CAP5625 - Introduction to AI

Project 2: Bayes' Theorem and Bayesian Network

For this project, you will use the Recipe1M dataset that contains around 1M recipes. Each recipe has the following structure. You need to use only the "ingredients" and "title" information. The dataset is provided in the files section on Canvas.

```
{
  "id",
  "instructions",
  "tokenized",
  "ingredients",
  "states",
  "images",
  "title"
}
```

Given the dish type and a few ingredients from a recipe, your task is to find another ingredient that is most likely to be used in that recipe. To do that, you have to use two methods separately: Bayesian Network and Bayes's Theorem.

Sample Input:

Dish type: salad,

Number of ingredients provided: 3

Ingredients: cucumber, onion, pepper

Sample Output:

Salt

Method 1: Bayesian Network

We have created a network by selecting 3 dish types and 10 ingredients from Recipe1M. It shows dish types-ingredients and ingredient-ingredient relationships. You have to find the most likely ingredient using the network given below.

Let's assume,

d = given dish type,

$\{x_1, x_2, \ldots, x_n\}$ = given ingredients

Steps:

1. Store the network in a data structure that you find suitable.
2. For each node, create its Conditional Probability Table (CPT).

Part 1 includes only the above two steps, everything else are included in Part 2

~~3. For each ingredient *t* in this network, find $P(t \mid d, \{x_1, x_2, \ldots, x_n\})$ using CPT.~~

~~4. Choose the ingredient that gives maximum probability.~~

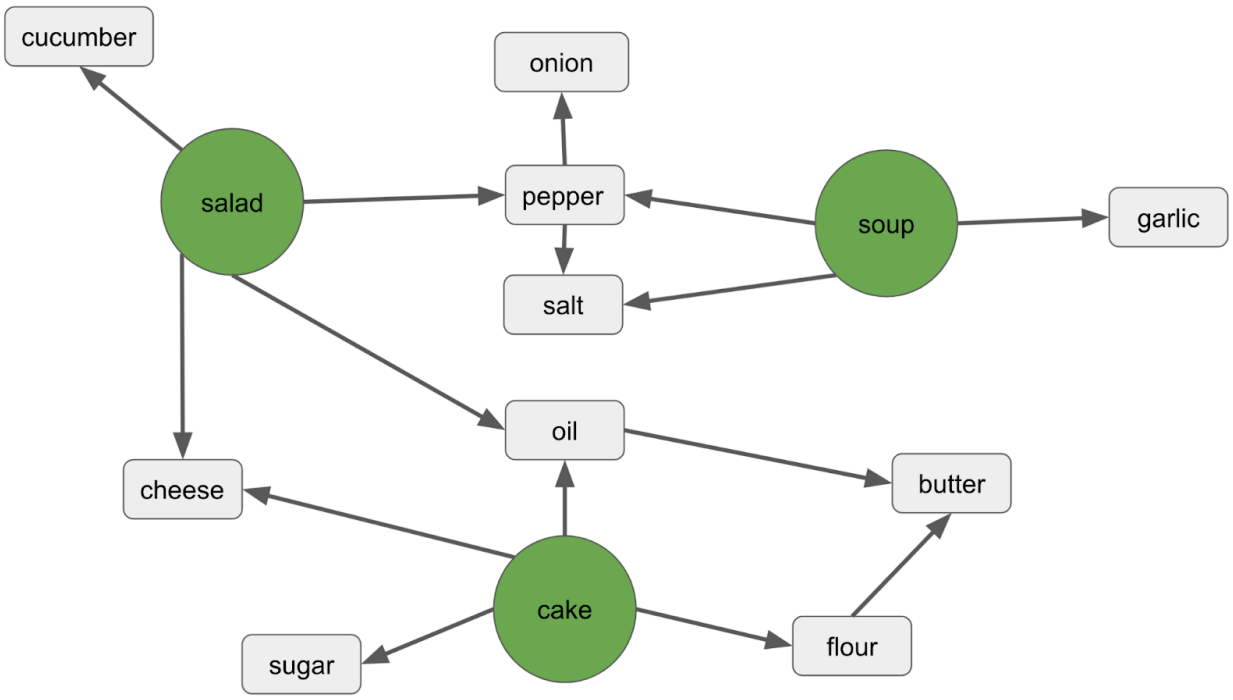

Figure: A Bayesian Network with 3 dish types and 10 ingredients

A sample data structure is given below for your implementation help. This is just an example with the CPT values randomly assigned for different combinations of True-False scenarios. To generate the CPT values, use the frequency based formula. For example:

$$P(+\text{garlic} \mid +\text{soup}) = \frac{\#\ soup\ recipes\ that\ contains\ garlic}{\#\ soup\ recipes}$$

```
{
   "ingredients": ["oil", "salt", "cheese", "onion", "pepper", ...],
   "dish type": ["salad", "cake", "soup"],
   "relationship": {
      "salad": {
         "parents": [],
         "cpt": [0.34, 0.66]
      },
      "salt": {
         "parents": ["soup", "pepper"],
         "cpt":   [0.25, 0.13, 0.72, 0.29, 0.16, 0.34, 0.55, 0.23]
      }
   }
}
```

Sample data structure to represent the Bayesian Network

Method 2: Bayes' Theorem

Steps:

Let's assume,

d = given dish type,

$\{x_1, x_2, \ldots, x_n\}$ = given ingredients

1. Find all recipes that have the given dish type in its title. Let R be the set of those recipes. Use R to do the following steps.
2. For each ingredient t in R, compute:

$$P(t \mid d, \{x_1, x_2, \ldots, x_n\}) = P(d, \{x_1, x_2, \ldots, x_n\} \mid t)\, P(t)$$

$$P(t) = \frac{\#\ recipes\ in\ R\ with\ t}{\#\ recipes\ in\ R}$$

$$P(d, \{x_1, x_2, \ldots, x_n\} \mid t) = \frac{\#\ recipes\ in\ R\ with\ \{x_1, x_2, \ldots, x_n, t\}}{\#\ recipes\ in\ R\ with\ t}$$

3. Choose the ingredient *t* that gives maximum $P(t \mid d, \{x_1, x_2, \ldots, x_n\})$.

Submission:

You can use any programming language for this project. Create a zip file including everything (**except the Recipe1M dataset**) that are required to run your program. Name the zip file with your UID (e.g. U13611582.zip). Submit it on Canvas.

You can use any method to take the input such as: file, console etc. Describe how to run your program in a readme file.

No report is required for this project.

Deadlines:

Part 1: Calculate Conditional Probability Tables

Deadline: Tuesday Oct 26, 2021

Weight: 50% of your final project grade

Part 2: Implement Bayesian Network and Bayes Theorem

Deadline: Friday Nov 5, 2021

Weight: 50% of your final project grade