1 Introduction

2 Assumptions and Symbols

2.1 Model Assumptions

* It is independent that each ingredient is used in each type of cuisine.

This is actually a very strict assumption as this may abandon lots of information in the original data set. The usage of chocolate and beef, which sound really weird if they are put together, would be much negatively correlated. But in our model we just make this independence assumption due to convenience and the lack of data, and actually the result is much acceptable.

* The selection of recipes from the website is unbiased.

That is to say, there is no preferences of ingredients in all those recipes. This concern can be improved if we select data from more websites around the world.

2.2 Parameters and Symbols

: The probability that one particular ingredient is used in one particular cuisine .

: Number of recipes of cuisine type .

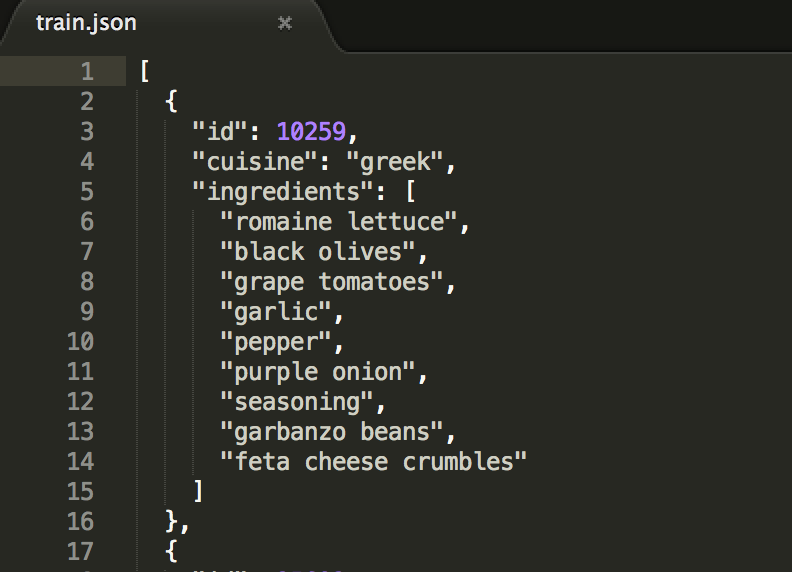
: Number of recipes of cuisine type that use ingredient .

3 Data cleaning

Original data is in ***json*** dataframe acquired from yummly.com and it consists of

1. known recipes with cuisine types and ingredients
2. test recipes with ingredients only

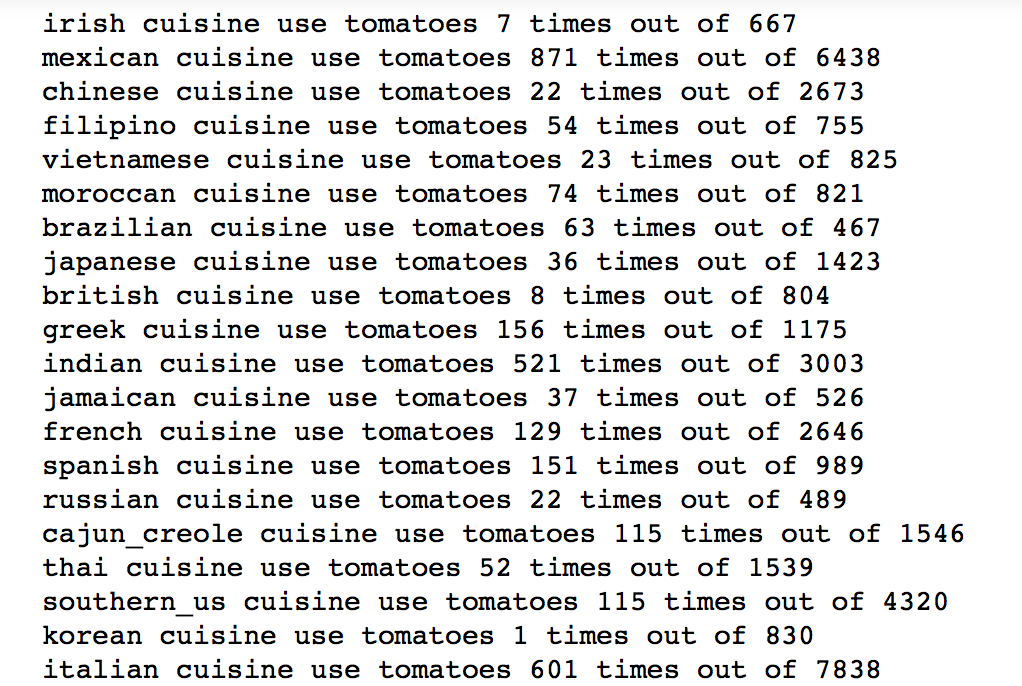
Example of original data is in the figure below.



There are 39774 known recipes and 9944 test recipes in total. All the known recipes are from 20 different types of cuisines, and within all these a total of 6714 different ingredients are used.

All the names of cuisines and ingredients are expected to be read in python as the string type. However, there are words like ‘purée’ that may not be directly manipulated as the string type. We encode all words with UTF-8 coding style for this kind of situation. That is, for example ‘purée’ would be considered as ‘pur\xc3\xa9e’.

What we do next with the data is some basic statistics. We count how many times an ingredient is used in all recipes and in each type of cuisine. We also count the number of each kind of cuisine with all known recipes. Sample outputs of the statistics are in the figure below.



4 Statement of the Model

4.1 Model Overview

Our issue is a supervised learning problem and our aim is to establish a classification method for accurate predictions. The main idea of our model is the maximization of probability. That is, we calculate the probability that all ingredients in the recipe are used in each certain kind of cuisine. The cuisine type with higher probability would be more likely to be the actual cuisine type of the recipe.

The first step is to infer the probability that one particular ingredient is used in one particular cuisine . We use two difference methods on the inference, naïve MLE – the more natural method and Bayesian inference – the more scientific method. Then we multiply the probabilities under the assumption of independence and compare these products to make the final prediction.

4.2 Naïve MLE

Our cleaned data gives that within all recipes of the cuisine type , of which use the ingredient . The most ‘natural’ inference would be that the probability that ingredient is used in cuisine type is the ratio of the usage:

On the other way, we may use a more mathematical language to translate this inference. Using or not using the ingredient, we consider that as a Bernoulli random variable with parameter . Now that we have experiments and among those experiments are 1 while others are 0. By applying MLE on the estimation of the parameter , we want to find that maximizes

By taking logarithm and find the maximum using derivatives, we can get that the optimal is

It is exactly what the ‘naïve’ inference has shown.

Simple calculation shows that the moment estimation leads to the same result as well.

4.3 Problems and improvements with naïve MLE

The result shown above seems reasonable and applicable at first sight. However, if we take more considerations, we may find some shortages of this method. If any ingredient is never used in one cuisine type, the inference above gives that , which further indicates that the product for comparison is 0. That is, this cuisine type is absolutely rejected whatever other ingredients are. This may exclude the conditions of possible recipes that tries novel ingredients that might be seldom used in its original cuisine type. Another trouble related to the 0 is that, there would be some ingredients (the accurate number is 422) in the test data that never appear in the train data. Then if we analysis recipes related to these ingredients, the product we get for all cuisine types would be 0 and we cannot find the best-fit cuisine type. We make the following modification on the MLE for these issues:

This modification will change the number by very little amount for non-zero while solve the problem of the zero prediction of probability.

4.4 Bayesian Inference

Bayesian inference is another approach for estimating the parameter of a Bernoulli distribution. The most essential advantage of this method is that we will not have a zero estimation of the probability, thus the problems mentioned above can be effectively resolved.

We assume the prior distribution on is beta. Then if our data gives of 1 among total experiments, the posterior distribution would be beta, which gives us the Bayes estimator of

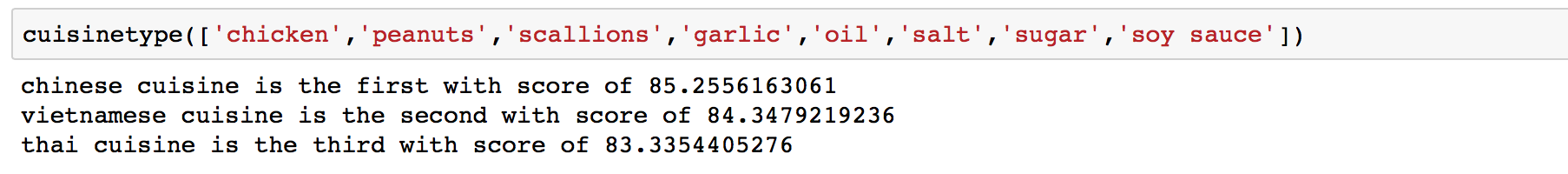
Back to our problem, now we would need to establish a prior belief of the distribution of , the probability of one ingredient used in one cuisine type. We use all the naïve MLE estimations of each ingredient used in each cuisine type, that is a total of 20\*6714 data to fit a beta distribution. By comparing the sample mean and the mean of the distribution we choose beta as our prior distribution. Then we can estimate in our model as

5 Results

Our model gives a great prediction on both train data and test data. One unexpected issue is that the Bayesian inference do not return a better result, and this might due to that the choose of the prior belief is not so reasonable. The table shows the accuracy of the table using different methods for estimating probability.

|  |  |  |
| --- | --- | --- |
| Accuracy of the Model | | |
|  | Train Data | Test Data |
| MLE | 79% | 69% |
| Bayesian | 74% | 61% |

We also apply our model on recipes that we ourselves may cook at home everyday. For example, we have made Kung Pao Chicken, a famous Chinese cuisine and we know what ingredients would be used in it.



Fantastic model!

6 Conclusions

7 References