Machine Learning Engineer Nanodegree

Capstone Proposal

What's Cooking?

Albert Pan
September 3, 2018

Proposal

Domain Background

Food plays a major part in any culture around the world. In fact, the geographical characteristics and cultural associations of a region directly influence their cuisines. By investigating the ingredients used in various cuisines, we can gain a better understanding of the geographical and cultural landscape of different regions.

Han Su et. al. [1] has worked on investigating if recipe cuisines could be identified by their ingredients, using data from food.com. They treated each ingredient as a feature and examined the common ingredients for each cuisine. Their study provides good insight on how to approach this project and what results we could expect to achieve.

For this project, we can use machine learning techniques to attain useful predictions with our data. This project will give me an opportunity to work with real-world datasets and to learn about the relation between ingredients and cuisines. This kind of research can be expanded to other fields of study involving text classification.

Problem Statement

Using the data provided by Yummly (https://www.kaggle.com/c/whats-cooking-kernels-only/data), the challenge is to predict the cuisine of the dish from its list of ingredients. More specifically, this would be multi-class classfication problem, as we have 20 different cuisines we can predict. We can use a machine learning model that utilizes this data to predict the appropriate cuisine.

My strategy for solving this problem is as follows:

1. Download data from Kaggle
2. Explore data with visualizations
3. Preprocess data and extract features
4. Train and test model
5. Tune hyperparameters

Datasets and Inputs

The dataset is composed of two JSON files. This dataset (https://www.kaggle.com/c/whats-cooking-kernels-only/data) is made available from a Kaggle competition provided by Yummly.

The most important file is the `train.json` file, which has 39774 rows of data and contains three fields, `id`, `cuisine`, and `ingredients`. The `ingredients` field consists the list of ingredients for each dish, while our target variable `cuisine` denotes the cuisine. The `id` field is a unique integer identifier for that particular dish. The `test.json` file has all the fields mentioned in the `train.json` file with the exception of our target variable, `cuisine`, and it has 9944 rows of data. In particular, I will be using the `ingredients` field as my features that I will pass into my model to train and predict the cuisine of the dish represented by the list of ingredients. Once I obtain the predictions of the data points in `test.json`, I will couple the `id` of the dish alongside with its respective prediction and submit it for the Kaggle What's Cooking (https://www.kaggle.com/c/whats-cooking-kernels-only) competition.

The distribution of the training dataset is shown below:

Cuisine	Count	Cuisine	Count
italian	7838	spanish	989
mexican	6438	korean	830
southern_us	4320	vietnamese	825
indian	3003	moroccan	821
chinese	2673	british	804
french	2646	filipino	755
cajun_creole	1546	irish	667
thai	1539	jamaican	526
japanese	1423	russian	489
greek	1175	brazilian	467

It seems that this data is unbalanced, as there are much more "italian" and "mexican" cuisines than any of the the other cuisines. This is something that we will have to be careful about when we train and evaluate our model.

Solution Statement

Dishes of the same cuisine will likely share a common set of ingredients. I will be taking a look at the most frequent ingredients used for each cuisine, and use the individual ingredients as input features for the model I will be training.

I will begin by using a variety of classification models that we have learned throughout the course, such as SVMs, Decision Trees, Random Forest, etc., as well as the XGBoost model and possibly deep learning models. I will additionally explore the viability of clustering algorithms and see if dishes of the same cuisine cluster together.

Benchmark Model

For our baseline benchmark, we can use the metric obtained by predicting the most common cuisine in the training and testing datasets. We can use a metric such as accuracy and compare the results of our to-be-trained model with the results of this baseline benchmark.

Evaluation Metrics

As previously mentioned, our training data seems to be unbalanced, and thus I will be using the F_1-score to evaluate the model. The F_1-score is the weighted average of the precision and recall, and can be expressed mathematically with the following form:

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

We can also take a look at the precision and recall metrics on their own. Precision can be expressed with:

$$precision = \frac{\#\ cuisine\_x\ dishes\ classified\ as\ cuisine\_x}{\#\ dishes\ classified\ as\ cuisine\_x}$$

and recall can expressed with:

$$recall = \frac{\#\ cuisine\_x\ dishes\ classified\ as\ cuisine\_x}{\#\ cuisine\_x\ dishes\ classified\ as\ cuisine\_x + \#\ cuisine\_x\ dishes\ incorrectly\ classified}$$

Project Design

This project will consist of the following steps:

1. **Data Exploration and Feature Extraction**
   I will visualize the dataset and remove any outliers, and I will clean up the dataset by identifying the ingredients that are not indicative of our target classes and removing them. I can visualize the ingredients that are associated with their respective cuisines, and decide whether or not those ingredients can be used as features. I will then split the dataset into a training and testing set.
2. **Train Model**
   I will train a variety of machine learning models and select the best model using cross validation. I will then optimize their hyperparameters using GridSearchCV.
3. **Test and Optimize Model**
   Once the model is trained, I will use the model to make predictions on the test set given from Kaggle. Depending on the results, I can adjust the model's parameters to improve the score.

[1]: Su, Han & Lin, Ting-Wei & Li, C.-T & Shan, Man-Kwan & Chang, Janet. (2014). Automatic recipe cuisine classification by ingredients. UbiComp 2014 - Adjunct Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing. 565-570. 10.1145/2638728.2641335.