Aprendizado de Máquina

Trabalho de Inteligência Artificial (INF1771)

Guilherme Dantas & Nagib Suaid

Base de dados: ‘Flags’

O objetivo é predizer a **religião** de um país a partir dos seguintes atributos:

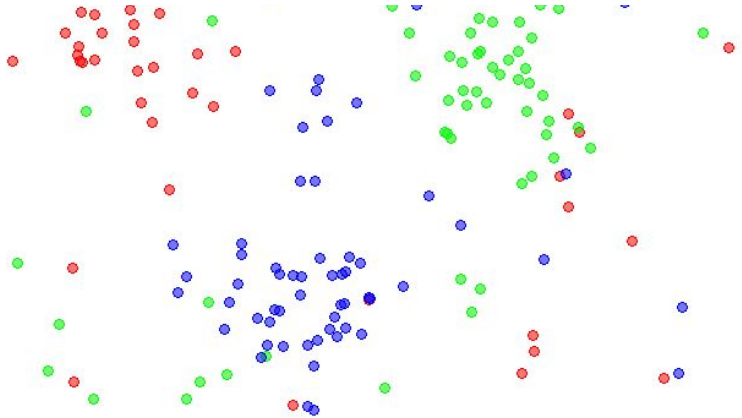
- Bandeira (23 atributos)
- Língua oficial
- Continente
- População
- Área
- Quadrante

Algoritmos implementados

K-Nearest Neighbor (KNN)

Com dois cálculos de distância:

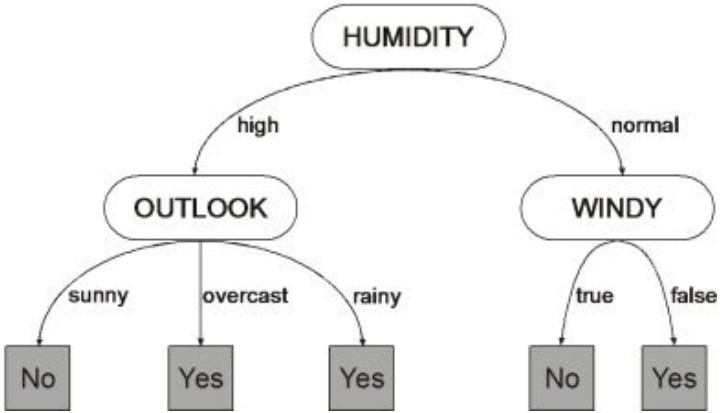
- Distância euclidiana
- Distância de Hamming



Árvores de decisão

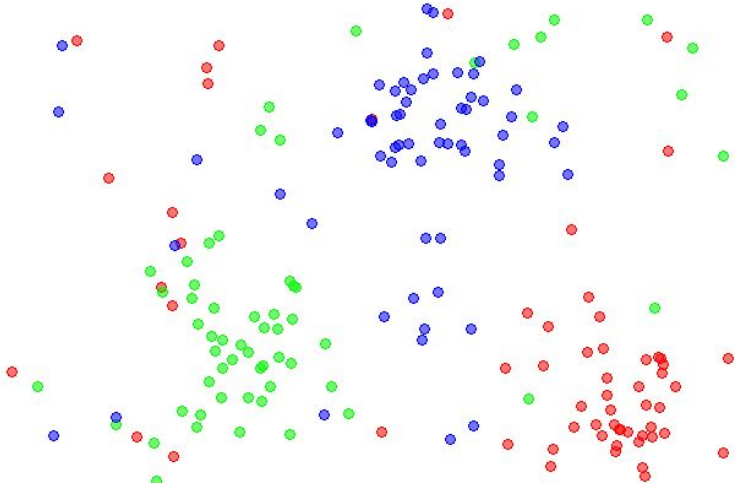
Com duas métricas de impureza:

- Impureza de Gini
- Entropia



K-Nearest Neighbor (KNN)

O rótulo de uma instância desconhecida é dada pelo rótulo mais comum entre as outras K instâncias mais “próximas” da base de treinamento, adotando algum critério de distância entre instâncias.



Foram implementadas as distâncias:

- Euclidiana
- De Hamming

KNN - Distância Euclidiana

A distância euclidiana entre duas instâncias é dada pela seguinte equação:

$$D_{euc}(x, y) = \sum_{a \in A} d(a_x, a_y)$$

E a distância entre dois atributos depende se é categórico ou contínuo:

$$d(a_x, a_y) = \begin{cases} 1, \text{if a is cathegorical and } a_x \neq a_y \\ 0, \text{if a is cathegorical and } a_x = a_y \\ (a_x - a_y)^2, \text{if a is continuous} \end{cases}$$

KNN - Distância de Hamming

A distância de Hamming entre duas instâncias é dada pela seguinte equação:

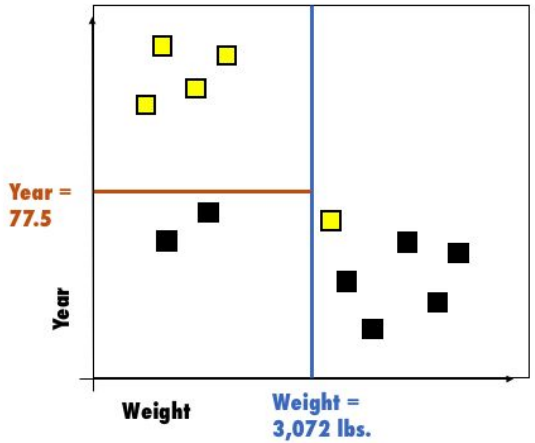
$$D_{euc}(x, y) = \sum_{a \in A} d(a_x, a_y)$$

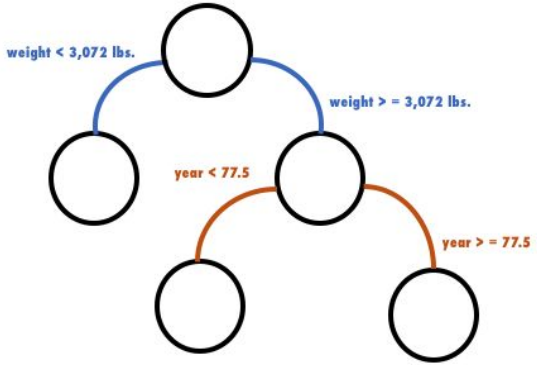
E a distância entre dois atributos depende se é categórico ou contínuo:

$$d(a_x, a_y) = \begin{cases} 1, \text{if } a_x \neq a_y \\ 0, \text{if } a_x = a_y \end{cases}$$

Árvores de Decisão

A árvore é construída subdividindo o conjunto de treinamento avaliando um atributo de cada vez, e cada subconjunto gerado a partir dessa avaliação é usado para gerar as sub-árvores filhas do nó, até que todos os elementos do conjunto de um nó tenham a mesma classificação.



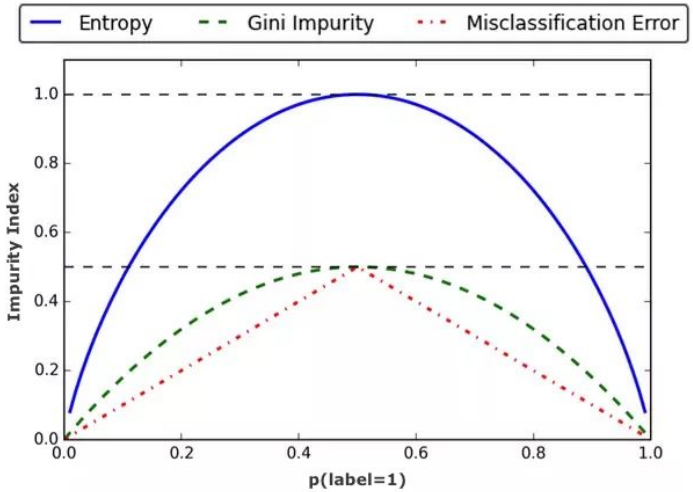


Árvores de Decisão

Para fazer a escolha de qual atributo utilizar para subdividir o conjunto de treino, avalia-se todas as divisões possíveis e seleciona-se a que gera subconjuntos com maior pureza média (ou menor impureza média)

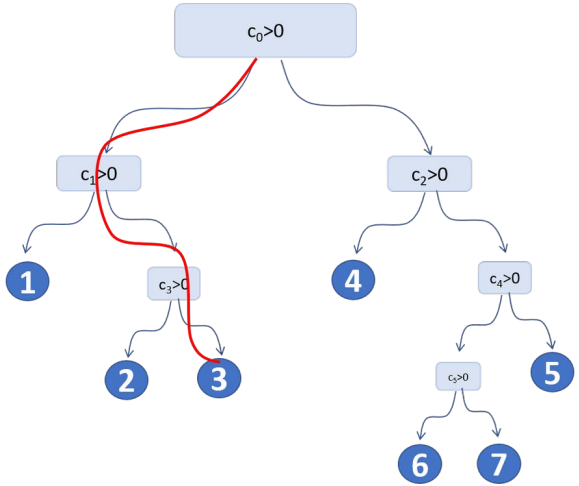
Foram implementadas as medidas de impureza:

- Impureza de Gini $1 - \sum_{i=1}^{J} {p_i}^2$
- Entropia $-\sum_{i=1}^{J} p_i \log_2 p_i$



Árvores de Decisão

Construída a árvore de decisão, para classificar uma instância desconhecida, basta percorrer a árvore a partir da raíz visitando o filho do nó corrente que corresponde ao conjunto para o qual a nova instância iria baseado no critério de divisão do nó



Seleção de atributos

Foi implementado um algoritmo que seleciona os atributos de forma a maximizar a acurácia de um dado algoritmo qualquer. Ele funciona da seguinte forma:

1. É medida a acurácia usando todos os atributos
2. É medida a acurácia sem um atributo que estava sendo usado
3. Realiza o passo 2 para todos os atributos usados até o instante e seleciona o atributo que, ao ser ignorado, melhora mais significativamente a acurácia.
4. Caso haja tal atributo, é realizado o passo 2 agora sem levar em conta o atributo selecionado. Caso contrário, o algoritmo cessa.

Resultados finais - KNN

	Distância de Hamming	**Distância euclidiana**
Atributos	Todos exceto “name”, “religion”, “colours”, “mainhue”, “saltires”, “icon” e “text”.	Todos exceto “name”, “population”, “religion”, “red”, “crosses”, “sunstars”, “crescent”, “animate” e “text”
Tempo gasto no processo de treinamento (em média, por instância)	7.1 ns (apenas armazena dados de treinamento)	
Tempo gasto no processo de classificação (em média, por exemplo desconhecido)	30.8 μs	1 ms
Taxa de reconhecimento	68,0%	66,0%

Resultados finais - Árvores de Decisão

	Índice de Gini	**Entropia**
Atributos	Todos exceto “name”, “religion”,'area', 'population','blue'	Todos exceto “name”, “religion”, 'zone','area', 'population','stripes' 'mainhue','topleft'
Tempo gasto no processo de treinamento (em média, por instância)	5.17ms	4.26ms
Tempo gasto no processo de classificação (em média, por exemplo desconhecido)	0.94ms	0.67ms
Taxa de reconhecimento	53.61%	61.85%