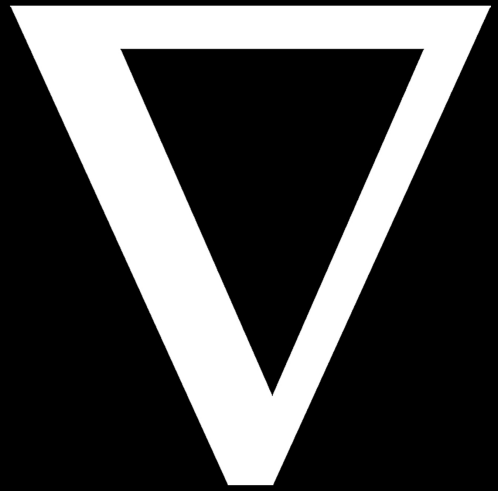


k-Nearest Neighbors (kNN)

Gustavo Sutter
@suttergustavo

k-Nearest Neighbors (ou k-Vizinhos mais próximos)

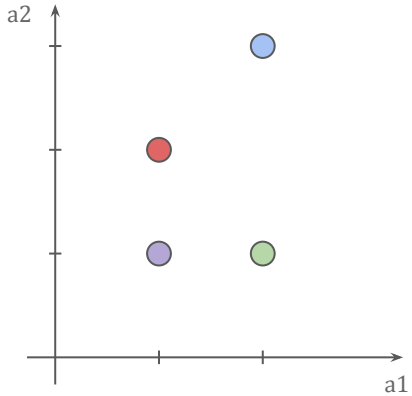
- Chamado geralmente de kNN
- É um algoritmo de classificação (mas com adaptações pode ser utilizado para regressão)
- Seus dados devem ser todos numéricos (deve ser realizada uma transformação de dados categóricos para dados numéricos)
- Parte do seguinte pressuposto: **exemplos de mesma classe estão localizados próximos no espaço**

Exemplos de mesma classe estão localizados próximos no espaço

Localização no espaço

- Como uma instância se localiza no espaço?
  - Cada instância é interpretada como um vetor que indica suas coordenadas

a1	a2
1	2
1	1
2	1
2	3



Vídeo com explicação mais detalhada dessa interpretação (link)

Localização no espaço

- Evidentemente esse conceito é generalizado para espaços com qualquer número de dimensões
- Esse tipo de interpretação é essencial para a compreensão de diversos outros conceitos de aprendizado de máquina

Exemplos de mesma classe estão localizados próximos no espaço

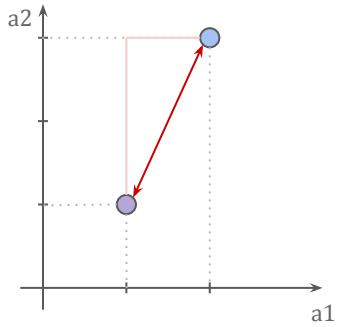
Proximidade entre instâncias

- A proximidade é calculada através de uma medida de distância entre as instâncias no espaço
- Dois tipos de medidas de distância são as mais comuns:
  - Distância Euclidiana
  - Distância Manhattan

Medidas de distância

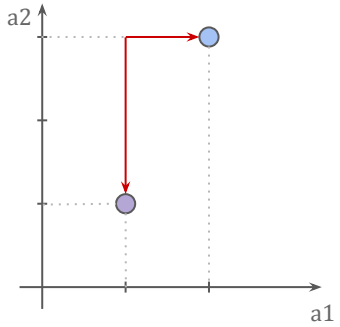
Distância Euclidiana

$$d_E(\vec{x}, \vec{y}) = \sqrt{\sum_i (\vec{x}_i - \vec{y}_i)^2}$$



Distância Manhattan

$$d_M(\vec{x}, \vec{y}) = \sum_i |\vec{x}_i - \vec{y}_i|$$



Algoritmo: Nearest neighbor ou 1NN (simplificação)

1. Calcular a distância do exemplo de teste para cada instância do conjunto de treino
2. Descobrir qual elemento do treino está mais próximo da instância de teste
3. Retornar a classe que o exemplo de treino mais próximo pertence

Algoritmo: k-Nearest neighbors

1. Calcular a distância do exemplo de teste para cada instância do conjunto de treino
2. Descobrir quais k elementos do treino estão mais próximos da instância de teste
3. Dentre os k exemplos mais próximos descobrir a classe mais frequente
4. Retornar a classe mais frequente entre os vizinhos mais próximos

Escolha do k

- O número de vizinhos é o que chamamos de **hiperparâmetro (ou *hyperparameter*)**, isto é, um parâmetro que é escolhido antes do treino
- Para a maioria das tarefas não existem hiperparâmetros mágicos que sempre funcionam. É necessário **testar vários valores e descobrir qual funciona melhor**
- Esse processo de teste de hiperparâmetros é realizado utilizando o conjunto de validação

Complexidades de treino e teste

- Idealmente queremos que o teste seja o mais rápido possível e aceitamos que o treino demore mais
- KNN é o inverso disso
- Treino é instantâneo e teste precisa olhar todos os exemplos
- Muitas vezes não roda por falta de eficiência