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

```
deals <- read.csv("./data/deals.csv")
head(deals)
```

```
##   Customer.Last.Name Offer
## 1              Smith     2
## 2              Smith    24
## 3            Johnson    17
## 4            Johnson    24
## 5            Johnson    26
## 6           Williams    18
```

Offers description

```
offers <- read.csv("./data/offers.csv")
head(offers)
```

```
##   Offer Campaign           Varietal Minimum.Qty Discount
## 1     1  January             Malbec           72       56
## 2     2  January         Pinot Noir           72       17
## 3     3 February          Espumante          144       32
## 4     4 February          Champagne           72       48
## 5     5 February Cabernet Sauvignon          144       44
## 6     6    March           Prosecco          144       86
##   Past.Peak
## 1     FALSE
## 2     FALSE
## 3      TRUE
## 4      TRUE
## 5      TRUE
## 6     FALSE
```

Your task

- Create a clustering model from the deals data.
- How can you set up the correct number of clusters?
- To get more stable clusters, use the `nstart=25` argument to the `kmeans` function.
- 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

```
mcdo <- read.csv("./data/McDonalds-Yelp-Sentiment-DFE.csv")
str(mcdo)
```

```
## 'data.frame':    1525 obs. of  3 variables:
##  $ policies_violated: Factor w/ 147 levels "","BadFood",
##  $ city             : Factor w/ 10 levels "","Atlanta","
##  $ review           : Factor w/ 1518 levels "\"And on th
```

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?