Crystal Gong (cjg5uw)
Cynthia Zheng (xz7uy)

AI 4710 Homework 4

In order to predict the cuisine of a dish given the ingredients, we implemented Naïve Bayes classifier, which is commonly used as a text classifier. Naïve Bayes is simple to implement, powerful, and effective on a large data set. We first shuffled the data and divided it into 6 disjoint sets. We calculated the P(Cuisine), P(Ingredient), and P(Ingredient | Cuisine) using data from the training set, 5/6 of the data. The following formula was used to calculate the cuisine prediction:

$$P(c \mid \mathrm{X}) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

where c is the cuisine and X is the ingredient in the given list of ingredients. The denominator in the Bayes' theorem was ignored as its presence didn't affect the final prediction. We then validated the classifier on the last 1/6 of the dataset to obtain its generalization accuracy. We repeated this process 6 times, changing the validation set each time, to perform 6-fold cross validation.

Because multiplying many small probabilities together is prone to underflow error, we used log(Probability) when calculating the probability of the cuisine to mediate this problem. Furthermore, if an ingredient did not exist in the given cuisine, we set its P(ingredient | cuisine) to 0.00000000000000000001 instead of 0 to prevent ignoring the probabilities of all other ingredients, because if the probability of one ingredient is zero then the probability of the entire set of ingredients given a particular cuisine would be zero.

The time required for training was 0.533 seconds on a 2.9 GHz processor with 16 GB memory. The training set was not extremely large, so training proceeded quickly. The following

table gives the accuracy that the model obtained for each of 6 runs. Variation between each run is due to variation of cuisines/ingredients in each validation set.

Run #	Generalization Error (%)
1	42.80936454849498
2	**49.83277591973244 (maximum)**
3	44.81605351170568
4	44.48160535117057
5	46.48829431438127
6	**42.14046822742475 (minimum)**
Average	**45.094760312151614**