Capstone Project

Machine Learning Engineer Nanodegree

Albert Pan

November 11, 2018

1 Definition

1.1 Project Overview

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.

1.2 Problem Statement

Using the data provided by Yummly[2], the challenge is to predict the cuisine of the dish from its list of ingredients. More specifically, this would be multi-

[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.

[2] https://www.kaggle.com/c/whats-cooking-kernels-only/data

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

1.3 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:

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{1}$$

We can also take a look at the precision and recall metrics on their own. Precision can be expressed with the following equation, where $dish_x$ is a dish of a particular cuisine x:

$$precision = \frac{dish_x\, correct}{total\ dish_x} \tag{2}$$

where $dish_x\, correct$ is the number of dishes of cuisine x that are correctly classified as x, and $total\ dish_x$ is the the total number of dishes that were classified as x.

Recall can expressed with:

$$recall = \frac{dish_x\, correct}{dish_x\, correct + dish_x\, incorrect} \tag{3}$$

where $dish_x$ *correct* is the number of dishes of cuisine x that are correctly classified as x, and $dish_x$ *incorrect* is the number of dishes of cuisine x that are incorrectly classified.

2 Analysis

2.1 Data Exploration

2.2 Exploratory Visualization

2.3 Algorithms and Techniques

2.4 Benchmark

For our baseline benchmark, we can use the metric obtained by predicting the most common cuisine in the training and testing datasets. The most common cuisine is *Italian*, and our benchmark model will predict Italian to all recipes. This would give us a benchmark F_1-score of 0.3292.

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.

3 Methodology

3.1 Data Preprocessing

3.2 Implementation

3.3 Refinement

4 Results

4.1 Model Evaluation and Validation

4.2 Justification

5 Conclusion

5.1 Free-Form Visualization

5.2 Reflection

5.3 Improvement