H&M Personalized Fashion Recommendations

Contribution: Kristina Denisova

The purpose of the competition is to recommend items from H&M shop according to previous purchases made throughout the week. There are simple cases when we can suggest just the most popular clothes categories. However, ML is not about this. So, we decided to provide some clustering methods, such as K-means, to make a recommendation system. It can be done from the consumer case: to define the group of customers who are interested in similar purchases. Or cluster items: what group is interested to a particular customer

Contribution: Tatyana Shelovanova

By working on the project, we managed to:

1. Explore the data.
2. Learn the possible machine learning algorithms that can be applied in a recommendation task .
3. Try KMeans as the first algorithms
4. Learn how Matrix Factorization works.

Contribution: Tatyana Shelovanova

Data description

1) **Articles**. This set includes the data about all items of clothes that H&M sells. They offer 25 features, but we selected 9 of them that are the most interesting ones because other ones just indexes that are unmeaningful.

2) **Customers**. This set includes some information about all H&M customers. The most useful features are 1) club_member_status, 2) fashion_news_frequency and 3) age.

3) **Transactions**. This set includes the information about purchases of a particular article by a particular customer from Sep 20, 2018 to Sep 22, 2020. Here is also the information about the prices of purchases. In total, we have **1 362 281 unique customers** and **104 547 unique articles**. Also there is a sales_channel_id that does not have a description.

Contribution: Tatyana Shelovanova

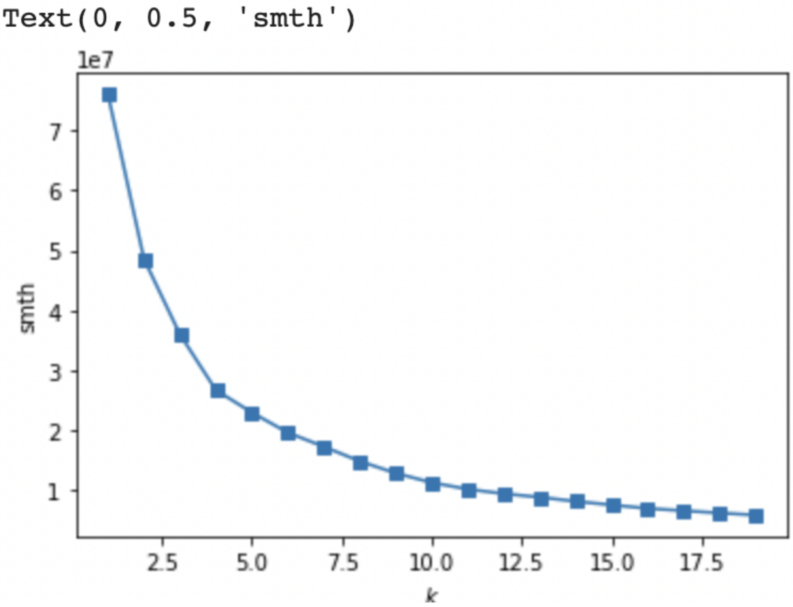
Methods

Our Task: to predict what articles each customer will purchase. For this task we are going to build a Recommender system.

We started from exploring what clustering is and what kind of ML tools can be used for it.

1. Kmeans

We have chosen several features of articles and customers from transaction data. For identifying the optimal number of clusters we decided to use **Elbow Method.** However, we got that the optimal number of clusters is 2, but it seems that we incorrectly specified the features or Kmeans does not work for the complicated task.



```
Text(0, 0.5, 'smth')
```

Contribution: Tatyana Shelovanova

Next steps

As a next step we suggest to implement **Matrix Factorization**. This method should help us to identify the relationship between items' and users' entities. **Matrix Factorization suggests that we should find the parameters based on SGD.**

However, there is a possibility that SGD will not work. In such a case, we would try ALS - Alternative Least Squares. For ALS, we will have to use 2 steps: 1) Use OLS given fixed Q 2) Use OLS given fixed P. Q is the vector for articles, P is the vector for customers.

Then, we will be able to calculate a scalar product of P and Q.

We will use the following metrics:

1) Hitrate - having the right recommendation;
2) Precision - accuracy(average, mean);
3) recall