Recipe Identification: Multiclass Classification

Project Proposal: Stanford University CS229

Course	CS229: Machine Learning (Fall 2018)
Project	Where is the Chef From?
Category	General Machine Learning, Multiclass classification, Food and Drinks
Team	Manish Pandit (manish7), Annie Pitkin (apitkin), Hermann Qiu (hq2128)

Objective:

In this project we aim to predict the category of a dish's cuisine given a list of its ingredients. Food is such a large part of the daily human experience. Our strongest geographic and cultural associations are tied to a region's local foods. Linking recipes to the country of origin is valuable, especially under multiclass classification conditions.

''If you're in Northern California, you'll be walking past the inevitable bushels of leafy greens, spiked with dark purple kale and the bright pinks and yellows of chard. Across the world in South Korea, mounds of bright red kimchi greet you, while the smell of the sea draws your attention to squids squirming nearby. India's market is perhaps the most colorful, awash in the rich hues and aromas of dozens of spices: turmeric, star anise, poppy seeds, and garam masala as far as the eye can see. Some of our strongest geographic and cultural associations are tied to a region's local foods.''

Challenges:

There are overall thousands of features in the dataset, however only a small number of feature values are true for each sample. There is a feature engineering challenge in that some of the ingredients are commonly used across multiple cuisines. In addition, the high level analysis of data reveals that there is going to be quite of bit of data cleansing required. For example, the dataset contains same ingredients listed with small variations in the spelling. Some of the ingredients includes special characters and numbers (measures 25g etc.).

Dataset:

The dataset from Kaggle challenge: What's for dinner? The data are stored in JSON format.

- train.json - the training set containing recipes id, type of cuisine, and list of ingredients
- test.json - the test set containing recipes id, and list of ingredients

https://www.kaggle.com/kaggle/recipe-ingredients-dataset/home

Method and Algorithm:

- **Data preparation:** Data cleansing and visualization.
- **Model selection:** We plan to research across following options and implement few options that we think will yield best results.
    - Naive Bayes
    - K-Nearest Neighbor
    - Deep Neural Networks
    - Decision Trees
    - Random Forest
    - Extended Support Vector Machine
- **Training:** Train the model against the training dataset.
- **Evaluation:** Evaluate model in terms of accuracy and other relevant metrics.
- **Parameter tuning:** Tune and optimize the hyper parameters of the model.
- **Predictions:** Predict against the test dataset and publish the results.

Expected results, qualitative & quantitative:

The output of the algorithm is the country of origin. The program would be able to take in a recipe with any number of ingredients and identify the country of origin. Quantitatively, we would evaluate the results by the classification accuracy (misclassification error) of the country of origin against the test dataset. Additionally could output a confusion matrix to identify patterns causing errors.

Reference Articles:

1. *Multiclass classification. https://en.wikipedia.org/wiki/Multiclass_classification*
2. *Softmax function: https://en.wikipedia.org/wiki/Softmax_function*
3. *Naive Bayes Classifier: https://en.wikipedia.org/wiki/Naive_Bayes_classifier*
4. *Mohamed, Aly (2005). "Survey on multiclass classification methods" (PDF). Technical Report, Caltech.*