

WHAT'S COOKING: PREDICTING CUISINES BASED ON INGREDIENTS OF RECIPES

He Huang, Ye Liu, Lichao Sun, Zhu Wang, Congying Xia, Fan Zhu

Department of Computer Science

The University of Illinois at Chicago

Chicago, IL, 60607, USA

{hehuang, yliu279, lsun29, zwang260, cxia8, fzhu20}@uic.edu

ABSTRACT

“There exists no problem in the universe that delicacies cannot solve”, said an ancient prophet. As mortal humans, we are all slaves of food, and some of us are so captivated by food that we will do anything to taste all possible cuisines on this planet and beyond. In this project, we explore the recipes of 20 different cuisines, and use several machine learning algorithms to predict the cuisine of each recipe based on its ingredients. We also compare the performance of these algorithms and discuss the factors that affect their performance.

1 INTRODUCTION

Our project is based a fun and delicious competition on Kaggle¹.

Picture yourself strolling through your local, open-air market...What do you see? What do you smell? What will you make for dinner tonight?

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, but sometimes even the local people cannot distinguish some local dishes from exotic ones. Therefore, we try to design a machine learning model that can automatically recognize the category of a dish's cuisine given its ingredients.

1.1 DATA DESCRIPTION

We use the data distributed on Kaggle. The training data contain 39774 recipes that are stored in Json format, and each recipe contains three attributes:

- id: the cuisine ID
- cuisine: the category of cuisine that this recipe belongs to.
- ingredients: the ingredients of the recipe

There are 20 categories of cuisines, and 7137 different ingredients in total.

1.2 MACHINE LEARNING TASK DEFINITION

We formulate this task as a multi-class classification problem. Each recipe can be represented as a 7137-length vector, which can be used as the input of the machine learning algorithm. The output

¹<https://www.kaggle.com/c/whats-cooking>

vector is the probability distribution of the recipes on 20 categories, where the probabilities sum to 1. Finally we pick the one with highest probability as our prediction.

The main evaluation metric we use is the overall prediction *Accuracy*.

1.3 DATA PREPARATION

Although Kaggle provides a testing dataset, it has no ground-truth labels, and we can only submit our results 5 times per day to the Kaggle online judge to get the accuracy. To deal with this situation, we randomly sample 10% of the training data as our testing data, while maintaining the proportion of corresponding cuisines across two datasets as the same, so that the performance will not be significantly affected by different sampled testing data. The proportions of 20 categories of cuisines are shown in Table 1.

Table 1: Proportions of each category of cuisine

brazilian	british	cajun creole	chinese	filipino
1.2%	2.0%	3.9%	6.7%	1.9%
french	greek	indian	irish	italian
6.7%	3.0%	7.6%	1.7%	19.7%
jamaican	japanese	korean	mexican	moroccan
1.3%	3.6%	2.1%	16.2%	2.1%
russian	southern us	spanish	thai	vietnamese
1.2%	10.9%	2.5%	3.9%	2.1%

1.4 DATA CLEANING AND PREPROCESSING

All artwork must be neat, clean, and legible. Cooking is the supreme art, and so that the ingredients should be as neat, clean and legible as well. We first change all letters into lower cases, then remove digits and special characters, remove stop-words and do the stemming. After doing so, we calculate the Bag-of-Word (BoW) feature vectors as well as the TF-IDF feature vectors for each recipe.



Figure 1: Cooking is the supreme art of the universe

2 DATA EXPLORATION

Before diving into the magical black hole of machine learning, let's first have a look at some interesting things we discover in the dataset.

2.1 HISTOGRAM OF THE COUNTS OF CUISINES

As is shown in Figure 2, we can see that Italian cuisine has the largest number of recipes, so there is no wonder why people say that Italy has some of the best food on this planet. Mexican food comes second, which may be due to the fact that Mexico shares much of its boundary with the United States. Affected by the Mexican foodies, people in southern U.S. also have a lot of ideas about food,

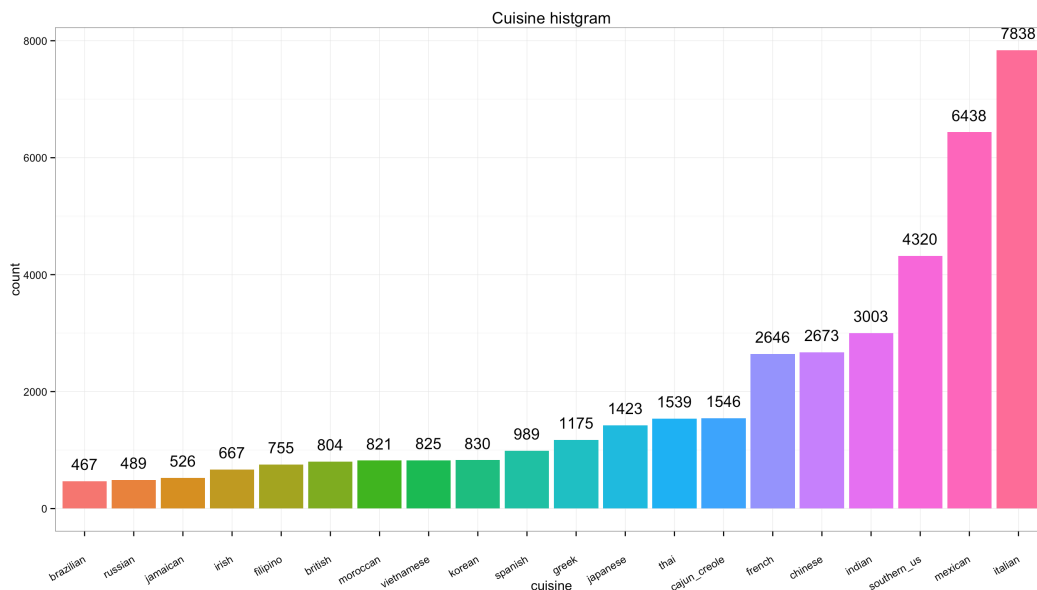


Figure 2: Histogram of the number of cuisines in each category

so we can see that southern U.S. cuisine has the third largest number of recipes. With the special spice and curry, as well as the large amount of immigrants to the U.S., Indian cuisine takes the fourth place. Maybe because of the long distance, Chinese and French dishes only take the fifth and sixth places, but we should never underestimate the taste of Chinese and French food, since China and French are both praised as “The Country of Food”.

2.2 CUISINE CLUSTERING

We use TF-IDF feature vectors, and then apply Principle Component Analysis (PCA) on the dataset to reduce the features into 2-dimensional vectors. k-Means algorithm with $k = 5$ is performed to get the clusters. We visualize the results in a 2-dimensional space, where clusters are shown in different colors, and the sizes of circles are calculated by the Jaccard Similarity (one vs. the other cuisines in this cluster).

As we can see in Figure 3, Chinese, Japanese and Korean cuisines are grouped into the same cluster, which may represent the fact that they are all Asian countries and use similar ingredients in their dishes. Although Vietnam and Thailand are in a purple-color cluster different than the blue group, they are still very close to it, since they are still part of Asia.

A surprise is that the European countries split into two non-adjacent groups, with France and Italy leading one of them. We can easily understand that Russian and southern U.S. cuisines are in a group with British cuisine, since those countries have a long related history. However, it is hard to understand why Indian and Brazilian food is closer to the Italy group than other groups, and we should remember that India is an Asian country and Brazil is very very far from Italy.

2.3 WORD CLOUD

We plot the word cloud using the ingredients that have frequencies more than 1000, as is shown in Figure 4. The larger the word, the higher the frequency of that ingredient.

3 MACHINE LEARNING ALGORITHMS

Now let's come to the fanciful machine learning part. We use five machine learning algorithms:

- Random Forest

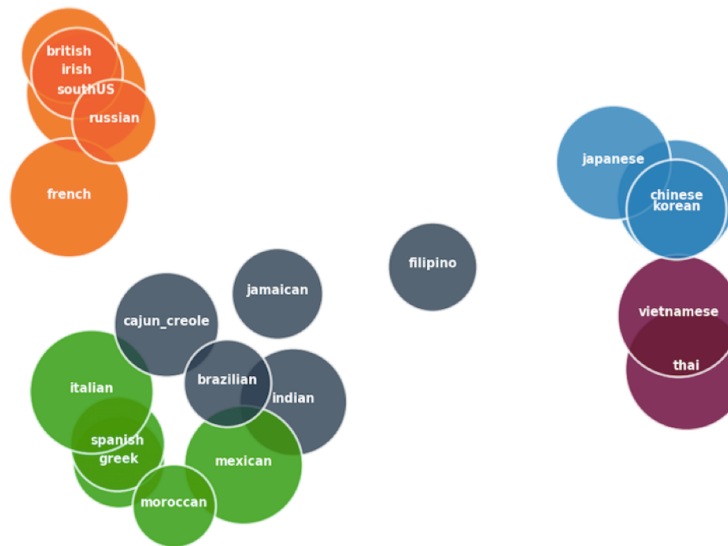


Figure 3: Visualization of cuisines after PCA and k-Means



Figure 4: Word cloud of ingredients

- Naive Bayes
- Linear Support Vector Machine (LinearSVM)
- Latent Dirichlet Allocation (LDA)
- Neural Network

3.1 RANDOM FOREST

According to Wikipedia, random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean

prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of over-fitting to their training set.

In our project, we use random forest for multi-class classification. We use "RandomForestClassifier" function from the Python package *sklearn.ensemble* to build random forest classifiers. The most important part we have done in using random forest is parameter tuning.

We take four parameters into consideration. They are:

- *n_estimators*. It represents the number of trees in the forest. When we increase the number of trees in the forests, the performance was improved, but the speed of the code becomes slower. So we set a series of *n_estimators* to do experiments to find out the optimal *n_estimators*. This is the result of our experiment, shown in Figure 5

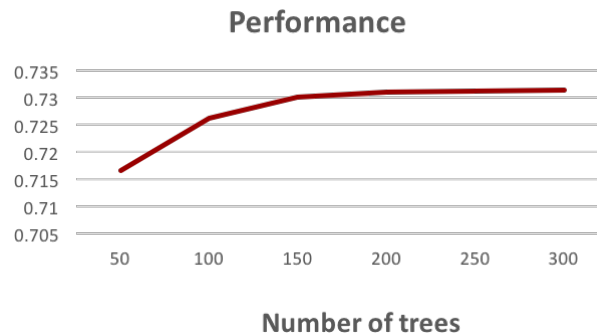


Figure 5: Accuracy w.r.t number of trees

In order to obtain a good performance and a fast speed, we choose *n_estimators* = 200 as our optimal parameter.

- *max_features*. The number of features to consider when looking for the best split. There are different types of *max_features* we can choose. The most frequent used tree types are:
 - *sqrt()*, then *max_features* = *sqrt(n_features)*
 - *log2()*, then *max_features* = *log2(n_features)*
 - None, then *max_features* = *n_feature*

If the *max_features* is too large, then the speed would be too slow. Also, reduce the maximum feature number it can have, avoids over-fitting. Empirical good default values are *max_features* = *n_features* for regression problems, and *max_features* = *sqrt(n_features)* for classification tasks *max_features*. In our experiments, we compare *sqrt()* and *log2()*. It turns out that *log2()* yields better performance. The results are shown in Figure 6.

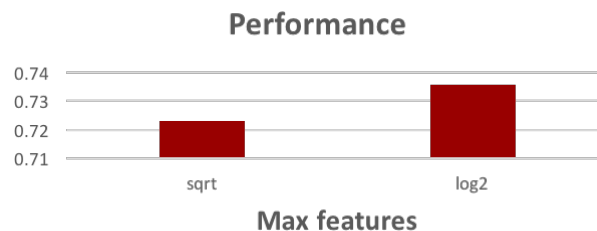
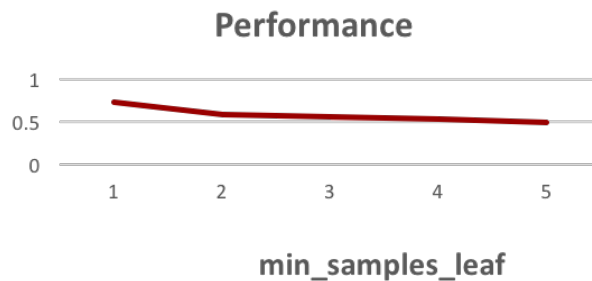
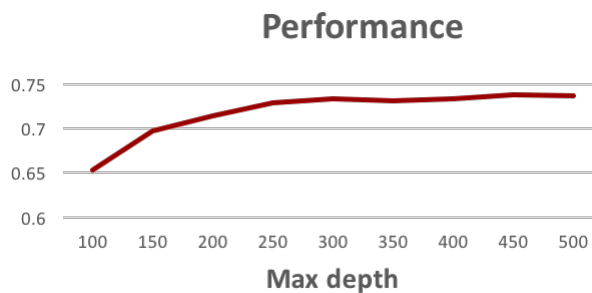


Figure 6: Accuracy w.r.t different *max_features*

- *Min_samples_leaf*. The minimum number of samples required to be at a leaf node. If int, then *consider_min_samples_leaf* as the minimum number. The results are shown in Figure 7. In our project, we set *min_samples_leaf* to 1.

Figure 7: Accuracy w.r.t different *Min_samples_leaf*

- *max_depth*. The maximum depth of the tree. If None, then nodes are expanded until all leaves are pure or until all leaves contain less than *min_samples_split* samples. The results are shown in Figure 8.

Figure 8: Accuracy w.r.t different *max_depth*

In this figure, it shows that add the depth of tree can improve the performance. In our case, since we only have 200 trees in the forest, we can just set this parameter as default and let nodes expanded until all leaves are pure or until all leaves contain less than *min_samples_split* samples.

Overall, the best performance of Random forest is **73.7%**.

3.2 NAIVE BAYES

Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. It is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle: all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable. For example, a fruit may be considered to be an apple if it is red, round, and about 10 cm in diameter. A naive Bayes classifier considers each of these features to contribute independently to the probability that this fruit is an apple, regardless of any possible correlations between the color, roundness, and diameter features.

Table 2: Results of Naive Bayes algorithm

	Accuracy	F1-score (micro)	F1-score (Macro)
Train	0.826601	0.826601	0.788229
Test	0.762946	0.762946	0.685015

The accuracy and F1-score of the results are shown in Table 2, and the ROC curve is shown in Figure 9.

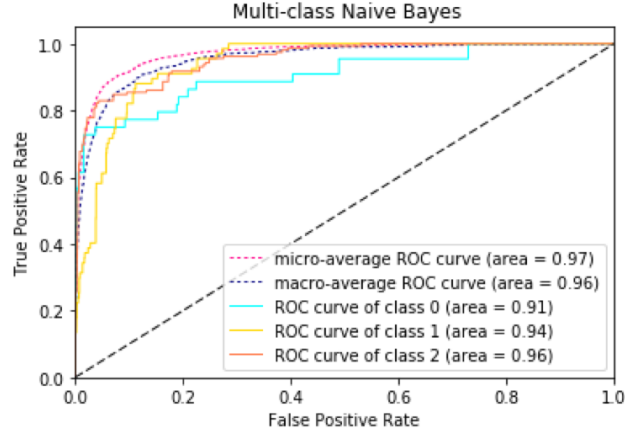


Figure 9: ROC curve of the results using Naive Bayes algorithm

3.3 LINEAR SUPPORT VECTOR MACHINE

Linear SVM is the fast machine learning algorithm for solving multi-class classification problems from ultra large data sets that implements an original proprietary version of a cutting plane algorithm for designing a linear support vector machine. LinearSVM is a linearly scalable routine meaning that it creates an SVM model in a CPU time which scales linearly with the size of the training data set.

In our project, we use Linear SVM (LSVM) for multi-class classification. We used “LinearSVC()” from the Python package *sklearn.svm* to build multi-class classifiers. The most import part we did in random forest is parameter tuning.

- C : Penalty parameter of the error term, default is 1. The results are shown in Table 3.

Table 3: Accuracy of LSVM w.r.t. C

	Train	Test
0.5	0.922187	0.77382
1.0	0.933123	0.764395
2.0	0.940383	0.752325

- Loss: “hinge” or “squared_hinge”, the results are shown in Table 4.

Table 4: Accuracy of LSVM w.r.t. different loss function

Loss	Train	Test
Hinge	0.908516	0.772567
Squared Hinge	0.933123	0.764395

As for the overall ROC, please refer to Figure 10.

3.4 LATENT DIRICHLET ALLOCATION

The Latent Dirichlet allocation(LDA) is an example of a topic model. Each document may be viewed as a mixture of various topics where each document is considered to have a set of topics that are assigned to it via LDA. Specifically, in our experiment, we set each cuisine is a document and the ingredients inside of each cuisine regarded as words in the document. By using the LDA, we can get the probability for document belonging to each topics. Thus we can make a decision about which type of cuisine this dish is. And simultaneously, we can get the main topic words for each topic, which contains the highest percent relevance word to the related topic.

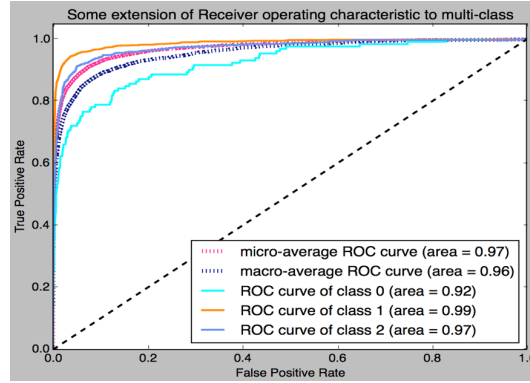


Figure 10: ROC curve of the results using Linear SVM

We use Latent Dirichlet Allocation (LDA) to learn the latent topics from the recipes. Although LDA is an unsupervised algorithm, it still reveals some interesting findings. We pick four of the learned topics and show the five most probable words in Table 5. From the table, we can see that, Topic 1 is most probable to contain different *tortilla* dishes, since there is an key ingredient of *flour tortillas*. Topic 11 may be concerned with cakes, since it has the key ingredients *all-purpose flour* and *baking powder*. Topic 14 may also be about dessert, since most of its ingredients can be used in making desserts. As for topic 11, we cannot make a reasonable guess based on the five ingredients, but among the 20 most probable ingredients, there are *orange juice*, *lime* and *ice cubes*, so we can guess that it may be a topic about drinks.

Table 5: Most probable words of four topics

Topic 0	sour cream	chili powder	salsa	black beans	flour tortillas
Topic 5	salt	all-purpose flour	baking powder	sugar	large eggs
Topic 11	sugar	water	ground cinnamon	honey	raisins
Topic 14	butter	salt	eggs	sugar	milk

3.5 NEURAL NETWORK

The last algorithm we use is neural network (NN), and its structure is shown in Figure 11. This is a neural network with three fully-connected hidden layers, and the activation function applied to the first three hidden layers is Rectify Linear Unit (ReLU) while soft-max function is used for the final output layer that generates the probability distribution on all categories of the input sample. We also attach a batch normalization (BN) layer followed by a dropout (DP) layer after each hidden layer.

The neural network with dropout and batch-normalization works best among all algorithms we use, and its results are shown in Figure 12. The test accuracy is **78.9%**, while the training accuracy is **98.7%**.

We also test on neural network without batch-normalization or dropout, and the results are shown in Figure 14 and Figure 15 respectively. As we can see in Figure 14, the test accuracy is **74.1%**, and the curves are smoother than the curve in Figure 12. If we train the model with dropout but without batch-normalization for another 400 epochs, the test accuracy can be as high as the model in Figure 12. For Figure 15, the test accuracy is **75.5%**, and the curves are steeper than Figure 14. We also run the model without dropout nor batch-normalization, and the results are compared in Table 6.

Table 6 shows that, if the neural network does not use batch-normalization, it will take a much longer time for the model to converge; if the neural network does not have dropout, the neural network will be more likely to over-fit the training data, as we can see that the training accuracies are very high while the testing accuracies are relatively lower in the last two rows than the previous rows. This shed lights on two things:

- Batch-normalization makes the model converge faster.

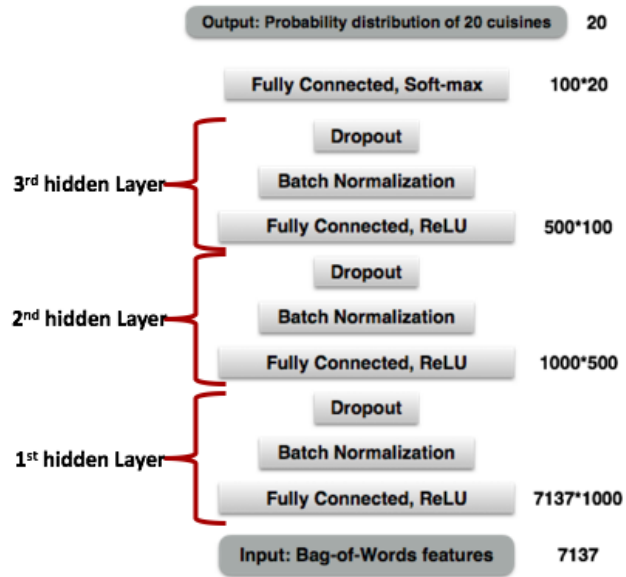


Figure 11: Structure of the neural network used

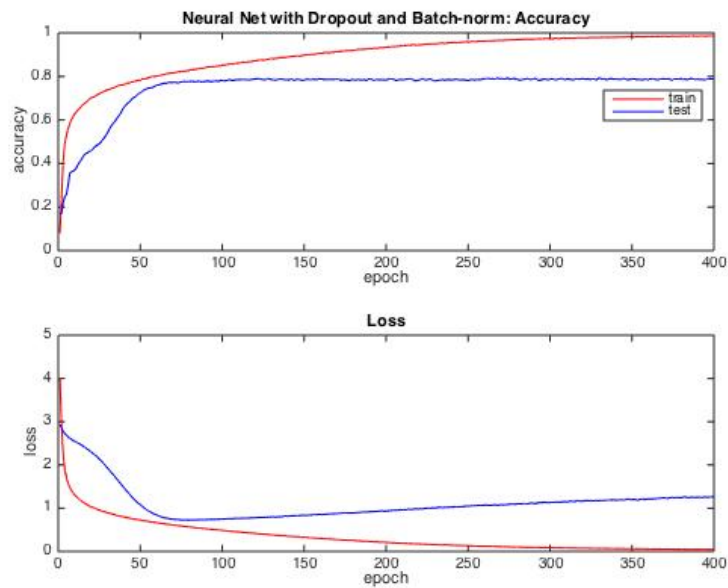


Figure 12: Results of NN with BN and DP

Table 6: Accuracies of different neural network models

BN	DP	Train	Test
Yes	Yes	0.987010	0.789090
No	Yes	0.780003	0.741327
Yes	No	0.999609	0.754651
No	No	0.999497	0.746858

- Dropout reduce the over-fitting of the model.

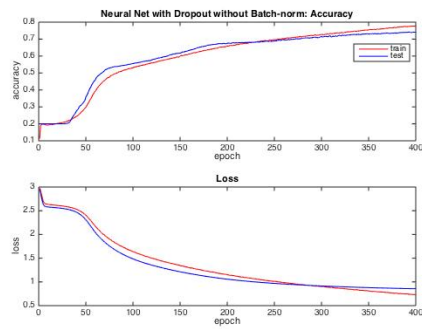


Figure 14: NN with DP but without BN

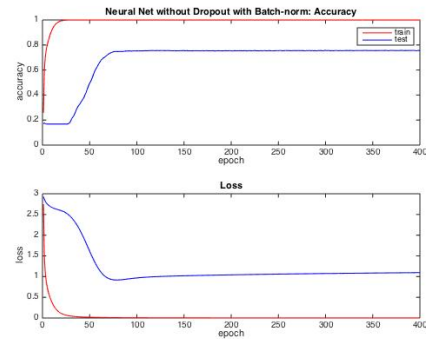


Figure 15: NN with BN but without DP

4 CONCLUSIONS

In this project, we, as foodies learning machine learning, explore several machine learning techniques and try to predict the category of cuisine based on recipes' ingredients. We implement Random Forest, Naive Bayes, Linear SVM and Neural Network, and find that neural network works best among those methods. We try different ways to tune our model, and learn the effects of different parameters or factors have on the prediction performance. We also apply Latent Dirichlet Allocation to cluster the cuisines, use different techniques to visualize the data, and discover some interesting things in the data. This has been a very fruitful practice of machine learning that we can apply what we have learned in class to real-world (delicious) problem.

"I always thought something was fundamentally wrong with the universe"², but what's wrong? Well, the problem is, why is the universe so empty instead of full of food? It is said that "a man is what he eats as much as what he reads". Lets eat the universe so that we can become master of the universe.



Figure 15: The Universe

ACKNOWLEDGMENTS

We want to thank professor Brian Ziebart for his excellent teaching in CS412, and we give our greatest love to all the food in the universe.

²The Hitchhiker's Guide to the Galaxy