Machine Learning Engineer Nanodegree

Capstone Proposal : Finding the cuisine

Mandeep Kaur
March 27, 2017

Proposal

Domain Background

Cuisine is the style of cooking. Every culture in this world has its own cuisine. Every cuisine has its own unique taste. Everyone can agree that this uniqueness comes from the choice of ingredients. Nobody can deny that few flavours are strongly related to some certain parts of world. Many times people, including the food critics, try to define a cuisine entirely on the bases of its ingredients.

In this project, I am going to create model that finds a relationship between a cuisine and its ingredients.

Problem Statement

The goal is to create a model that can, given the ingredients of a recipe, find the cuisine of that recipe it belongs to.

Datasets and Inputs

For this project the dataset is publically available on Kaggle's competition "What's cooking?"[1] The dataset includes the recipe id, cuisine and list of ingredients.

An example is:

```
{
 "id": 24717,
 "cuisine": "indian",
 "ingredients": [
     "tumeric",
     "vegetable stock",
     "tomatoes",
     "garam masala",
     "naan",
     "red lentils",
     "red chili peppers",
     "onions",
     "spinach",
     "sweet potatoes" ]
 }
```

This dataset contains a total of 6714 unique ingredients and 39774 sets of recipes' ingredient lists with their cuisines. This dataset has 20 different cuisines.

Solution Statement

This is a classification problem. We have to classify the recipe into one of the 20 cuisines based on the list of ingredients it has. First I will breakdown the ingredient list into different columns and then by using a classification algorithm I can train the model.

Measurable: with the help of the below defined evaluation metric we can easily evaluate the efficiency of our model. Replicable: This problem is easily reproducible. Anyone can define a list of ingredients of a recipe and replicate it. Quantifiable: The problem can easily be expressed in terms of mathematics. We can encode the list of ingredients because this a categorical data and then run it through a classification algorithm.

Benchmark Model

For the benchmark model my choice is to use decision tree with the following parameters:
criterion : gini,
min_samples_split:2
min_samples_leaf=1
(This is the default configuration of the decision tree classifier in sklearn)

Decision tree classifier is a classic algorithm for this type of problem, so this is the best choice to create a benchmark model. The accuracy of predictions made on test set with this model will be used as a result to compare with our model.

Evaluation Metrics

The most suitable evaluation metric for this problem is Accuracy. This is a classification problem and the efficiency of the solution can be evaluated on the number of cuisines it can identify accurately.

So we can simply define Accuracy in mathematical terms as:
Acc = Number of recipe ingredients lists correctly classified / Total number of recipes

Our test set will be the one on which we never trained our model. So this provide a very good parameter to check for accuracy of our model because it has never been seen before. During training we divide the data into train and validation set and we check the accuracy on that validation set.

Project Design

The workflow for this solution using python is :

1. Download the datasets
2. Split the columns into two sets: X for ingredient list and Y for respective cuisines
3. Now encode the ingredients list into proper multiple columns with the ingredient present as 1 and not present as 0
4. Divide the data into 3 parts train, validation and test
5. For solution I will be using random forest classifier and grid search with different parameters.
6. Train this model on train data and check the efficiency against the validation set.
7. Once we have optimal parameters and trained model now we can check the results on test model and compare it with our benchmark model.
   To validate it further, Kaggle provides a test dataset. By passing that testdata through this model we can submit the results to check the efficiency of model on public leaderboard[2].

----

citations:

[1] https://www.kaggle.com/c/whats-cooking

[2] https://www.kaggle.com/c/whats-cooking/leaderboard