Vagueness, uncertainty and degrees of belief

In some situations, it is not suitable to express general statements using logical universals. We have seen different ways to represent defaults, but sometimes other extensions for knowledge representation are necessary.
Besides the intrinsec imprecision of statements like "someone is somewhat tall", the way the conclusions are formulated may also be imprecise (e.g. in medical field, a rule may not be applicable in 100% of cases).

Noncategorical reasoning

We distinguish three ways to "relax" the categorical nature of classical logic, in order to make the universal $\forall x\, P(x)$ more flexible:

1. Relaxation of the strength of the quantifier - instead of "for all x" we say "for % of x".
   "95% of the persons in this group are master students"
   This is a statistical interpretation. The use of probability in these sentences is objective (it is an assertion about the frequency of an event, it is not a subjective interpretation or a degree of confidence).
2. Relaxation of the applicability of a predicate - instead of a statement like "Everyone in my group is (absolutely) tall", we say "Everyone in my group is (moderately) tall".
   The predicate "tall" applies to an individual to a greater or lesser extent. These predicates are called vague.
   A person may be considered both tall (moderately) and short (weakly) at the same time.

3. Relaxation of the degree of belief in a sentence – instead of having the statement "Everyone in this group is a master student" we say "I believe that everyone in this group is a master student, but I am not sure". This is uncertain knowledge and it can be quantified by using the concept of subjective probabilities.

Objective probability

It refers to the frequency of a single event happening and it does not depend on who is assessing the probability. It is best applied to situations like "coin flipping" or "card drawing".

Subjective probability

The degree of confidence (also called subjective probability) in a sentence is separable from the content of the sentence. The degree of belief in a sentence can vary, regardless of how vague or categorical the sentence may be. For example, we may be absolutely certain that Bill is quite tall, while we may only suspect that he is married.

With subjective beliefs, we express degrees of confidence rather than black-and-white conclusions.

Subjective probabilities can be mechanically computed like the objective ones, but they are used in a different way. We are interested in how evidences combine to change our degree of confidence in a belief, rather than simply deriving new conclusions.

Def The prior probability of a sentence α involves the prior state of information (or background knowledge) β.
We write it $Pr(\alpha|\beta)$.

For example, suppose that we know that 0.2% of the population has hepatitis. Based on just that, our degree of belief that John (a randomly chosen individual) has hepatitis is 0.002.

Def A posterior probability is derived when new evidence is considered: $P(\alpha \mid \beta \wedge \gamma)$ where γ is the new evidence.
For example, if John is yellowish, given the symptoms and the prior probability, we may conclude that the posterior probability of John having hepatitis is 0.65.

The key problem is how to combine evidences from different sources to reevaluate our beliefs.

A basic Bayesian approach

Suppose that we have a number of atomic sentences of interest $p_1, \ldots, p_n$ (e.g. Eric is tall, Ane is married, George is a teacher and so on). In different interpretations, different combinations of these sentences will be true. Let J be an interpretation that specifies which sentences are true/false.

Def. The joint probability distribution $\mathcal{J}$ is the specification of the degree of belief for each of the 2^n truth assignments

$$\sum_J \mathcal{J}(J) = 1 \quad \text{and} \quad \mathcal{J}(J) \in [0,1].$$

The degree of belief in any sentence α is defined as

$$P_{\mathcal{J}}(\alpha) = \sum_{J \models \alpha} \mathcal{J}(J).$$

Knowing that $P_{\mathcal{J}}(\alpha \mid \beta) = \dfrac{P_{\mathcal{J}}(\alpha \wedge \beta)}{P_{\mathcal{J}}(\beta)}$, the degree of belief that John is tall given that he is male from California is

$$\frac{P_{\mathcal{J}}(\text{John is tall, John is male, John is from California})}{P_{\mathcal{J}}(\text{John is male, John is from California})}.$$

For n atomic sentences, we need to specify $2^n - 1$ values. This is unachievable for any practical applications.

Belief (or Bayesian) networks

Suppose that we have the atomic sentences $p_1, \ldots, p_n$. We can specify an interpretation using $\langle P_1, \ldots, P_n \rangle$, where each P_i is p_i (when the sentence is true) or $\neg p_i$ (when the sentence is false). We have that

$$J(\langle P_1, \ldots, P_n \rangle) = P_r(P_1 \wedge P_2 \ldots \wedge P_n)$$

because there is only one interpretation that satisfies $P_1 \wedge \ldots \wedge P_n$.

We represent all the variables p_i in a directed acyclic graph, called a belief (or Bayesian) network.

There is an arc from p_i to p_j if the truth of p_i directly affects the truth of p_j (the former is a parent of the letter). We assume that the variables are numbered such that the parents of any p_j appear earlier in the sequence than p_j (we can do that because the graph is acyclic).

According to the chain rule in probabilities, we have:

$$J(\langle P_1, \ldots, P_n \rangle) = P_r(P_1) \cdot P_r(P_2 \mid P_1) \ldots P_r(P_n \mid P_1 \wedge \ldots \wedge P_{n-1}).$$

To compute the joint probability distribution, we still need $2^n - 1$ values because for each term $P_r(P_{j+1} \mid P_1 \wedge \ldots \wedge P_j)$ there are 2^j conditional probabilities to specify and $\sum_{j=0}^{n-1} 2^j = 2^n - 1$.

In order to reason about subjective probabilities, some simplifying assumptions are necessary.

We will assume that each propositional variable in the belief network is conditionally independent from the nonparent variables, given the parent variables.

$$Pr(P_{j+1} \mid P_1 \wedge \ldots \wedge P_j) = Pr(P_{j+1} \mid \text{parents}(P_{j+1}))$$

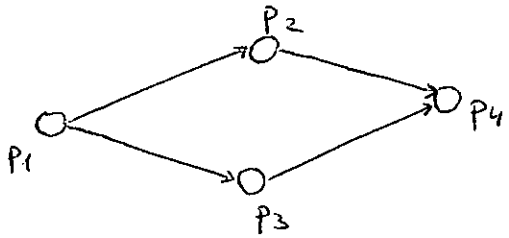
With these independence assumptions, it follows that

$$J(\langle P_1, \ldots, P_n \rangle) = Pr(P_1 \mid \text{parents}(P_1)) \cdot \ldots \cdot Pr(P_n \mid \text{parents}(P_n)).$$

To fully specify J, we need to know $Pr(P \mid \text{parents}(P))$ for each variable p.

If K is the maximum number of parents for any node, then we have no more than $n \cdot 2^K$ values to specify.

For the belief network in the figure below



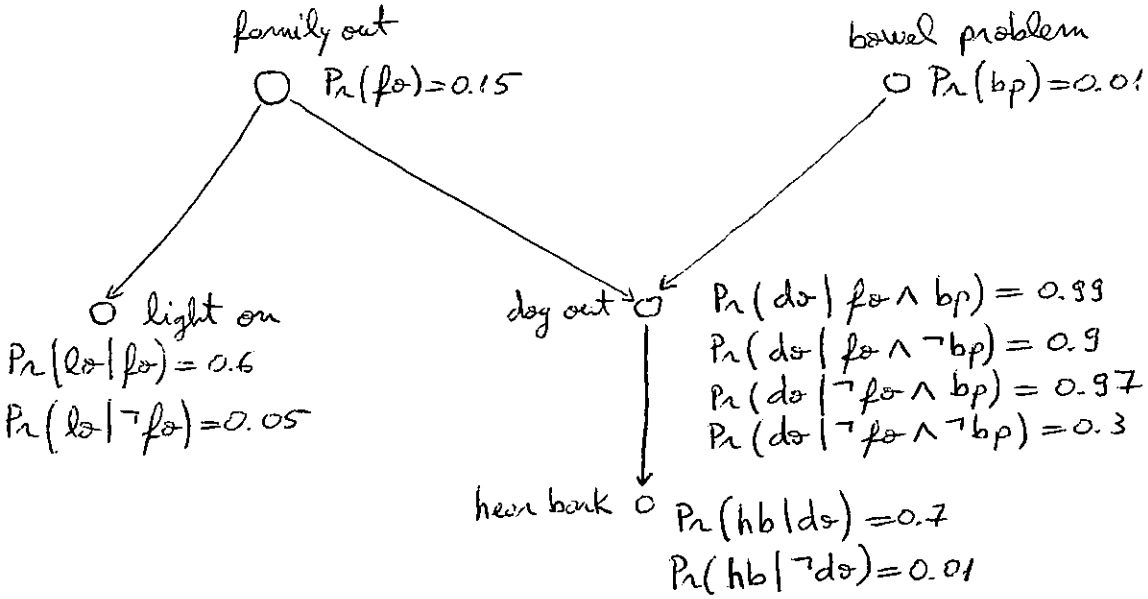
we have $J(\langle P_1, P_2, P_3, P_4 \rangle) = Pr(P_1) \cdot Pr(P_2 \mid P_1) \cdot Pr(P_3 \mid P_1) \cdot Pr(P_4 \mid P_2 \wedge P_3)$.

We now need $(1+2+2+4) = 9$ values rather than 15 (without the independence assumption) to compute J.

An example (due to Eugene Charniak)

We have a family with a dog. We usually put the dog out (do) when the family is out (fo). We also put the dog out when it has bowel problem (bp). A reasonable proportion of time when the dog is out, you can hear it barking (hb) when you approach the house. We usually leave the light on (lo) outside the house when the family is out.

Using these facts, we can construct the following belief network:



We assume the following about the joint probability distribution:

$$J(\langle FO, LO, BP, DO, HB\rangle) = Pr(FO) \cdot Pr(LO|FO) \cdot Pr(BP) \cdot Pr(DO|FO \wedge BP) \cdot Pr(HB|DO).$$

We need $1+2+1+4+2 = 10$ values to specify the joint probability distribution.

Using this belief network, we want to calculate the probability that the family is out, given that the light is on and we don't hear barking.

$$Pr(fo|lo \wedge \neg hb) = \frac{Pr(fo \wedge lo \wedge \neg hb)}{Pr(lo \wedge \neg hb)} = \frac{\sum_{BP,DO} J(\langle fo, lo, BP, DO, \neg hb\rangle)}{\sum_{FO,BP,DO} J(\langle FO, lo, BP, DO, \neg hb\rangle)}$$

1. $J(\langle fo, lo, bp, do, \neg hb\rangle) = Pr(fo) \cdot Pr(lo|fo) \cdot Pr(bp) \cdot Pr(do|fo \wedge bp) \cdot (1 - Pr(hb|do))$
$= 0.15 \cdot 0.6 \cdot 0.01 \cdot 0.99 \cdot 0.3$
2. $J(\langle fo, lo, bp, \neg do, \neg hb\rangle) = 0.15 \cdot 0.6 \cdot 0.01 \cdot 0.01 \cdot 0.99$
3. $J(\langle fo, lo, \neg bp, do, \neg hb\rangle) = 0.15 \cdot 0.6 \cdot 0.99 \cdot 0.9 \cdot 0.3$
4. $J(\langle fo, lo, \neg bp, \neg do, \neg hb\rangle) = 0.15 \cdot 0.6 \cdot 0.99 \cdot 0.1 \cdot 0.99$

5. $J(\langle \neg fo, lo, bp, do, \neg hb \rangle) = 0.85 \cdot 0.05 \cdot 0.01 \cdot 0.97 \cdot 0.3$

6. $J(\langle \neg fo, lo, bp, \neg do, \neg hb \rangle) = 0.85 \cdot 0.05 \cdot 0.01 \cdot 0.03 \cdot 0.99$

7. $J(\langle \neg fo, lo, \neg bp, do, \neg hb \rangle) = 0.85 \cdot 0.05 \cdot 0.99 \cdot 0.3 \cdot 0.3$

8. $J(\langle \neg fo, lo, \neg bp, \neg do, \neg hb \rangle) = 0.85 \cdot 0.05 \cdot 0.99 \cdot 0.7 \cdot 0.99$

So, $Pr(fo \mid lo \wedge \neg hb) = \dfrac{1. + 2. + 3. + 4.}{1. + 2. + 3. + 4. + 5. + 6. + 7. + 8.}$

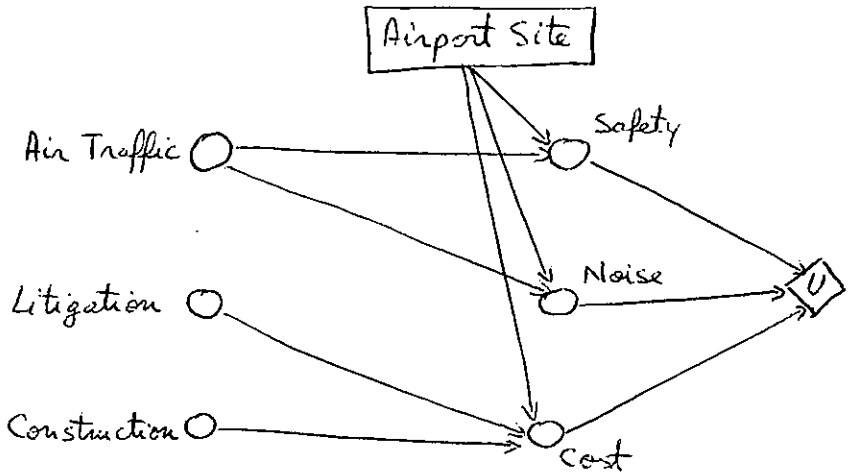
Decision networks (influence diagrams)

They are general decision mechanisms that combine Bayesian networks with additional node types for actions and utilities.

A decision network represents information about an agent's current state, its possible actions, the resulting state from the agent's action and the utility (value) of that state.

There are three types of nodes:

- Chance nodes (as circles) represent probabilistic variables, just like they do in Bayesian networks.
  In decision networks, the parent nodes can include chance nodes as well as decision nodes.
- Decision nodes (as rectangles) represent decisions made by the agent. In the following figure, the AirportSite action can take on a different value for each site under consideration.
  The choice influences the cost, safety and noise that will result.
- Utility (value) node (as diamond) – represent the agent's utility function – there is only one such a node.
  It has parents all variables describing the outcome that directly affect utility. The utility is expressed as a function of the parents attributes.



Evaluating decision networks

For each possible value of the decision node, the resulting utility is calculated. The action (decision) with the highest utility (value) will be chosen.
The algorithm is the following:

1. set the evidence variables for the current state.
2. for each possible value of the decision node:
   - calculate the posterior probabilities for the parent nodes of the utility node (using, for example, a Bayesian network).
   - calculate the resulting utility for that action.
3. return the action with the highest utility.

Representing ignorance – Dempster-Shafer Theory

It is designed to deal with the distinction between uncertainty and ignorance. Rather than computing the probability of a proposition, it computes a lower and an upper bound on the probability of a proposition.

if we have an unbiased coin, the degree of belief that we get heads if we flip it would be 0.5.
if we have a biased coin, due to lack of information, we may want to say only that the degree of belief lies between some limits within $[0, 1]$.

These limits are called belief and plausibility.
For an unbiased coin, we have 0.5 belief and 0.5 plausibility that the result is heads. For an unknown coin, we have 0 belief that we get heads and 1 plausibility.

Thus, the value of a propositional variable is represented by a range, called the possibility distribution of the variable.

Example - suppose that we have a database with names of people and their believed ages. In the case of complete knowledge, the ages would be values. But if we do not know the exact age, we may specify it by a range.

Mary	[18, 22]
Ana	[20, 24]
George	[35, 40]
David	[27, 33]
Cris	[20, 23]

Given an interval Q, rather than asking if $age(x) \in Q$, it is more natural to ask about the possibility of $age(x) \in Q$.
For example, if $Q = [19, 24]$ then it is possible that $age(Mary) \in Q$; it is not possible that $age(George) \in Q$; and it is certain that $age(Cris) \in Q$.

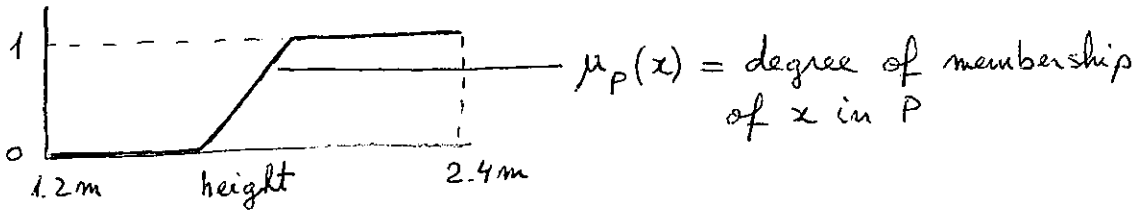
Consider now that we ask what is the probability that the age of a randomly selected individual is in Q?
The belief in this proposition is 2/5 (because of Ana and Cris) and the plausibility is 3/5 (Mary, Ana and Cris). So, the answer is [0.4, 0.6].

The Dempster - Shafer rule combines multiple sources of information with varying levels of knowledge and confidence.

Vagueness

It refers to the degree to which certain predicates are satisfied. For each vague predicate, there is a corresponding base function in terms of which the predicate is understood. (for "tall" the base function is "height").

Def. The degree curve is a function that capture the relationship between a vague predicate and its base function.



An important thing is that an object's degree of satisfaction can be nonzero for multiple predicates over the same base function (e.g. "short" and "tall").

Negation, conjunction and disjunction of vague predicates:

$$\mu_{\neg P} = 1 - \mu_P$$

$$\mu_{P \wedge Q} = \min(\mu_P, \mu_Q)$$

$$\mu_{P \vee Q} = \max(\mu_P, \mu_Q)$$

In a typical application, called fuzzy control, vague predicates are used in production rules.

Unlike standard production systems where a rule either applies or not, here the antecedent of a rule will apply to some degree and the action will be affected to a proportional degree. Such a system enables inferences even when the antecedent conditions are only partially satisfied.

Example – We are given the following rules:

1. If the service is (poor) or the food is (rancid) then the tip is (stingy).
2. If the service is (good) then the tip is (normal).
3. If the service is (excellent) or the food is (delicious) then the tip is (generous).

Assume that service and food quality are described by numbers on a linear scale (e.g. a number from 0 to 10). The amount of tip is represented as a percentage of the cost of the meal (e.g. 10%).

For each of the eight predicates in the example, we are given a degree curve. The base functions are: service, food quality or tip.

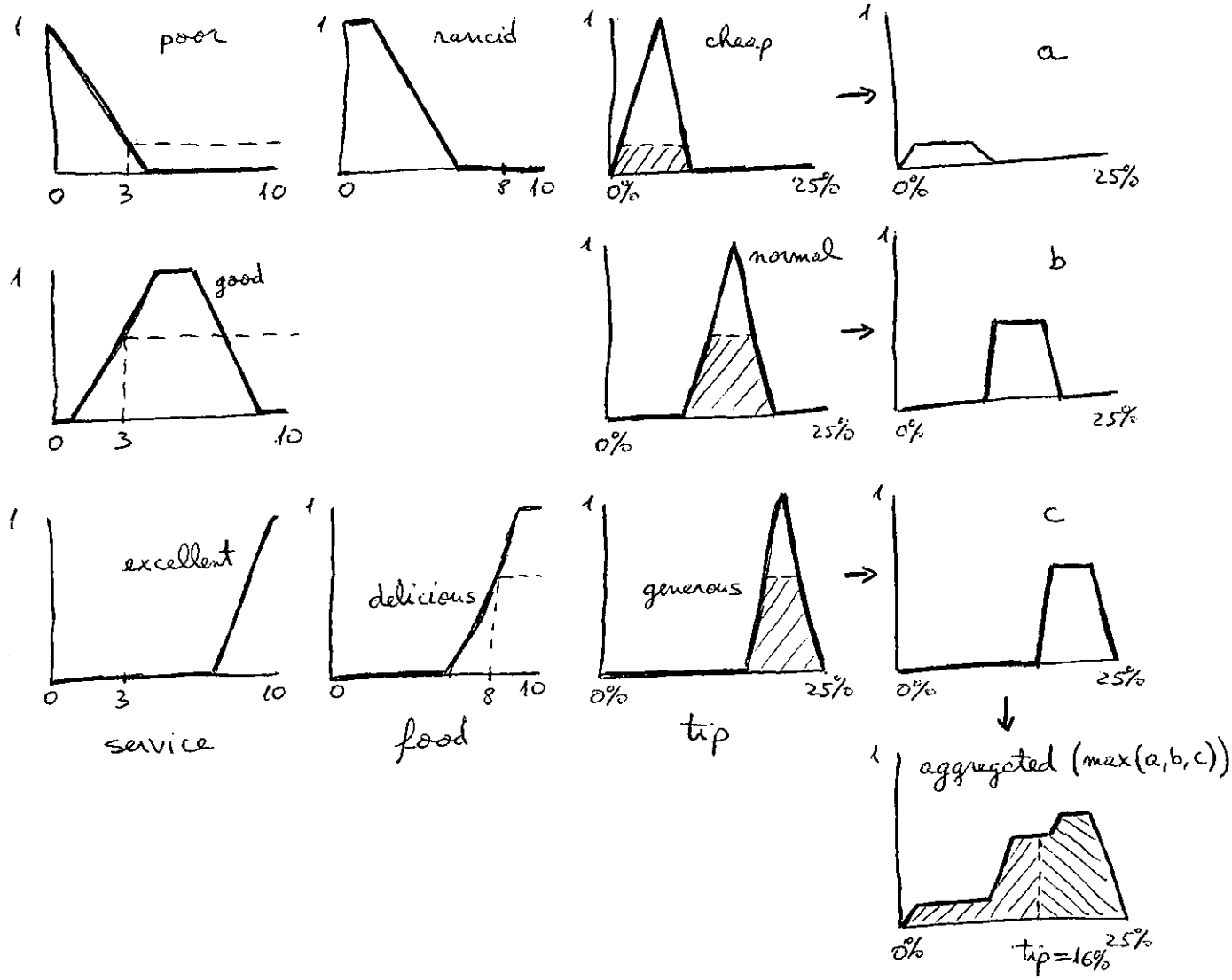
Problem: Given the ratings for the service and for the food, calculate the tip, subject to the rules above.

e.g. service = 3, food = 8, tip = ?

Algorithm

1. transform the inputs into the degrees to which each of the vague predicates used in the antecedents hold.
2. evaluate the antecedents – combine the degrees of applicability of all the predicates in the antecedent of a rule.
3. evaluate the consequents by determining the degrees to which the predicates of each consequent side should be satisfied.
   An intuitive way is that the consequent should hold only to the degree that the rule is applicable.
4. aggregate the consequents - obtain a single degree curve for the "tip" base function.

5. defuzzify the output – generate a value for the tip from the aggregated degree curve at step 4.
One way to do that is to take the center of the area under the curve.



A Bayesian reconstruction

Much of the reasoning with vague predicates can be formulated in terms of subjective probability.
The vague predicates are now treated as ordinary ones, true in some interpretations, false in others.

For a predicate (e.g. Tall), we associate a base measure (e.g. height). We have sentences like Tall(bill) and height(bill) = n, where n is a number.

$$\sum_{n=1.2}^{2.4} Pr(\text{height}(\text{bill}) = n) = 1$$

We reinterpret the "degree of tallness for height of x" as "degree of belief in tallness given the height of x"

$$Pr(\text{Tall}(x) \mid \text{height}(x) = n).$$

If α and β are not independent, we assume that

$$Pr(\alpha \wedge \beta \mid \gamma) = \min\{Pr(\alpha \mid \gamma), Pr(\beta \mid \gamma)\}$$
$$Pr(\alpha \vee \beta \mid \gamma) = \max\{Pr(\alpha \mid \gamma), Pr(\beta \mid \gamma)\}$$

For the example about tips, in subjective terms we are interested to calculate

$$\text{AveragedTip} = \sum_z z \cdot Pr((\text{tip} = z) \mid (\text{food} = x) \wedge (\text{service} = y)).$$

We have that $Pr((\text{tip} = z) \mid (\text{food} = x) \wedge (\text{service} = y)) =$

$$\sum_{G,N,S} Pr((\text{tip} = z) \wedge G \wedge N \wedge S \mid (\text{food} = x) \wedge (\text{service} = y)) =$$

$$\sum_{G,N,S} Pr((\text{tip} = z) \mid G \wedge N \wedge S \wedge (\text{food} = x) \wedge (\text{service} = y)) \cdot$$
$$Pr(G \wedge N \wedge S \mid (\text{food} = x) \wedge (\text{service} = y)),$$

where G is Generous or its negation, N is Normal or its negation, and S is Stingy or its negation.

We assume that the tip is completely determined given G, N and S, therefore

$$Pr((\text{tip} = z) \mid G \wedge N \wedge S \wedge (\text{food} = x) \wedge (\text{service} = y)) =$$
$$Pr((\text{tip} = z) \mid G \wedge N \wedge S).$$

From the Bayes' rule, we have that

$$Pr((\text{tip} = z) \mid G \wedge N \wedge S) = \frac{Pr(G \wedge N \wedge S \mid (\text{tip} = z)) \cdot Pr((\text{tip} = z))}{Pr(G \wedge N \wedge S)} =$$

$$\frac{Pr(G \wedge N \wedge S \mid (\text{tip} = z)) \cdot Pr((\text{tip} = z))}{\sum_u Pr(G \wedge N \wedge S \wedge (\text{tip} = u))} =$$

$$\frac{Pr(G \wedge N \wedge S \mid (\text{tip} = z)) \cdot Pr((\text{tip} = z))}{\sum_u Pr(G \wedge N \wedge S \mid (\text{tip} = u)) \cdot Pr((\text{tip} = u))}.$$

Assuming that all tips are a priori equally likely

$$Pr((tip = z)) = Pr((tip = u)),$$

we obtain that $Pr((tip = z) \mid G \wedge N \wedge S) = \frac{Pr(G \wedge N \wedge S \mid (tip = z))}{\sum_u Pr(G \wedge N \wedge S \mid (tip = u))}$.

$$Pr(G \wedge N \wedge S \mid (tip = u)) = \min \{ Pr(G \mid (tip = u)), Pr(N \mid (tip = u)), Pr(S \mid (tip = u)) \}$$

can be calculated from the given degree curves for the predicates Stingy, Generous and Normal.

$$Pr(G \wedge N \wedge S \mid (food = x) \wedge (service = y)) = \min \{ Pr(G \mid (food = x) \wedge (service = y)), Pr(N \mid (food = x) \wedge (service = y)), Pr(S \mid (food = x) \wedge (service = y)) \}.$$

From the production rules, we assume that

$$Pr(G \mid (food = x) \wedge (service = y)) = \max \{ Pr(Excellent \mid (food = x) \wedge (service = y)), Pr(Delicious \mid (food = x) \wedge (service = y)) \}$$

Considering the food quality to be independent of the service quality, we obtain that

$$Pr(G \mid (food = x) \wedge (service = y)) = \max \{ Pr(Excellent \mid (service = y)), Pr(Delicious \mid (food = x)) \}$$

that can be calculated from the degree curves for Excellent and Delicious.

Similarly,

$$Pr(N \mid (food = x) \wedge (service = y)) = Pr(Good \mid (service = y))$$

and

$$Pr(S \mid (food = x) \wedge (service = y)) = \max \{ Pr(Poor \mid (service = y)), Pr(Rancid \mid (food = x)) \}$$

that can be calculated from the degree curves for Good, Poor and Rancid.