Nicolas Drizard
nicolasdrizard@g.harvard.edu
CS281-F15

Project Proposal

Due: 23:59pm October 14, 2015

Virgile Audi
vaudi@g.harvard.edu

Introduction

Looking for an applied project, we decided to take part in Yelp Dataset Challenge round 6 ! The dataset is constituted of 1.6 million reviews for 61 thousands business in the UK, the US, Germany and Canada. It also contains 481 thousands attributes such as hours, parking availability or ambience as well as aggregated check-ins over time for the entire base of business. We chose to analyse this data set because it gave us the opportunity to come to conclusions that could eventually be implemented on Yelps platform. In more details, we focus in this project on the classification by categories. When looking for a restaurant, Yelp users enter keywords and get results corresponding to these tag words. But how are these tags associated to each venue? Until now, each owner manually associates the tags to their venue when creating the profile along many attributes that users can then update based on their experiences. Can we use Machine Learning techniques to create a more refine category system? Would it also be possible to find some new categories that are not as obvious as Pizzeria or Fast Food based on the features and characteristics of the venues? How much information can reviews give us on the type of restaurants ? These are some of the questions we are looking to answer in this project.

We now present a road map to the project with possible extensions if time permitted:

1 Data Preparation

The given dataset gathers many information under several formats. We first need to extract relevant features of our data and/or pre-process them to apply our algorithms.

1.1 Numerical Feature Extraction

This part works with two tables of the Yelp data: business and checkin. The first one stores information about the business (localizations, name, categories ...) and the second contains the average number of customers checkins for each hour of the week.
We joined and converted these two tables in order to have only numerical or binary features.
We applied different operations to build the features:

- Dropping irelevant features
- Identifying categorical features and converting them into orthogonal vectors of a N dimensional space, with N = *number of categories*
- One attribute of each business stores different information in an unstructured way (each business does not have the same number of information stored by this attribute). It may concern, for instance, the price, the ambience, if alcohol is offered... We needed to identify the most shared informations to avoid too many missing values

On top of these steps, e chosed to focus first on the restaurants (we filtered the dataset to keep only the entries with 'Restaurants' in their list of categories). Also to build easily our first baselines we chosed not to handle the missing values and droped the business with not enough information.

In conclusion, we built in this part the data set of the Yelp restaurants with 205 features relatives to the customers checkins and the inherent properties of the restaurants. We will test these features under supervised learning algorithms.

Some improvement might be considered:

- Combining features and/or applying them polynomial or cosinus base function to increase their complexity
- Considering more attributes and the business where we don't have the checkin information while filling the missing values. This inference can be done through the average or median values over the entire data set, or through a nearest neighbors estimation.

This preliminary work lead to ve

1.2 Text Processing

Virgile

2 Supervised Learning

2.1 Multinomial Logistic Regression

Considering the data set built with the processing with numerical features. We applied a multiclass logistic regression from scikit learn.

The first question was to narrow down our targets. In the restaurants entries, there are still 261 different categories shared. As the text mining of the reviews seemed more complicated, we needed to consider the most diffenriating categories to have decent models. As a result, we focused on the nationality of the food to test among the restaurants.

With 3 classes: 1 without specific nationality and the two most represented nationalities, Mexican and Chinese, the multiclass logistic regression provided poor results on the test set (the data set was separated in the schema 80:20, train:test), only slightly better than predicting no nationality for each entry (dummy model). The explanations could be that the features do not contain relevant information to differentiate the style and the restaurant set without missing values contains only 13000 entries with around 1000 positive for the two classes.

Precision score:

$$\rho_{dummy} = 0.671423029551$$
$$\rho_{logreg} = 0.726111608768$$

2.2 Dimensionality Reduction

Given the poor results of our model, applying a principal component analysis did not improve the performence. The intrinsic quality need to be improved first.

3 Infering New Subcategories

Once we studied the current label, we would infer latent correlations among the existing categories to come up with new subcategories.

3.1 K-Nearest-Neighbors

The preliminary work on the features will embed the examples in a multi-dimensional space from which we can extract cluster based on the local geometry. We could apply an unsupervised method as the k-nearest-neighbors to infer new subcategories. For instance this clustering method could be used locally in each already known category to infer subcategories or more globally. Another approach if time could be to apply a decision tree, which has the advantage to work on categorical features. One drawback could be its exposure to overfitting.

3.2 Latent Dirichlet Allocation with network component

Using the results about the venues' clustering using the regular K-NN algorithm and taking them into account when performing LDA, we will look at a possible improvement of the performance. In other words, we will try to add an extra node in the graphical model representing the structure of the graph that has an influence on the topic node.

4 Evaluation methods

To assess the performance of the algorithms used in part B, we will divide the data set into a training and test sets in order to make prediction on the test set. We will evaluate the logistic regression through the classical classifiers metrics considering each class separately; we will compute the confusion matrix and extract the accuracy and the recall. Evaluating the new subcategories will be a challenge as it remains very subjective. We can always assess the performance of the LDA model using the perplexity metric but how good this will translate in terms of category classification still needs to be investigate.

5 Possible Extensions

- Online variational inference for LDA (https://www.cs.princeton.edu/ blei/papers/HoffmanBleiBach2010b.pdf) to speed up convergence
- Use Hierarchical Dirichlet Process to relax the assumptions of the number of categories

6 Work Division

Given our personal affinity, we subdivided each part.

- **Nicolas** : *Feature Extraction*, *Dimensionality Reduction*, *Multinomial Logistic Regression* and *K-nn*
- **Virgile**: *Text Processing*, *LDA* and *LDA with network component*