada árvore em uma floresta aleatória escolhe atributos apenas de um subconjunto aleatório o que permite uma maior variação entre as árvores no modelo de conjunto.

## Predição Final

A final, a predição do modelo é obtida como uma média do resultado de cada árvore do conjunto (modelos de regressão) ou o *majority-voting* (classificação).

# Um exemplo clássico

Antes de prosseguir talvez você queira revisar aqui alguns pontos de como empregar o classificador DecisionTreeClassifier no exemplo clássico de Mitchell, Tom Michael (1997). Machine Learning.

df = pd.read\_csv('http://meusite.mackenzie.br/rogerio/ML/PlayBallcsv.csv')  
df.head()

Day Outlook Temperature Humidity Wind Play ball  
0 D1 Sunny Hot High Weak No  
1 D2 Sunny Hot High Strong No  
2 D3 Overcast Hot High Weak Yes  
3 D4 Rain Mild High Weak Yes  
4 D5 Rain Cool Normal Weak Yes

Primeiramente você precisa fazer o encode dos dados, uma restrição da implementação do scikit-learn.

from sklearn.preprocessing import OneHotEncoder, LabelEncoder  
  
df\_labels = df.drop(columns=['Day','Play ball']).apply(LabelEncoder().fit\_transform)  
  
pd.concat([ df\_labels, df[['Play ball']] ], axis=1).head()

Outlook Temperature Humidity Wind Play ball  
0 2 1 0 1 No  
1 2 1 0 0 No  
2 0 1 0 1 Yes  
3 1 2 0 1 Yes  
4 1 0 1 1 Yes

O uso do estimador segue do mesmo modo que já empregamos para o estimador logístico, o Knn e é a forma de empregarmos qualquer modelo de aprendizado supervisinado.

from sklearn.tree import DecisionTreeClassifier  
  
X = df\_labels  
y = df['Play ball']  
  
clf = DecisionTreeClassifier()   
  
clf.fit(X,y)  
y\_pred = clf.predict(X)  
print( clf.score(X,y) )

1.0

Você pode achar útil visualizar a árvore de decisão criada com o graphviz.

import graphviz   
dot\_data = tree.export\_graphviz(clf, out\_file=None,   
 feature\_names=list(df\_labels.columns.values),   
 class\_names=list(sorted(df['Play ball'].unique())),   
 filled=True, rounded=True,   
 special\_characters=False,  
 proportion=False, impurity=False, node\_ids=False,label=None)   
graph = graphviz.Source(dot\_data)   
graph.render('graph\_cartoons') # para gravação em .pdf  
graph

<graphviz.files.Source at 0x7f05ae17acd0>

E a informação mútua e o peso dos atributos no classificador fornecem informações úteis para análises dos modelos em muitos casos.

print(X.columns)  
print(mutual\_info\_classif(X, y, discrete\_features=True))  
print(clf.feature\_importances\_)

Index(['Outlook', 'Temperature', 'Humidity', 'Wind'], dtype='object')  
[0.17103394 0.02025554 0.10524435 0.03335912]  
[0.31555556 0.15555556 0.28 0.24888889]

# Selecionando Diferentes Modelos

Diferentes modelos criam fronteiras de decisão diferentes para os mesmos conjuntos de dados. Eles empregam critérios diferentes e, por isso, produzem classificações de modo diferente o que torna difícil você escolher empregar um modelo ou outro, muitas vezes, essa é a parte mais difícil de resolver um problema de aprendizado de máquina: *qual o estimador certo para empregar?* Não há um melhor estimador *apriori* para quaisquer dados e estimadores diferentes são mais adequados para diferentes tipos de dados e problemas. O scikit-learn fornece em <https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html> um diagrama, e que fornecemos abaixo, que funciona como um guia, embora aproximado, de como abordar problemas com relação a quais estimadores empregar.

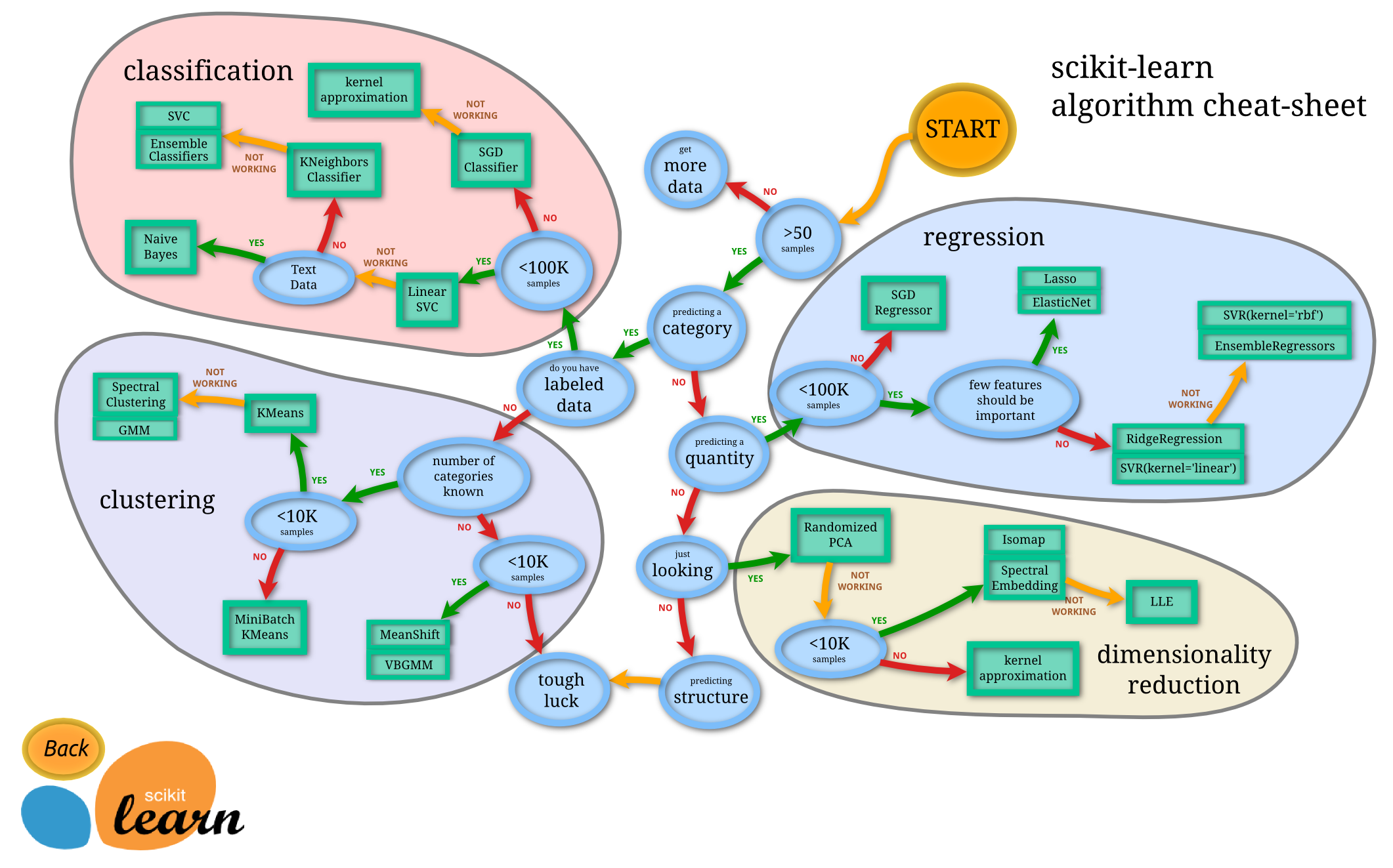


Figura 8. Diagrama do scikit-learn para seleção de modelos. (Fonte: <https://scikit-learn.org/>)

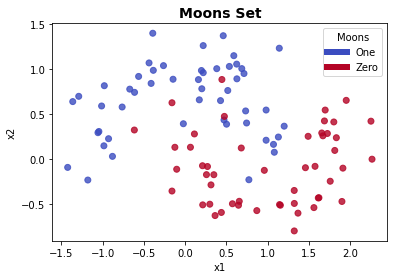
De qualquer modo uma abordagem comum e que você encontra em sistemas de *Auto ML*, consiste em adotarmos alguma métrica, como a acuracidade ou a precisão, para selecionar diferentes modelos, experimentando cada um deles, muito à exemplo do que fizemos na trilha anterior para a escolha dos melhores hiperparâmetros.

Vamos adaptar então o código modelo final da trilha anterior para explorar agora, não só diferentes hiperparâmetros, mas diferentes estimadores. Afinal você já estudou aqui ao menos 4 estimadores para classificação e está na hora de você buscar escolher o melhor modelo dentre esses classificadores para um conjunto de dados de sua escolha. Vamos adionar ainda um quinto modelo, que você ainda não estudou, para que você entenda que você já pode explorar quaisquer modelos de classificação com o que aprendeu aqui.

Abaixo os principais classificadores do scikit-learn e os que empregaremos, com (*), para seleção de um estimador para o nosso conjunto de dados* moons\*:

from sklearn.linear\_model import LogisticRegression # (\*)  
from sklearn.neighbors import KNeighborsClassifier # (\*)  
from sklearn.tree import DecisionTreeClassifier # (\*)  
from sklearn.model\_selection import KFold  
from sklearn.model\_selection import cross\_val\_score  
from sklearn.naive\_bayes import GaussianNB  
from sklearn.tree import DecisionTreeClassifier  
from sklearn.svm import SVC  
from sklearn.neural\_network import MLPClassifier  
from sklearn.ensemble import RandomForestClassifier  
from sklearn.ensemble import AdaBoostClassifier  
from sklearn.ensemble import GradientBoostingClassifier # (\*)

from sklearn.datasets import make\_moons  
cmap\_data = plt.cm.coolwarm   
  
X, y = make\_moons(n\_samples=100, noise=0.25, random\_state=1234)  
  
df = pd.DataFrame({'x1':X[:, 0], 'x2':X[:, 1], 'y':y})  
  
plt.scatter(df.x1, df.x2, c=df.y, cmap=cmap\_data, alpha=0.8)  
plt.title('Moons Set',weight='bold',fontsize=14)  
plt.xlabel("x1")  
plt.ylabel("x2")  
  
custom\_lines = [Line2D([0], [0], color=cmap\_data(0.), lw=6),  
 Line2D([0], [0], color=cmap\_data(1.), lw=6)]  
plt.legend(custom\_lines, ['One', 'Zero'], loc='upper right',title='Moons')  
  
plt.show()  
  
print(df.head())



x1 x2 y  
0 -0.763251 0.577354 0  
1 -0.391942 1.395641 0  
2 1.324561 -0.492319 1  
3 0.271295 -0.082532 1  
4 2.253887 0.420281 1

df.to\_csv('moons.csv',index=None)

def border(clf):  
 x\_min, x\_max = X['x1'].min() - .5, X['x1'].max() + .5  
 y\_min, y\_max = X['x2'].min() - .5, X['x2'].max() + .5  
 xx, yy = np.meshgrid(np.arange(x\_min, x\_max, 0.2),  
 np.arange(y\_min, y\_max, 0.2))  
  
 if hasattr(clf, "decision\_function"):  
 Z = clf.decision\_function(np.c\_[xx.ravel(), yy.ravel()])  
 else:  
 Z = clf.predict\_proba(np.c\_[xx.ravel(), yy.ravel()])[:, 1]  
  
 Z = Z.reshape(xx.shape)  
 plt.contourf(xx, yy, Z, cmap=cmap\_data, alpha=0.8)   
 plt.scatter(X['x1'], X['x2'], c=y, cmap=cmap\_data, alpha=0.8)  
 plt.title(str(clf)[0:str(clf).index('(')] + ' Moons Predictions',weight='bold',fontsize=14)  
 plt.xlabel("x1")  
 plt.ylabel("x2")  
  
 custom\_lines = [Line2D([0], [0], color=cmap\_data(0.), lw=6),  
 Line2D([0], [0], color=cmap\_data(1.), lw=6)]  
  
 plt.legend(custom\_lines, ['One', 'Zero'], loc='upper right',title='Moons')  
  
 plt.show()  
 return

Para obter os hiperparâmetros de cada modelo você pode empregar:

from sklearn.ensemble import GradientBoostingClassifier  
GradientBoostingClassifier().get\_params

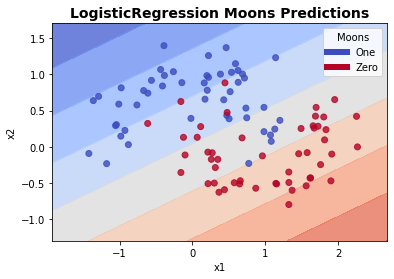
<bound method BaseEstimator.get\_params of GradientBoostingClassifier()>

Abaixo um esquema geral de como avaliar os diferentes regressores com seus diferentes parâmetros e você poderá empregar outros se quiser.

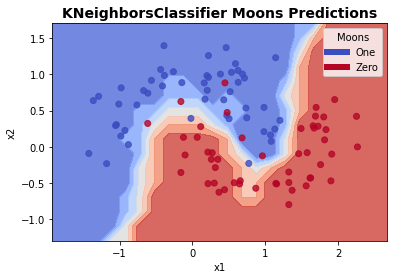
from sklearn.linear\_model import LogisticRegression  
from sklearn.neighbors import KNeighborsClassifier  
from sklearn.tree import DecisionTreeClassifier  
from sklearn.model\_selection import KFold  
from sklearn.model\_selection import cross\_val\_score  
from sklearn.naive\_bayes import GaussianNB  
from sklearn.tree import DecisionTreeClassifier  
from sklearn.svm import SVC  
from sklearn.neural\_network import MLPClassifier  
from sklearn.ensemble import RandomForestClassifier  
from sklearn.ensemble import AdaBoostClassifier  
from sklearn.ensemble import GradientBoostingClassifier  
  
from sklearn.model\_selection import train\_test\_split  
from sklearn import neighbors  
from sklearn.preprocessing import MinMaxScaler  
from sklearn.preprocessing import OneHotEncoder, LabelEncoder  
from sklearn.model\_selection import GridSearchCV  
from sklearn.metrics import classification\_report  
  
X = df[['x1','x2']]  
y = df.y  
  
X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, stratify=y, test\_size=0.3, random\_state=123)  
  
  
base\_estimators = [ LogisticRegression(),  
 neighbors.KNeighborsClassifier(),  
 DecisionTreeClassifier(),  
 RandomForestClassifier(),  
 GradientBoostingClassifier()]  
   
   
param\_grids = [ {},  
 {'n\_neighbors': [3,4,5], 'metric': ['euclidean','manhattan']},  
 {'max\_depth': [2,3,4,5]},  
 {'n\_estimators':[3,4,5,6]},   
 {'n\_estimators':[3,4,5,6]}]   
   
for i in range(len(base\_estimators)):  
 clf = GridSearchCV(base\_estimators[i], param\_grids[i], cv=5, scoring='accuracy')  
 clf.fit(X\_train, y\_train)  
 # print(clf.cv\_results\_)  
 print(clf.best\_estimator\_)  
 border(clf.best\_estimator\_)  
 print()  
 print("Detailed classification report:")  
 print()  
 y\_pred = clf.predict(X\_test)  
 print(classification\_report(y\_test, y\_pred))  
 print()

LogisticRegression()

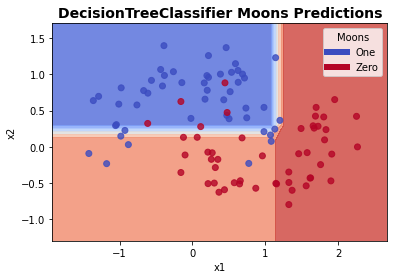
/usr/local/lib/python3.7/dist-packages/sklearn/base.py:446: UserWarning: X does not have valid feature names, but LogisticRegression was fitted with feature names  
 "X does not have valid feature names, but"



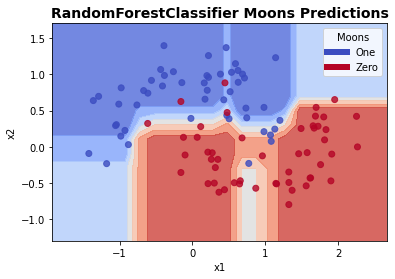
Detailed classification report:  
  
 precision recall f1-score support  
  
 0 0.93 0.93 0.93 15  
 1 0.93 0.93 0.93 15  
  
 accuracy 0.93 30  
 macro avg 0.93 0.93 0.93 30  
weighted avg 0.93 0.93 0.93 30  
  
  
KNeighborsClassifier(metric='euclidean', n\_neighbors=4)



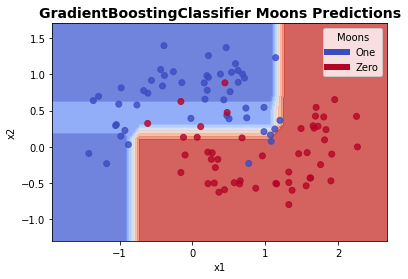
Detailed classification report:  
  
 precision recall f1-score support  
  
 0 0.83 1.00 0.91 15  
 1 1.00 0.80 0.89 15  
  
 accuracy 0.90 30  
 macro avg 0.92 0.90 0.90 30  
weighted avg 0.92 0.90 0.90 30  
  
  
DecisionTreeClassifier(max\_depth=2)



Detailed classification report:  
  
 precision recall f1-score support  
  
 0 0.92 0.80 0.86 15  
 1 0.82 0.93 0.87 15  
  
 accuracy 0.87 30  
 macro avg 0.87 0.87 0.87 30  
weighted avg 0.87 0.87 0.87 30  
  
  
RandomForestClassifier(n\_estimators=5)



Detailed classification report:  
  
 precision recall f1-score support  
  
 0 0.93 0.93 0.93 15  
 1 0.93 0.93 0.93 15  
  
 accuracy 0.93 30  
 macro avg 0.93 0.93 0.93 30  
weighted avg 0.93 0.93 0.93 30  
  
  
GradientBoostingClassifier(n\_estimators=3)



Detailed classification report:  
  
 precision recall f1-score support  
  
 0 0.93 0.93 0.93 15  
 1 0.93 0.93 0.93 15  
  
 accuracy 0.93 30  
 macro avg 0.93 0.93 0.93 30  
weighted avg 0.93 0.93 0.93 30

Analisando os dados os melhores resultados são os resultados dos modelos Gradiente Boosting e Regressão Logística e, neste caso, deveríamos empregar o modelo mais simples de Regressão Logística seguindo o princípio da parcimônia. Este entretanto é um exemplo de brinquedo, mas serve para mostrar que nem sempre modelos mais elaborados irão fornecer a melhor predição dos dados.

# CASO: Classifying Political Parties Based on Congressional Votes

Este conjunto de dados inclui votos para cada um dos congressistas da Câmara dos Representantes dos EUA nos 16 votos principais identificados pelo CQA. Nosso objetivo é construir selecionar um melhor modelo para prever a filiação partidária dos congressistas com base nos registros de votação.

Fonte: <http://archive.ics.uci.edu/ml/datasets/Congressional+Voting+Records>

df = pd.read\_csv('http://archive.ics.uci.edu/ml/machine-learning-databases/voting-records/house-votes-84.data',header=None,prefix='Vote\_')  
df.rename(columns={'Vote\_0':'class'},inplace=True)  
df.head()

class Vote\_1 Vote\_2 Vote\_3 ... Vote\_13 Vote\_14 Vote\_15 Vote\_16  
0 republican n y n ... y y n y  
1 republican n y n ... y y n ?  
2 democrat ? y y ... y y n n  
3 democrat n y y ... y n n y  
4 democrat y y y ... y y y y  
  
[5 rows x 17 columns]

df.shape

(435, 17)

## Preparação dos Dados

Não existem dados ausentes, a as abstenções '?' são um grande número de casos. Vamos esperar que elas possam também ajudar a identificar o partido do congressista uma vez que isso parece ser mais uma informação importante do que exatamente ausência de informação.

df.isnull().sum().sum()

0

(df == '?').sum()

class 0  
Vote\_1 12  
Vote\_2 48  
Vote\_3 11  
Vote\_4 11  
Vote\_5 15  
Vote\_6 11  
Vote\_7 14  
Vote\_8 15  
Vote\_9 22  
Vote\_10 7  
Vote\_11 21  
Vote\_12 31  
Vote\_13 25  
Vote\_14 17  
Vote\_15 28  
Vote\_16 104  
dtype: int64

Agora precisamos fazer o encode dos dados. A rigor deveríamos fazer o hot encode, mas por simplificidade será suficiente aqui fazermos apenas o label encode do dados.

from sklearn.preprocessing import OneHotEncoder, LabelEncoder  
  
df\_labels = df.drop(columns=['class']).apply(LabelEncoder().fit\_transform)  
  
pd.concat([ df\_labels, df[['class']] ], axis=1).head()

Vote\_1 Vote\_2 Vote\_3 Vote\_4 ... Vote\_14 Vote\_15 Vote\_16 class  
0 1 2 1 2 ... 2 1 2 republican  
1 1 2 1 2 ... 2 1 0 republican  
2 0 2 2 0 ... 2 1 1 democrat  
3 1 2 2 1 ... 1 1 2 democrat  
4 2 2 2 1 ... 2 2 2 democrat  
  
[5 rows x 17 columns]

Antes de testarmos diferentes classificadores vamos aplicar um modelo básico para verificar o formato dos dados para aplicação do modelo.

from sklearn.tree import DecisionTreeClassifier  
  
X = df\_labels  
y = df['class']  
  
clf = DecisionTreeClassifier()   
  
clf.fit(X,y)  
y\_pred = clf.predict(X)  
print( clf.score(X,y) )

1.0

Parece tudo certo e podemos ainda verificar a árvore produzida neste treinamento.

import graphviz   
dot\_data = tree.export\_graphviz(clf, out\_file=None,   
 feature\_names=list(df\_labels.columns.values),   
 class\_names=list(sorted(df['class'].unique())),   
 filled=True, rounded=True,   
 special\_characters=False,  
 proportion=False, impurity=False, node\_ids=False,label=None)   
graph = graphviz.Source(dot\_data)   
graph.render('graph\_cartoons') # para gravação em .pdf  
graph

<graphviz.files.Source at 0x7f05b7c59650>

e os valores de informações mútua e importância dos atributos.

print('Mutual information and the attribute with higher Mutual Info: \n')  
print(mutual\_info\_classif(X, y, discrete\_features=True))  
print(X.columns[ mutual\_info\_classif(X, y, discrete\_features=True).argmax() ])  
print()  
print('Feature Importances and the attribute with higher Feature Importances: \n')  
print(clf.feature\_importances\_)  
print(X.columns[ clf.feature\_importances\_.argmax() ])

Mutual information and the attribute with higher Mutual Info:   
  
[8.73872293e-02 2.49962316e-04 2.99660509e-01 5.12951549e-01  
 2.92820364e-01 1.02055280e-01 1.37023508e-01 2.35826473e-01  
 2.15261646e-01 3.52248115e-03 7.43690939e-02 2.59411124e-01  
 1.57899640e-01 2.32400929e-01 1.52771138e-01 7.06865546e-02]  
Vote\_4  
  
Feature Importances and the attribute with higher Feature Importances:   
  
[9.69769930e-03 1.53546906e-02 2.81948666e-02 8.27422108e-01  
 0.00000000e+00 8.08141609e-04 0.00000000e+00 4.84884965e-03  
 2.95109584e-02 1.41486946e-02 2.65326564e-02 1.20136023e-02  
 8.55160757e-03 0.00000000e+00 7.38511255e-03 1.55310124e-02]  
Vote\_4

Vamos então aplicar os classificadores que aprendemos até aqui variando alguns de seus parâmetros.

from sklearn.linear\_model import LogisticRegression  
from sklearn.neighbors import KNeighborsClassifier  
from sklearn.tree import DecisionTreeClassifier  
from sklearn.ensemble import RandomForestClassifier  
  
from sklearn.model\_selection import train\_test\_split  
from sklearn import neighbors  
from sklearn.preprocessing import MinMaxScaler  
from sklearn.preprocessing import OneHotEncoder, LabelEncoder  
from sklearn.model\_selection import GridSearchCV  
from sklearn.metrics import classification\_report  
  
X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, stratify=y, test\_size=0.3, random\_state=123)  
  
base\_estimators = [ LogisticRegression(),  
 neighbors.KNeighborsClassifier(),  
 DecisionTreeClassifier(),  
 RandomForestClassifier() ]  
   
param\_grids = [ {},  
 {'n\_neighbors': [3,4,5,6,7]},  
 {'max\_depth': [2,3,4,5,6,7,8,9,10], 'criterion': ['gini','entropy']},  
 {'n\_estimators':[3,4,5,6]}]   
   
save\_estimators = []  
  
for i in range(len(base\_estimators)):  
 clf = GridSearchCV(base\_estimators[i], param\_grids[i], cv=5, scoring='accuracy')  
 clf.fit(X\_train, y\_train)  
 # print(clf.cv\_results\_)  
 print(clf.best\_estimator\_)  
 save\_estimators.append(clf.best\_estimator\_)  
 print()  
 print("Detailed classification report:")  
 print()  
 y\_pred = clf.predict(X\_test)  
 print(classification\_report(y\_test, y\_pred))  
 print()

LogisticRegression()  
  
Detailed classification report:  
  
 precision recall f1-score support  
  
 democrat 0.93 0.97 0.95 80  
 republican 0.96 0.88 0.92 51  
  
 accuracy 0.94 131  
 macro avg 0.94 0.93 0.93 131  
weighted avg 0.94 0.94 0.94 131  
  
  
KNeighborsClassifier(n\_neighbors=4)  
  
Detailed classification report:  
  
 precision recall f1-score support  
  
 democrat 0.94 0.94 0.94 80  
 republican 0.90 0.90 0.90 51  
  
 accuracy 0.92 131  
 macro avg 0.92 0.92 0.92 131  
weighted avg 0.92 0.92 0.92 131  
  
  
DecisionTreeClassifier(max\_depth=2)  
  
Detailed classification report:  
  
 precision recall f1-score support  
  
 democrat 0.95 0.96 0.96 80  
 republican 0.94 0.92 0.93 51  
  
 accuracy 0.95 131  
 macro avg 0.95 0.94 0.94 131  
weighted avg 0.95 0.95 0.95 131  
  
  
RandomForestClassifier(n\_estimators=5)  
  
Detailed classification report:  
  
 precision recall f1-score support  
  
 democrat 0.92 0.97 0.95 80  
 republican 0.96 0.86 0.91 51  
  
 accuracy 0.93 131  
 macro avg 0.94 0.92 0.93 131  
weighted avg 0.93 0.93 0.93 131

Com os parâmetros acima o melhor estimador foi a Árvore de Decisão com o critério gini de seleção dos atributos e com uma profundidade máxima de 2 da árvore.

[x.score(X\_test, y\_test) for x in save\_estimators]

[0.9389312977099237,  
 0.9236641221374046,  
 0.9465648854961832,  
 0.9312977099236641]

best = np.array( [x.score(X\_test, y\_test) for x in save\_estimators] ).argmax()  
save\_estimators[best]

DecisionTreeClassifier(max\_depth=2)

clf = save\_estimators[best]  
  
import graphviz   
dot\_data = tree.export\_graphviz(clf, out\_file=None,   
 feature\_names=list(df\_labels.columns.values),   
 class\_names=list(sorted(df['class'].unique())),   
 filled=True, rounded=True,   
 special\_characters=False,  
 proportion=False, impurity=False, node\_ids=False,label=None)   
graph = graphviz.Source(dot\_data)   
graph.render('graph\_cartoons') # para gravação em .pdf  
graph

<graphviz.files.Source at 0x7f05ad55fb50>

Você pode acessar a documentação dos dados (<http://archive.ics.uci.edu/ml/machine-learning-databases/voting-records/house-votes-84.names>) e verificar a que eleições se referem os votos acima:

* Vote\_4: physician-fee-freeze
* Vote\_3: adoption-of-the-budget-resolution
* Vote\_11: synfuels-corporation-cutback

E, de fato, seria suficiente ainda somente a informação do Vote\_4: physician-fee-freeze para classificar os congressistas. Em 1984, o governo federal impõe um congelamento de um ano nos pagamentos para os serviços de saúde, sendo os serviços de saúde um tema polêmico que ainda hoje divide republicamos e democratas.

# Síntese

Nesta trilha aprendeu sobre o modelo de Árvores de Decisão. Elas são modelos bastante empregados e, além disso, permitiram introduzir uma série de conceitos importantes como entropia e ganho de informação que têm um papel importante, sobretudo, na seleção de atributos para modelos de aprendizado.

Árvores de decisão podem ser empregadas tanto para modelos de classificação, como empregamos aqui, como também para criar estimadores de regressão não lineares. Elas são eficientes preditores mas árvores muito profundas tendem a criar modelos com sobreajuste. Para reduzir isso podemos fazer uma *poda* da árvore, limitando a sua profundidade (parâmetro que empregamos aqui na seleção de melhores parâmetros do estimador). Outro método bastante empregado em aprendizado de máquina é o uso de *ensemble models* e, aqui, você aprendeu sobre sobre um dos mais importantes modelos conjuntos. As florestas aleatórias são modelos conjuntos que empregam vários estimadores de árvore não correlacionados e obtêm a predição através da média dos valores ou do *majority-voting* das classes obtidos dos diferentes estimadores.

# Para Saber Mais

* Acesso [https://ml-playground.com/#](https://ml-playground.com/) e experimente alguns modelos de classificação sem qualquer programação. Se estiver interessado ainda pode verificar como exportar e importar dados do [https://ml-playground.com/#](https://ml-playground.com/) para o Python. Dica, os dados são salvos e carregados em formato .json.
* O exemplo clássico Árvores de Decisão de Mitchell, Tom Michael (1997) em **Machine Learning** é também explorado em Sayad, Saed (2021). **An Introduction to Data Science**, Disponível em: <https://www.saedsayad.com/data_mining_map.htm> e ajuda a entender em um passo a passo como é construída uma árvore de decisão e são calculados o ganho de informação.
* Acesse Jake VanderPlas. **Python Data Science Handbook** Disponível em: <https://jakevdp.github.io/PythonDataScienceHandbook/> e veja o capítulo **In-Depth: Decision Trees and Random Forests**, em especial o exemplo: **Example: Random Forest for Classifying Digits** classificam dados de imagens que não exploramos aqui.
* Explore diferentes modelos de classificação com o scikit-learn. Você pode começar por aqui:
* **Choosing the right estimator** <https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html>, ou
* **Classifier comparison** <https://scikit-learn.org/stable/auto_examples/classification/plot_classifier_comparison.html>

crie *projetos* próprios, seus, empregando os mesmos dados que empregamos aqui mas com outros modelos, ou empregue outros dados de interesse. Procure explorar modelos mais elaborados como SVC (Support Vector Machines) e AdaBoostClassifier, e empregar dados com mais classes de saída e maior volume de casos. Você também pode empregar modelos Árvore ou mesmo de K-vizinhos mais próximos para regressões não lineares. Os modelos código e o que você aprendeu até aqui são bastante suficientes para você avançar bastante rapidamente nesses projetos.

# Referências

Weinberger, Kilian. **Machine Learning for Intelligent Systems**, Lecture 28: Decision / Regression Trees Disponível em: <http://www.cs.cornell.edu/courses/cs4780/2018fa/page18/> Acesso: 14 de Novembro de 2021.

Mitchell, Tom Michael (1997). **Machine Learning**. New York, NY: McGraw-Hill. ISBN 0070428077.

Subasi, A. (2020). **Machine learning techniques. Practical Machine Learning for Data Analysis Using Python**, 91–202. <doi:10.1016/b978-0-12-821379-7.00003-5>

Machado G, Mendoza MR, Corbellini LG. (2015) **What variables are important in predicting bovine viral diarrhea virus? A random forest approach.** Veterinary Research. Jul 24;46(1):85. doi: 10.1186/s13567-015-0219-7.

Larose, Chantal D.; Larose, Daniel T. **Data Science Using Python and R** Hoboken: Wiley, c2019. E-book (259 p.) (Wiley Series on Methods and Applications in Data Mining Ser.). ISBN 9781119526834 (electronic bk.). Disponível em: <https://www3.mackenzie.br/biblioteca_virtual/index.php?tipoBiblio=ebookcentral&flashObg=n>

Kotu, Vijay; Deshpande, Balachandre **Data Science: concepts and practice**. 2nd ed. Cambridge, [England]: Morgan Kaufmann, c2019. E-book (570 p.) ISBN 9780128147627 (electronic bk.). Disponível em: <http://pergamum.mackenzie.br:8080/pergamumweb/vinculos/00003c/00003cef.jpg>.

Jake VanderPlas. **Python Data Science Handbook** O'Reilly Media, Inc. (2016). ISBN: 9781491912058. Disponível em: <https://jakevdp.github.io/PythonDataScienceHandbook/>. Acesso: 06 de Novembro de 2021.