[EN] Introduction

1. Clustering

"Customer Segmentation is the subdivision of a market into discrete customer groups that share similar characteristics. Customer Segmentation can be a powerful means to identify unsatisfied customer needs. Using the above data, companies can then outperform the competition by developing uniquely appealing products and services.

You are owing a supermarket mall and through membership cards, you have some basic data about your customers like Customer ID, age, gender, annual income and spending score. You want to understand the customers, such as who are the target customers, so that some guidelines can be given to marketing team and plan the strategy accordingly."

According to the previous dataset description, please perform your clustering analysis and draw your conclusions regarding the main customer segments. For this case, you should be using at least the following algorithms: 1) K-Means; 2) DBSCAN; 3) Agglomerative Clustering. Use the file named "*segmentation data.csv*" to perform your analysis. The conclusions that might be taken should be according to following dimensions:

- Number of customers profiles;
- Main characteristics of those customers;
- Minority and majority customer profiles;
- Are there clusters well separated or is there any fuzzy distinction between customers?

2. Collaborative Filtering:

As previously explained, Collaborative Filtering is one way of analysing explicit data from users and items so one can perform any relevant recommendations to existing and new customers. For this case, a data set containing rated jokes from close to 60,000 users is given. By using Collaborative Filtering, either item-based CF or Matrix Factorization, pick 3 users and suggest the top 5 jokes per user. Please make sure that the users have different tastes for jokes and justify why those tastes are different as a proof of your successful implementation. Additionally, describe your approach for joke recommendation and justify why it should be the considered the best option.

Dataset Content

The dataset contains over 1.7 million continuous ratings (-10.00 to +10.00) of 150 jokes from 59,132 users. The dataset is collected between November 2006 - May 2009. The complete dataset has two CSV files:

- ratings.csv: Each row is formatted as [User ID] [Item ID] [Rating]
- items.csv: Maps item ID's to jokes

The ratings are real values ranging from -10.00 to +10.00.

As of May 2009, the jokes {7, 8, 13, 15, 16, 17, 18, 19} are the "gauge set".

[PT] Introdução

1. Clustering

"Segmentação de Clientes pode ser descrita como a subdivisão do mercado em grupos de clientes discretos que partilham características semelhantes. Segmentação de Clientes pode ser uma abordagem bastante poderosa para identificar clientes menos satisfeitos. Usando o tipo de dados apresentados, empresas podem explorar vantagens competitivas ao desenvolver produtos e serviços únicos para esses grupos de clientes.

Imaginado que fazem a gestão de um centro comercial e através do uso de cartões de clientes, existem alguns dados compostos por ID do Cliente, idade, género, salário anual e pontuação de gastos no centro comercial. O objetivo é o de perceber melhor os clientes, principalmente que tipo de clientes existem para melhor informar a equipa de marketing e planear uma estratégia para abordar esses clientes."

Tendo em consideração a descrição anterior sobre o dataset a utilizar, faça uma análise de clustering e retire algumas conclusões sobre a segmentação de clientes. Para isso, devem ser utilizados pelo menos os seguintes algoritmos: 1) K-Means, 2) DBSCAN; 3) Agglomerative Clustering. Utilize o ficheiro com o nome "*segmentation data.csv*" para realizar a sua análise. As principais conclusões que devem ser retiraram devem ter em consideração as seguintes dimensões:

- Número de perfis de clientes segmentados;
- Principais características dos clientes segmentados;
- Perfil dos clientes maioritários e minoritários / qual o perfil mais dominante e menos dominante;
- Existem clusters bem separados ou onde a sua diferenciação não é tão clara?

2. Collaborative Filtering:

Como explicado anteriormente no curso, Collaborative Filtering é uma das formas / abordagens para analisar dados explícitos de clientes e items, para que seja possível realizar alguma recomendação relevante para novos clientes e já existentes. Para este caso, o data set a ser utilizado é constituído por anedotas classificadas por cerca de 60,000 utilizadores. Usando Collaborative Filtering, quer CF baseado em items ou Fatorização de Matrizes, escolha 3 utilizadores e faça a sugestão de um top 5 de anedotas por utilizador. Tenham em consideração que os utilizadores escolhidos devem ter gostos por tipos de anedotas diferentes. Justifiquem a escolha dos utilizadores. Desta maneira será possível avaliar se as anedotas sugeridas realmente foram bem-sucedidas. Adicionalmente, descrevam a abordagem escolhida para recomendação de anedotas e justifiquem o porquê dessa abordagem ser a melhor neste contexto.

Conteúdo do dataset

O dataset contem mais de 1.7 milhões de avaliações (-10.00 a +10.00) de 150 anedotas por parte de 59,132 utilizadores. O dataset foi recolhido entre Novembro de 2006 e Maio de 2009. O dataset completo contem 2 ficheiros CSV:

- ratings.csv: Cada linha está formatada como [User ID] [Item ID] [Rating]

- item.csv: Mapeia os IDs das anedotas com as anedotas em si