Interim report
CS6350 Machine Learning

Montgomery Carter Shaobo He

October 29, 2015

1 Recap

The goal of our proposed project is to predict the cuisine type of a recipe based on the recipe's ingredient list[1]. In the dataset we are given a list of recipes. For each recipe we are given cuisine type and a list of ingredients.

2 Milestones Achieved

Our first task was to acquire the training and test data[2] from kaggle.com and ensure its formatting was proper for machine processing.

Next we wrote a python program to read the data stored in JSON format and convert it into a vector data structure so that it can be processed smoothly by machine learning algorithms. To be more specific, each recipe becomes a vector where each dimension represents a possible ingredient. If the dimension's value is 0, the ingredient is not present in the recipe; if the value is 1, the ingredient is present in the recipe. The final dimension of the vector is the label. Because we are dealing with multiple cuisines, we have a multi-class classification problem (which is addressed in further detail below). For now, we mapped each possible cuisine to an integer value, and this value is inserted as the last dimension in each recipe vector. We recognize that when we get further into analysis using specific algorithms, it is likely that we will have to change the representation of the label.

Having transformed the data, we also did some preliminary analysis of the data. For example, we noticed that there are about 6000 different possible ingredients (meaning 6000 dimensions per vector), which may prove to be a challenge for linear classification. However, we also noticed that there are a number of different ingredients which seem to be only slight spelling variations of the same actual ingredient. Condensing such variants into a single ingredient may help with reducing dimensionality.

We realize that perhaps our biggest obstacle with this project is going to be the multi-class classification. We haven't really addressed such algorithms in class, so we are reading about ways to address multi-class classification.[3]

3 Plan

The first thing of our plan is to refine the data so it can be processed more accurately by machine learning algorithms. The reason of this input data refinement is that we notice that there are around 6000 sorts of ingredients in our training data set, which can make it hard for K nearest neighbor algorithm to classify the data set. Moreover, duplicate ingredients exists with spelling variantions so we should be able to remove the naming redudancy and thus reduce the dimensionality of the training data set.

The next step is to apply machine learning algorithms on the training set. The first algorithm we would like to try is decision tree since it is a simple algorithm to learn data with multiple labels. Another reason is that we noticed that many ingredients are only associated with one individual cuisine. Thus it may be easy for decision tree to generalize the training data.

If decision tree does not work well, then we will turn to algorithms that convert the problem of multi-class classification into multiple binary-class classification problems. For instance, we can try one-vs-one multi-class classification[3] and one-vs-res multi-class classification[3].

References

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

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

[3] `https://en.wikipedia.org/wiki/Multiclass_classification`.