Machine Learning Lab

Pablo Maldonado

Customer Segmentation

Problem

Our customer, a wine shop, engaged us to create a customer segmentation model to get ready for the Christmas holidays. Instead of “spray and pray” campaigns, they want to identify customer segments among their most loyal wine tasters.

(Data from `http://eu.wiley.com/WileyCDA/WileyTitle/productCd-111866146X.html`)

Deals data

```
import pandas as pd
deals = pd.read_csv("./data/Lec6 deals.csv")
print(deals.head())
```

```
##   Customer Last Name  Offer
## 0              Smith      2
## 1              Smith     24
## 2            Johnson     17
## 3            Johnson     24
## 4            Johnson     26
```

Offers description

```python
import pandas as pd
offers=pd.read_csv("./data/Lec6 offers.csv")
print(offers.head())
```

```
##    Offer  Campaign            Varietal  Minimum Qty  Dis
## 0      1   January              Malbec           72
## 1      2   January          Pinot Noir           72
## 2      3  February           Espumante          144
## 3      4  February           Champagne           72
## 4      5  February  Cabernet Sauvignon          144
##
##   Past Peak
## 0     False
## 1     False
## 2      True
## 3      True
## 4      True
```

Your task

- Create a clustering model from the deals data using kmeans.
- First, you need to tidy up your data: use pivot tables in `pandas` to get your observations such that each row is one customer and each column is one of the offers.
- Kmeans requires a parameter, k. How can you set up the correct number of clusters?
- Join the data you obtain (offer number and cluster number) with the offers data.
- Interprete the clusters. What does it mean to belong to each cluster? Can you identify if there are clusters of french wine lovers? or of bubble fans?

Sentiment Analysis

McDonalds

We have been approached by McDonalds USA to create a predictive model to use in Yelp. The model should scan reviews and assign them a label (in this case, the type of problem the reviewer has) and redirect them to the appropriate customer support agent. The model should identify keywords associated with topics. Those keywords will be later sent to the Big Data engineering team, which will implement suitable search software. Our client is also wants to know if there are some branches that perform particularly bad in different topics.

(Data available from `https://www.crowdflower.com/data-for-everyone/`)

Review data

```python
import pandas as pd
mcdo = pd.read_csv("./data/Lec6 McDonalds-Yelp-Sentiment-DF
print(mcdo.head())
```

```
##                      policies_violated     city  \
## 0  RudeService\rOrderProblem\rFilthy  Atlanta
## 1                        RudeService  Atlanta
## 2           SlowService\rOrderProblem  Atlanta
## 3                                 na  Atlanta
## 4                        RudeService  Atlanta
##
##                                                review
## 0  I'm not a huge mcds lover, but I've been to be...
## 1  Terrible customer service. I came in at 9:30pm...
## 2  First they "lost" my order, actually they gave...
## 3  I see I'm not the only one giving 1 star. Only...
## 4  Well, it's McDonald's, so you know what the fo...
```

Your task

- Create a text classification model. For the reviews that have multiple topics, choose the first one.
- Create a few visualizations for the location and review data. For example, which locations rate worse for bad food? What are the top issues per city?
- (Optional) In the case of multiple reviews, which issues go together more often?