[redacted]

Project 2: How long will it take to cook this?

[redacted]

1 Overview

The goal of this Project is to build and critically analyse supervised Machine Learning methods, to predict the cooking time for recipes based on their steps, ingredients and other features. The cooking time of a recipe has been categorised into three classes, corresponding to quick, medium and slow.

This assignment aims to reinforce the largely theoretical lecture concepts surrounding data representation, classifier construction, and evaluation, by applying them to an open-ended problem. You will also have an opportunity to practice your general problem-solving skills, written communication skills, and creativity.

This project has two stages. The main focus of these stages will be the written report, where you will demonstrate the knowledge that you have gained and the critical analysis you have conducted in a manner that is accessible to a reasonably informed reader.

2 Deliverables

[redacted]

1. **Report**: an **anonymous** written report, of 1000-1500 words (for a group of one person) or 2000-2500 words (for a group of two people)
2. **Output**: the output of your classifiers, comprising predictions of labels for the test instances, submitted to the Kaggle[1] in-class competition described below.
3. **Code**: one or more programs, written in Python, which implement machine learning models, make predictions, and evaluate the results.

[redacted]

[redacted]

[1] https://www.kaggle.com/

3 Terms of Use

The data has been collected from Food.com (formerly GeniusKitchen), under the provision that any resulting work should cite this resource:

> Generating Personalized Recipes from Historical User Preferences. Bodhisattwa Prasad Majumder, Shuyang Li, Jianmo Ni, Julian McAuley, in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019.

[redacted]

4 Data

[redacted]

The recipes are collected from Food.com[2], which is a platform that allows the user to publish recipes and comments on others' recipes. In our dataset, each recipe contains:

- recipe features: `name`, `ingredients`, `steps`, `number of steps`, and `number of ingredients`
- text features: produced by various text encoding methods for `name`, `ingredients`, and `steps`. Each feature is provided as a single file with rows corresponding to the file of recipe features.
- class label: the preparation time of a recipe `duration` (3 possible levels, 1, 2 or 3)

You will be provided with training set and a test set. The training set contains the recipe features, text features, and the `duration`, which is the "class label" of our task. The test set only contains the recipe and text features without the label.

The files provided are:

- *recipe_train.csv*: recipe features and class label of training instances.
- *recipe_test.csv*: recipe features of test instances.
- *recipe_text_features_*.zip*: preprocessed text features for training and test sets, 1 zipped file for each text encoding method. Details about using these text features are provided in README.

5 Task

You are expected to develop Machine Learning models to predict the preparation of a recipe based on its features (e.g. name, ingredients, steps etc.). You will implement and compare different machine learning models and explore the effective features for this task.

[2] `https://www.food.com/`

- **The training-evaluation phase**: The holdout or cross-validation approaches can be applied on the training data provided.
- **The test phase**: the trained classifiers will be evaluated on the unlabelled test data. The predicted labels of test cases should be submitted as part of the Stage I deliverable.

Various machine learning techniques have been (or will be) discussed in this subject (0R, Naive Bayes, Decision Trees, kNN, SVM, neural network, etc.); many more exist. You may use any machine learning method you consider suitable for this problem. *You are strongly encouraged to make use of machine learning software and/or existing libraries (such as* `sklearn`*) in your attempts at this project.*

In addition to different learning algorithms, there are many different ways to encode text for these algorithms. The files in *recipe_text_features_*.zip* are some possible representations of the name, ingredients and steps of recipes we have provided. For example, one of the encoding method is `CountVectorizer` in `sklearn`, which converts text documents into "Bag of Words" – the documents are described by word occurrences while ignoring the relative position information of the words. You can use these representations to develop your classifiers, but you should also feel free to extract your own features from the raw recipe features, according to your needs. Just keep in mind that any data representation you use for the text in the training set will need to be able to generalise to the test set.