Geometria Computacional 2020.2

Prof. Anselmo Antunes Montenegro

Primeiro Trabalho

Aplicação de conceitos sobre Fecho Convexo 2D ao problema de detecção de *outliers*

O problema da existência de dados, cujo comportamento destoa dos demais significativamente, ocorre em muitas áreas da ciência. Dados com esta característica são denominados *outliers* e tem sido alvo de estudo por muitos anos.

Uma das formas mais comuns de se detectar *outliers* é através de técnicas estatísticas como, por exemplo, a baseada no determinante de covariância mínimo [1].

Uma forma alternativa é utilizar técnicas baseadas em distâncias. Dentro desta visão, uma possível abordagem é considerar o uso de fechos convexos como meio de detectar dados extremos do conjunto de dados. A hipótese de métodos baseados em distâncias é a de que as camadas mais externas obtidas em uma sequência de fechos convexos têm maior chance de incluir *outliers* que as mais internas.

Com base no artigo [2], que propõe uma variação do método de *Onion Peeling* [3] para detecção de *outliers*, implemente uma solução para detecção dos *k*-principais *outliers*, em dados bidimensionais, provenientes de uma distribuição Gaussiana. O método implementado deve receber como entrada um conjunto *S* de pontos 2d, um valor *k* que indica o número de *outliers* a serem detectados e deve produzir como saída uma lista *L* com os *k outliers* mais significativos.

O trabalho será avaliado conforme os seguintes aspectos:

1) Implementação do método e aplicação sobre uma distribuição Gaussiana 2D sintética (7.0 pontos).

2) Aplicação sobre um conjunto de dados 2d real. Sugestão para dados reais de teste: os dois primeiros atributos (a segunda e terceira colunas) do conjunto de dados em https://archive.ics.uci.edu/ml/datasets/Wine (1.5 pontos).

3) Comparação com os métodos MCD (Minimum Covariance Determinant) e OCMSVD (One-class SVM) disponíveis na biblioteca Sklearn [3]. Especifique pontos fortes e as limitações da abordagem implementada em comparação com os métodos MCD e OCMSVD, caso existam. (1.5 pontos)

4) Bonus: modificação ou extensão do método (2.0 pontos)

Obs.: a visualização dos dados e dos *outliers* ajuda na análise do resultado.

Referências:

- [1] Hubert, M., Debruyne, M., and Rousseeuw, P. J., “Minimum Covariance Determinant and Extensions”, arXiv e-prints, 2017. https://arxiv.org/abs/1709.07045

- [2] Harsh, A., Ball, J. E., and Wei, P., “Onion-Peeling Outlier Detection in 2-D data Sets”, arXiv e-prints, 2018. https://arxiv.org/abs/1803.04964

- [3] O'Rourke, J. Computational Geometry in C. Cambridge University Press, 1994.

- [3] https://scikit-learn.org/stable/modules/outlier_detection.html#outlier-detection

Data de entrega: 30/10/2020.