



INF-317B - Tópicos em Inteligência Artificial: Machine Learning

Professores: André Kazuo Takahata e Ricardo Suyama

Quadrimestre: 2020.1

Data de entrega: 12/03/2020

Projeto 01

Elaborar um código computacional de um classificador com o método K -vizinhos mais próximos (K -nearest neighbors, KNN) e realizar testes conforme os itens abaixo:

- a) Realizar testes com os dados fornecidos nos arquivos dataset1.txt a dataset6.txt. Nos arquivos, a i -ésima linha representa (x_1^i, x_2^i, θ^i) , em que x_1^i e x_2^i são as características do i -ésimo dado e θ^i a respectiva classe. A Figura 1 mostra os vetores de características do arquivo dataset1.txt no espaço de características¹.

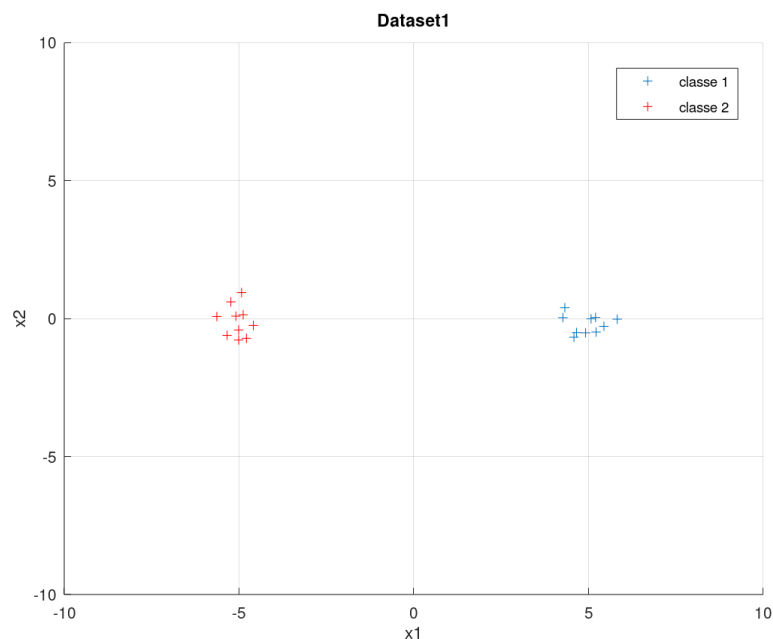


Figura 1: distribuição dos dados do Dataset1 no espaço de características.

¹ O arquivo `read_data.mat` gera o gráfico da Figura 1.

- b) Realizar testes em conjuntos de dados bidimensionais mais complexos como os mostrados na Figura 2.

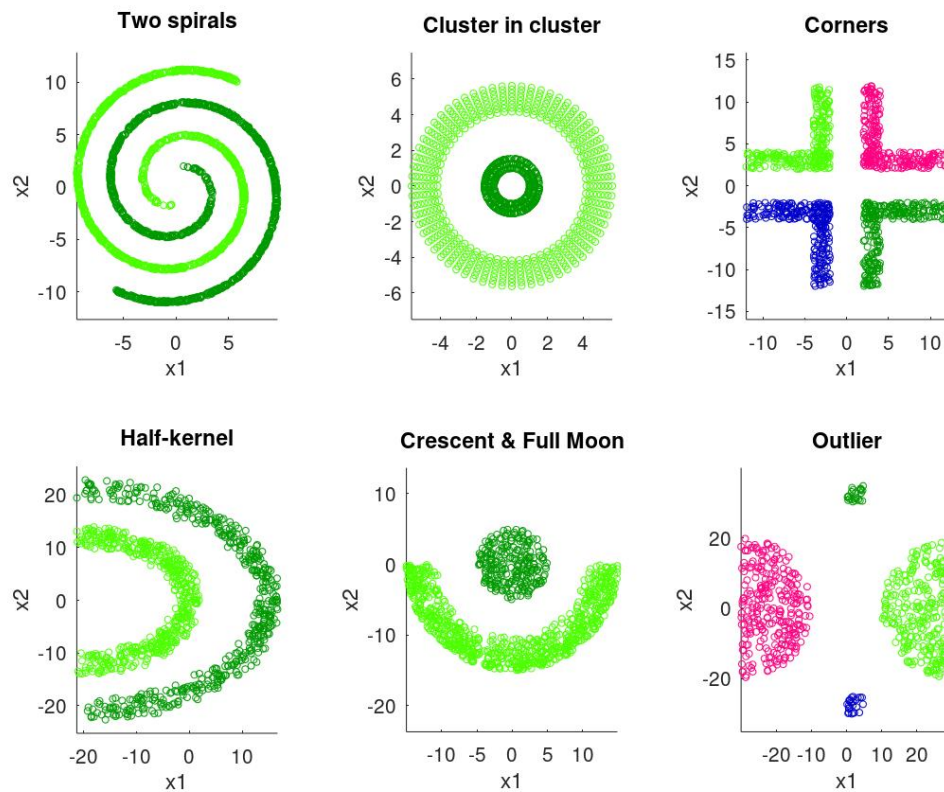


Figura 2: Conjuntos de dados bidimensionais artificiais

Os arquivos para se gerar os dados podem ser encontrados em <https://www.mathworks.com/matlabcentral/fileexchange/41459-6-functions-for-generating-artificial-datasets>. Outros conjuntos de dados bidimensionais equivalentes também podem ser utilizados.

- c) Testar o algoritmo implementado no conjunto de dados “*Olivetti Faces*” disponível em <https://cs.nyu.edu/~roweis/data.html>.

Opcionalmente é proposto a implementação de um código próprio de ordenação para a realização da localização dos K vizinhos mais próximos. Pode ser realizado uma comparação no tempo computacional entre essa opção e o uso de função de ordenação do pacote computacional utilizado na realização do Projeto. Nesse caso o método deve ser descrito na Seção de Materiais e Métodos e os resultados na Seção de Resultados.

Relatório

O relatório deverá ser de no máximo 12 páginas (sem contar capa e apêndices), com uso de fonte tamanho 12pt e espaçamento de 1,5 linhas e conter:

- Capa
- Resumo
- Introdução
- Materiais e métodos: devem ser descritos os conjuntos de dados utilizados, o algoritmo K -NN e o(s) método(s) de validação cruzada utilizado(s). No caso de dados bidimensionais devem ser mostrados gráficos de dispersão de modo a descrever os vetores de características no espaço de características.
- Resultados: mostrar matrizes de confusão e fazer análises a respeito da acurácia, precisão e revocação.
- Conclusão
- Referências bibliográficas
- Apêndices: anexar o código fonte utilizado