Relatório do Trabalho de Implementação, da matéria Otimização Combinatória (INF 2912)

Autores: Luis Marcelo Fonseca [luismarcelovaf@gmail.com](mailto:luismarcelovaf@gmail.com), Adriel Garcia Hernandez [adriel.gh871217@gmail.com](mailto:adriel.gh871217@gmail.com), Gabriel Pujol [gabrielpujol87@gmail.com](mailto:gabrielpujol87@gmail.com), Daniel Gribel [gribel.daniel@gmail.com](mailto:gribel.daniel@gmail.com)

Descrição do problema original:

Considere objetos descritos por *c* características binárias. Estes objetos podem vir de *g* grupos diferentes. Um grupo é descrito pelas probabilidades de ter cada uma das c características. Uma possível deﬁnição de um grupo é que este tenha *cy* características com probabilidade p (p > 0.5) de ocorrer, *cn* com probabilidade 1−p e as demais com ocorrência indiferente (p = 0.5). Sugere-se que os conjuntos das *cy* características de cada grupo formem conjuntos com interseção par a par vazia.

Dado que *n* objetos são gerados conforme descrito acima, propor um classiﬁcador que indique o grupo de cada objeto.

A abordagem deve considerar entre *nmin* e *nmax* objetos gerados a partir de cada um dos *g* grupos, determinando o conjunto para aprendizado. Em seguida, gera-se e clasiﬁca-se *n* objetos gerados sorteando com probabilidades iguais o grupo de origem e as características do objeto de acordo com o seu grupo.

Propõe-se executar os testes assumindo o conhecimento de *g* e sem conhecer *g*, quando um algoritmo para propor *g* deve ser testado. Um conjunto de instâncias será disponibilizado.

Methods.

The work consists on the implementation and testing of the following four methods:

1. K-Means
2. ULP – Uncapacitated Facility Location Problem
3. P-Center – Centroids Localization Problem
4. Propose an alternative method

Now following, we present all the implemented algorithms.

K-Means is a clustering method that aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.

The algorithm consists of an initialization and two main steps. We perform the initialization by randomly choosing k data points as the “means”, or centroids, and assigning to them the closest points of the dataset. Then, we execute the two main steps (Update and Assignment) until there are no changes in the assignment of clusters (our convergence criterion). The Update step consists of calculating the new means to be the centroids of the observation points in the new clusters. The Assignment step, assigns each observation to the cluster whose centroid has the minimum Euclidean distance to it: intuitively the "nearest" centroid.

ULP – Uncapacitated Facility Location Problem.

This problem is about selecting a subset of the data points known as “*facilities*”, such as this selection minimizes the *distances* to another set of data points known as “*clients*”, and taking into account that each facility has a *cost* for being selected. The problem assumes that each facility can produce and ship unlimited quantities of the commodity under consideration, and that we have no capacity restrictions while transporting and storing the commodity in the clients.

In our case study, we consider the facilities and the clients as the same set of data points. Since we do not take into account the cost of opening a facility, there is a term missing from the original formulation’s objective function, and that term was used to establish the amount of facilities that would be open. For this reason, our problem needs to receive the amount of groups as part of the input, and we had to add a restriction to force the amount of facilities to be equal to the amount of groups we need.

We solved this model using the GUROBI solver.

P-Center - Problema de Localização de Centroides: Consiste em resolver o P-Center determinar os objetos representantes de cada grupo e classiﬁcar cada objeto como sendo do grupo com representante mais próximo

K-Medians: é uma variante do K-means onde se calcula a mediana em vez da média para cada grupo para determinar a sua centróide. Isto tem o efeito de minimizar o erro sobre todos os grupos respeito à distância L1, ao contrário do que K-means faz com o quadrado da distância L2.

K-Medians: is a variation of k-means clustering where instead of calculating the mean for each cluster to determine its centroid, one instead calculates the median. This has the effect of minimizing error over all clusters with respect to the 1-norm distance metric, as opposed to the square of the 2-norm distance metric (which k-means does.)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Instância/CPU | K-Means | UFL | P-center | K-Medians |
| Inst1 |  |  |  |  |
| Inst2 |  |  |  |  |
| ... |  |  |  |  |