Sexta avaliação de GBC063 - AARE Etapa I / 2020

Trabalho de utilização da biblioteca Scikit-learn para agrupar empresas de chinesas (Taiwan) com base em suas informações financeiras

Este trabalho consiste em implementar na linguagem Python (v.3) um código que usa as funções e métodos apropriados da biblioteca Scikit-Learn para agrupar empresas de Taiwan com base em 95 variáveis financeiras.

Os dados são reais e já foram usados para testar a influência desses dados financeiros na acurácia da previsão de falência das empresas.

Entregas

Os trabalhos podem ser feitos em grupos de, no máximo 3 alunos, e as entregas consistem de:

- os códigos comentados referentes aos algoritmos.
- um relatório dos experimentos (conforme especificado abaixo).

As entregas devem ser realizadas via plataforma MS Teams. Em caso de impossibilidade de submissão através da plataforma MS Teams, as entregas poderão ser enviadas por email.

As entregas serão aceitas até às 13:00 do ***dia 30/09.***

Entregas fora desse prazo, serão consideradas com um valor 10% (por dia) menor.

Algoritmos referentes aos algoritmos de Clusterings e classificação .

Conforme demonstrado em sala de aula, a biblioteca scikit-learn contém as implementações dos algoritmos para criação de Clusterings, Classificadores baseados em casos e funções de Kernel.

Espera-se que os alunos implementem modelos e realizem testes, de modo que apresentem o melhor agrupamento possível utilizando-se o método K-means no dataset [https://scikit-learn.org]

Dataset e problema

O dataset selecionado foi obtido a partir de dados disponíveis no *Taiwan Economic Journal* do ano de 1999 a 2009 e publicado no E*uropean Journal of Operational Research* [https://isslab.csie.ncu.edu.tw/download/publications/1.pdf]

O código em Python (v.3) referentes ao problema listado abaixo deve ser implementados e, depois de testado, entregue em formato texto (com tabulação/indentação pronta para ser testado).

1. Encontrar agrupamentos para o dataset (desconsiderando a classe 'Bankrupt') usando o algoritmo K-means para os valores de K = {1, 2, 3, …., 30}

2. Fazer uma redução de dimensionalidade (para as duas principais componentes) para o dataset e mostrar os pontos coloridos por cluster em gráficos no plano bidimensional (PC1 x PC2)

3. Utilizar o método do *elbow* (cotovelo) para estimar qual é o 'K' que melhor representa esse dataset*

* **O método do cotovelo é um dos mais básicos métodos para se escolher uma quantidade de clusters na dispersão das instâncias dos clusters e da distância entre os outros clusters [https://en.wikipedia.org/wiki/Determining_the_number_of_clusters_in_a_data_set]**.

Relatório

Um documento contendo introdução, resultados e conclusão deve ser entregue em formato pdf.

Caso a estratégia usada para se obter a melhor quantidade de clusters não estiver explícita no código, ela deve ser explicada no relatório.

Na seção Resultados deve ser apresentado o gráfico da dispersão utilizado para a obtenção k ótimo.