Of recipes and ingredients

Redouane Elghazi
redouane.elghazi@epfl.ch

Enguerrand Prebet
enguerrand.prebet@epfl.ch

Pierre Mahmoud–Lamy
pierre.mahmoud-lamy@epfl.ch

Abstract

Can recipes be grouped into meaningful classes in terms of ingredients? Are these groups characterized by their distributions in carbs,fat and proteins?

1 Database Presentation

We gathered the data from Robert West et al. (2013). This data contains recipes from several cooking websites. A recipe entry has informations such as the calories (fat, proteins, carb) and the list of the ingredients.

To only extract the ingredients from the list, we used an external data base from openfood.schoolofdata.ch as a starting point. It contains a list ingredients but also a groups for these ingredients such as "vegetables" or "cereals".

We had to append words that were not in the database but that appear in recipes, especially for cheese names, pasta names and sauce names.

2 Preprocessing the database

Each recipe was represented as one long string. To only extract ingredients as opposed to quantities and others words that may appear in a recipe, we considered our external database with words sorted by decreasing length. For each ingredient, we then processed through the recipes and checked if it appeared in the string of ingredients. In that case, we added it to our list of ingredients and deleted it from the string.

By sorting and deleting ingredients, we ensure that combination of words like "black pepper" is found without also finding "pepper". Some words were still matched incorrectly, with "tea" found in recipes because of the use of "teaspoon", and "gin" found in "virgin" for instance. So before looking for our ingredients, we preprocessed the recipe by removing those words first. We also took into consideration plural and separators.

When looking at the nutritional informations, we saw that some recipes are too energetic to be normalized per serving, being over the recommended amount for one person **per day**. To be able to compare different recipes, we normalized our data by the total energy for the recipe, and thus considered proportion of fat, carb and proteins instead of their raw value.

3 Clustering

To reorganize our different recipes into groups according to their ingredients, we need a decent representation. Our choice was to consider boolean value for each food group if a recipe contains an ingredients from that food group. This gives us a N-dimensional vector with value in $\{0,1\}^N$, with N being the number of food group (11 in our case).

We use a K-means clustering algorithm to create our recipe groups. Starting from K random means, we assign each recipe to the closest mean (the distance being defined as the sum of differences over the N elements). Then each mean is updated to correspond to the effective mean of the elements assigned to this mean.

In practice, though recipes get values in $\{0,1\}^N$, the means have real values in $[0,1]^N$ to get the best possible clustering.

Moreover, after having tested different values of K to study the convergence of the groups, we set the option to fix the initial means, for reproducibility.

4 First results

After several tests, the more meaningful number of groups seems to be 4. By meaningful, we mean that we can visualize what the groups represent. Recall that the output of the algorithm is the assignement vector for each recipe, and the mean vectors. We then need to interpret the meaning of those means:

- the first group represents desserts based on complex preparations, mostly with eggs and flour or other cereal base, such as pies, cakes, muffins,...
- the second group represents more simple desserts, with fewer ingredients and more fruit-based, such as fruit salads, baked apples,... but also chocolate mousse,...
- the third group consists in complex prepared meals, with meat and vegetables. It uses a lot of different food groups (eggs, dairy, fats) and can for example be stews, boeuf bourguignon,...
- the last group is easy-to-do dishes with meat and vegetables. It uses more condiments (mustard, sauces,...) and can be bolognese pasta, hot-dogs,...

The split of the groups can be considered using 2 questions:

- Is this a main course ?
- Is it simple ?

5 Classification by the nutrients

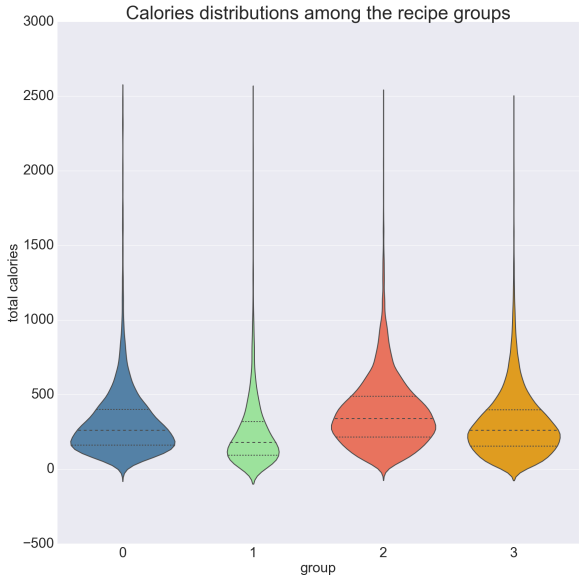

Figure 1

We see that more complex recipes (0,2) have more calories than their simpler counterparts (respectively 1,3).

Moreover, quite surprisingly, main courses recipes (2,3) are in average more calorific than dessert recipes (0,1). They also have more recipes with very high calories, compared to dessert recipes.

This could be linked to desserts often considered full of fats and sugar, thus pushing people to develop and publish a lot of low-fat, low-calories dessert recipes in the 2010's. It also comes with the fact that there are a lot of subtitutes for sugar, milk and eggs.

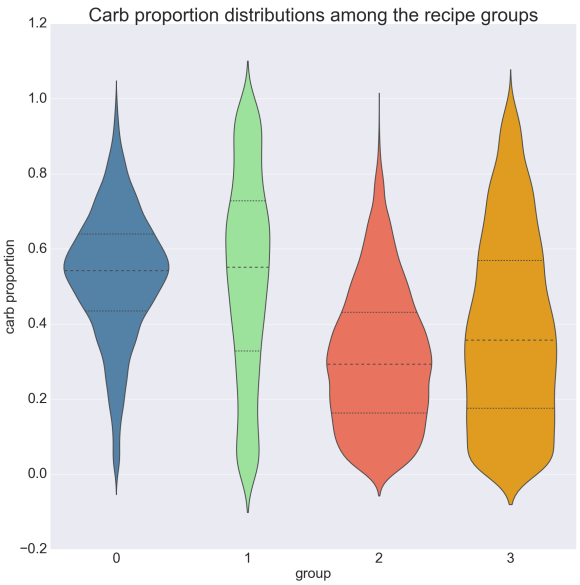

Figure 2

We observe very different distributions of the carb proportion:

group 0 is pretty gathered around its mean, which is quite high. It seems complex desserts have high rates of carbs. group 1 is quite regular: there is almost a uniform distribution of carb proportions. Thus simple desserts are not characterized by their carb rate. group 2 and 3 have pretty low carb rates compared to the desserts groups. However simple main courses (3) can get to a high rate of carbs.

For the proteins distribution, we can observe a great distinction between the desserts (0,1) and the main courses (2,3). The desserts have very low rates of proteins, and very centered around the mean.

The proteins proportion is sparse for the 2nd group, and even more spread for the 3rd group.

The fat proportions are quite different among groups too. For group 1, the distribution is relatively uniform, a bit "squeezed" aroun the first quartile.

The distribution for group 0 is symmetric

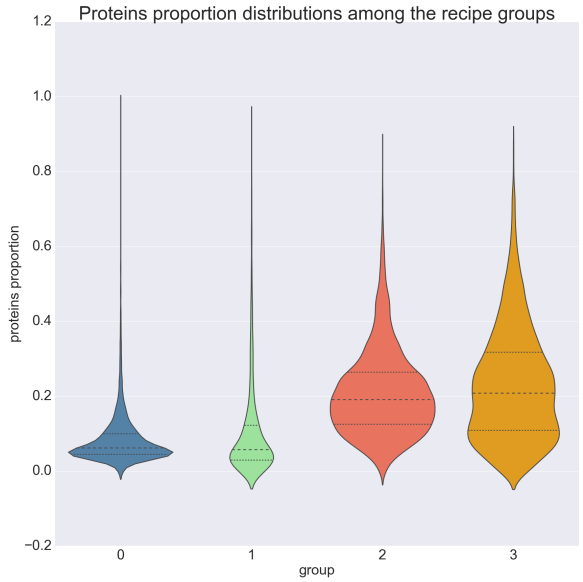

Figure 3

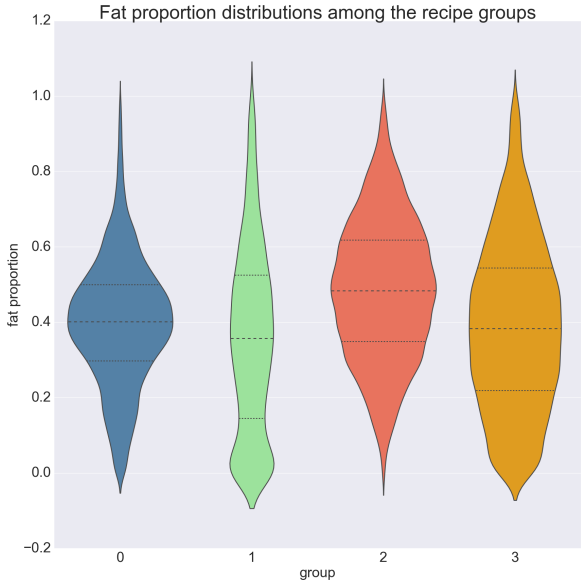

Figure 4

around the mean at 0.4, whereas distributions for groups 2,3 are more spread.

We also observe that complex recipes (0,2) have in average a higher fat rate than more simpler recipes (1,3).

To test how well the nutrients describe the four groups, we use logistic regressions to obtain propensity scores for our 2 questions.

Our results are shown in FIGURE 5. We notice mainly two things:

- the distinction main course/ dessert is pretty clear: groups 0 and 1 are both clearly identified as desserts. For groups 2 and 3, there is a residual tail but the main part is identified as main course
- however the distinction between simple and complex recipes is not as correct. We see recipes are mostly considered simple, even if group 3 is not that far of being identified. Group 1 is quite out of the target zone.

6 Conclusions

From our analysis, we can answer our original questions. It clearly appeared that the recipes could be grouped into meaningful classes just using their ingredients. As for the characterization of these classes regarding their distributions in carbs, fat, and protein, the results are less sharp: among the 4 classes that we could create, only 2 seem to be discriminated. More precisely, we cannot distinguish a simple recipe from a complex one by just looking at the proportion of carbs, fat, and protein. Still we managed to distinguish desserts from main courses.

References

[Robert West et al.2013] Robert West, Ryen W. White, and Eric Horvitz. From cookies to cooks: Insights on dietary patterns via analysis of web usage logs. In *Proceedings of the 22nd International World Wide Web Conference*, 2013.

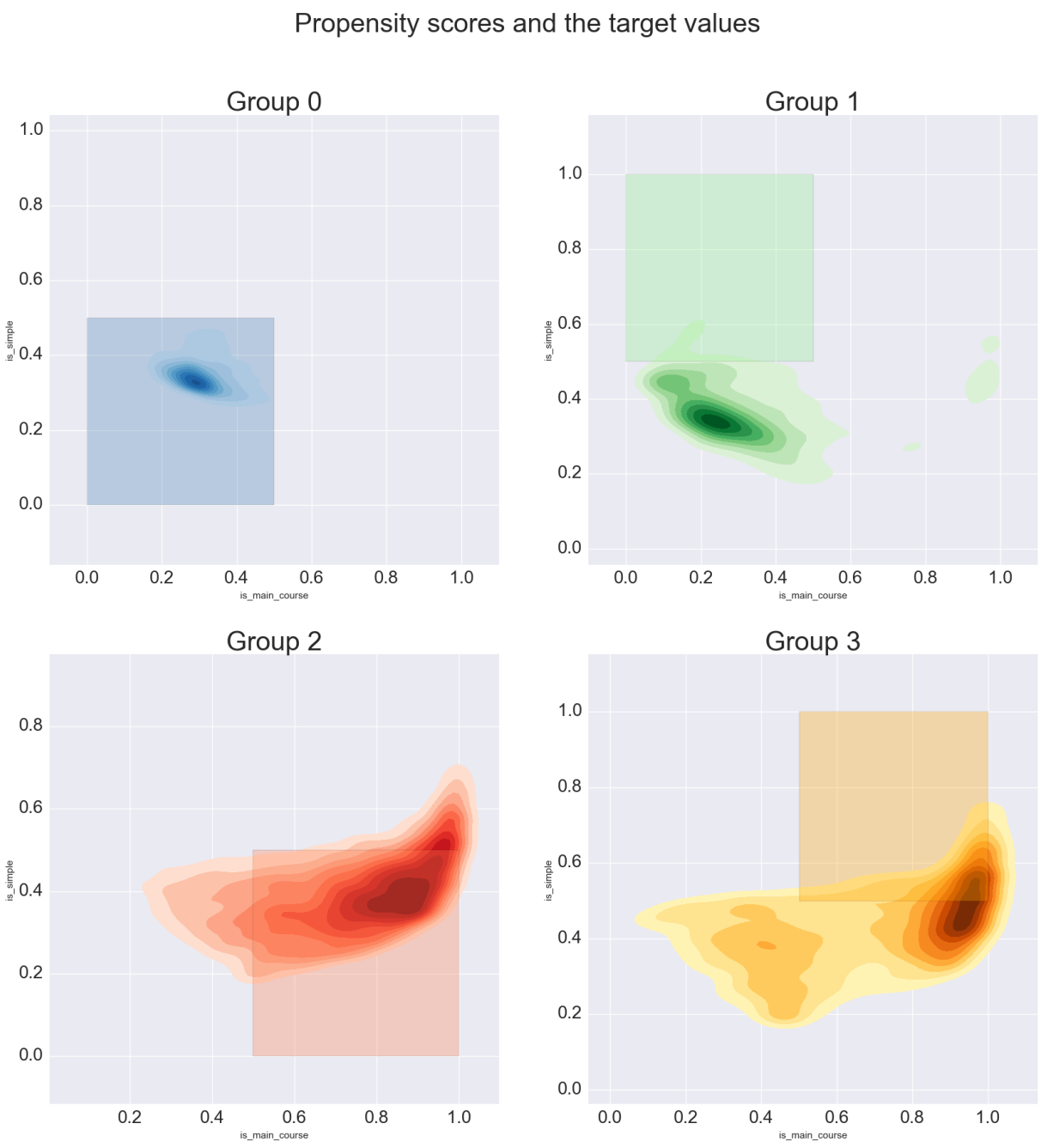
Propensity scores and the target values
Group 0
Group 1
Group 2
Group 3
is_simple
is_main_course
0.0
0.2
0.4
0.6
0.8
1.0

Figure 5