DS0103EN-1-1-1-From-Problem-to-Approach-v1.0

December 19, 2018

From Problem to Approach

0.1 Introduction

The aim of these labs is to reinforce the concepts that we discuss in each module's videos. These labs will revolve around the use case of food recipes, and together, we will walk through the process that data scientists usually follow when trying to solve a problem. Let's get started!

In this lab, we will start learning about the data science methodology, and focus on the **Business Understanding** and the **Analytic Approach** stages.

0.2 Table of Contents

1 Business Understanding

This is the **Data Science Methodology**, a flowchart that begins with business understanding.

Why is the business understanding stage important? Your Answer: To know and understand the requirements and goal(s).

Double-click **here** for the solution.

Looking at this diagram, we immediately spot two outstanding features of the data science methodology.

What are they? Your Answer: 1. The flowchart is highly iterative. 2. The flowchart never ends. Double-click **here** for the solution.

Now let's illustrate the data science methodology with a case study. Say, we are interested in automating the process of figuring out the cuisine of a given dish or recipe. Let's apply the business understanding stage to this problem.

Q. Can we predict the cuisine of a given dish using the name of the dish only? Your Answer: No you can't.

Double-click **here** for the solution.

Q. For example, the following dish names were taken from the menu of a local restaurant in Toronto, Ontario in Canada.

1. Beast

2. 2 PM

3. 4 Minute

Are you able to tell the cuisine of these dishes? Your Answer:NO

Double-click **here** for the solution.

Q. What about by appearance only? Yes or No. Your Answer: NO

Double-click **here** for the solution.

At this point, we realize that automating the process of determining the cuisine of a given dish is not a straightforward problem as we need to come up with a way that is very robust to the many cuisines and their variations.

Q. What about determining the cuisine of a dish based on its ingredients? Your Answer: Yes

Double-click **here** for the solution.

As you guessed, yes determining the cuisine of a given dish based on its ingredients seems like a viable solution as some ingredients are unique to cuisines. For example:

- When we talk about **American** cuisines, the first ingredient that comes to one's mind (or at least to my mind =D) is beef or turkey.
- When we talk about **British** cuisines, the first ingredient that comes to one's mind is haddock or mint sauce.
- When we talk about **Canadian** cuisines, the first ingredient that comes to one's mind is bacon or poutine.
- When we talk about **French** cuisines, the first ingredient that comes to one's mind is bread or butter.
- When we talk about **Italian** cuisines, the first ingredient that comes to one's mind is tomato or ricotta.
- When we talk about **Japanese** cuisines, the first ingredient that comes to one's mind is seaweed or soy sauce.
- When we talk about **Chinese** cuisines, the first ingredient that comes to one's mind is ginger or garlic.
- When we talk about **indian** cuisines, the first ingredient that comes to one's mind is masala or chillis.

Accordingly, can you determine the cuisine of the dish associated with the following list of ingredients? Your Answer:

yes Double-click **here** for the solution.

2 Analytic Approach

So why are we interested in data science? Once the business problem has been clearly stated, the data scientist can define the analytic approach to solve the problem. This step entails expressing the problem in the context of statistical and machine-learning techniques, so that the entity or stakeholders with the problem can identify the most suitable techniques for the desired outcome.

Why is the analytic approach stage important? Your Answer:

Because it helps identify what type of patterns will be needed to address the question most effectively.

Double-click **here** for the solution.

Let's explore a machine learning algorithm, decision trees, and see if it is the right technique to automate the process of identifying the cuisine of a given dish or recipe while simultaneously providing us with some insight on why a given recipe is believed to belong to a certain type of cuisine. This is a decision tree that a naive person might create manually. Starting at the top with all the recipes for all the cuisines in the world, if a recipe contains **rice**, then this decision tree would classify it as a **Japanese** cuisine. Otherwise, it would be classified as not a **Japanese** cuisine.

Is this a good decision tree? Yes or No, and why? Your Answer: NO!

Double-click **here** for the solution.

In order to build a very powerful decision tree for the recipe case study, let's take some time to learn more about decision trees.

- Decision trees are built using recursive partitioning to classify the data.
- When partitioning the data, decision trees use the most predictive feature (ingredient in this case) to split the data.
- **Predictiveness** is based on decrease in entropy - gain in information, or *impurity*.

Suppose that our data is comprised of green triangles and red circles. The following decision tree would be considered the optimal model for classifying the data into a node for green triangles and a node for red circles.

Each of the classes in the leaf nodes are completely pure – that is, each leaf node only contains datapoints that belong to the same class.

On the other hand, the following decision tree is an example of the worst-case scenario that the model could output.

Each leaf node contains datapoints belonging to the two classes resulting in many datapoints ultimately being misclassified.

A tree stops growing at a node when:

- Pure or nearly pure.
- No remaining variables on which to further subset the data.
- The tree has grown to a preselected size limit.

Here are some characteristics of decision trees: Now let's put what we learned about decision trees to use. Let's try and build a much better version of the decision tree for our recipe problem.

I hope you agree that the above decision tree is a much better version than the previous one. Although we are still using **Rice** as the ingredient in the first *decision node*, recipes get divided into **Asian Food** and **Non-Asian Food**. **Asian Food** is then further divided into **Japanese** and **Not Japanese** based on the **Wasabi** ingredient. This process of splitting *leaf nodes* continues until each *leaf node* is pure, i.e., containing recipes belonging to only one cuisine.

Accordingly, decision trees is a suitable technique or algorithm for our recipe case study.

2.0.1 Thank you for completing this lab!

This notebook was created by Alex Aklson. I hope you found this lab session interesting. Feel free to contact me if you have any questions!

This notebook is part of a course on **Coursera** called *Data Science Methodology*. If you accessed this notebook outside the course, you can take this course, online by clicking here.