# Homework 2

Your Name: \_Vikas Sanil\_\_\_\_\_\_\_\_\_\_\_\_

Student ID:\_\_A20511552\_\_\_\_\_\_\_\_\_\_\_\_\_

**1.** [40] Use Naïve Bayes Classifier to classify the objects

We conducted a survey to collect people’s daily diets and try to build a model to predict whether their diets result in healthy conditions or not. The final results could be Yes, No, Unsure

| **Breakfast** | **Lunch** | **Dinner** | **Healthy?** |
| --- | --- | --- | --- |
| Ham | Carnivorous | Beef | Y |
| Milk | Carnivorous | Beef | N |
| Bread | Veggie | Pork | U |
| Bread | Veggie | Veggie | Y |
| Ham | Veggie | Veggie | Y |
| Bread | Carnivorous | Beef | N |
| Ham | Veggie | Pork | N |
| Milk | Veggie | Pork | U |
| Milk | Carnivorous | Veggie | U |
| Noddle | Carnivorous | Pork | ? |

1). [15 points] What is laplace smoothing? And why we need it in the Naïve Bayesian classifier?

Laplace smoothing is used to smooth relative frequencies in empirical probability vectors to avoid issues with extreme values(0 or1). It’s called Laplace correction, after the French mathematician Pierre-Simon Laplace, who introduced it for the case k = 2 (also known as Laplace’s rule of succession).

In Naive Bayesian classifier, when there is a feature which has extreme values(0 or 1) probability then it will skew the classification.

2). [25 points] Using the Naive Bayesian Classification Hint: you may need to use laplace smoothing (use the formula in our slide) if you do have zero-conditional probabilities. Use the setting in the slide to solve the problems in this case. Note, only apply laplace smoothing to the ones you have zero-conditional probabilities.

| Breakfast P | healthy Yes | healthy N | healthy U | P(Breakfast|Y) | P(Breakfast|N) | P(Breakfast|U) |
| --- | --- | --- | --- | --- | --- | --- |
| Ham | 3 | 2 | 1 | 0.43 | 0.29 | 0.14 |
| Milk | 1 | 2 | 3 | 0.14 | 0.29 | 0.43 |
| Bread | 2 | 2 | 2 | 0.29 | 0.29 | 0.29 |
| Noodles | 1 | 1 | 1 | 0.14 | 0.14 | 0.14 |
| Total | 7 | 7 | 7 |  |  |  |

| Lunch P | healthy Yes | healthy N | healthy U | P(Lunch|Y) | P(Lunch|N) | P(Lunch|U) |
| --- | --- | --- | --- | --- | --- | --- |
| Carnivorous | 2 | 3 | 2 | 0.40 | 0.60 | 0.40 |
| Veggie | 3 | 2 | 3 | 0.60 | 0.40 | 0.60 |
| Total | 5 | 5 | 5 |  |  |  |

| Dinner P | healthy Yes | healthy N | healthy U | P(Dinner|Y) | P(Dinner|N) | P(Dinner|U) |
| --- | --- | --- | --- | --- | --- | --- |
| Beef | 2 | 3 | 1 | 0.33 | 0.50 | 0.14 |
| Pork | 1 | 2 | 3 | 0.17 | 0.33 | 0.50 |
| Veggie | 3 | 1 | 2 | 0.50 | 0.17 | 0.33 |
| Total | 6 | 6 | 6 |  |  |  |

| P(Y) | 0.33 |
| --- | --- |
| P(N) | 0.33 |
| P(U) | 0.33 |

X` =(Noodle, Carnivorous, Pork)

(P(Noodle|Y) \* P(Carnivorous|Y)\*P(Pork|Y)\*P(Y)) = 0.0032

(P(Noodle|N) \* P(Carnivorous|N)\*P(Pork|N)\*P(N)) = 0.0095

(P(Noodle|U) \* P(Carnivorous|U)\*P(Pork|U)\*P(U)) = 0.0095

P(X') = P(Noodle|Y) \* P(Carnivorous|Y)\*P(Pork|Y)\*P(Y)+ P(Noodle|N) \* P(Carnivorous|N)\*P(Pork|N)\*P(N)+ P(Noodle|U) \* P(Carnivorous|U)\*P(Pork|U)\*P(U) = 0.0222

P(X`|Y) = (P(Noodle|Y) \* P(Carnivorous|Y)\*P(Pork|Y)\*P(Y))/P(X') = 0.14

P(X`|N) = (P(Noodle|N) \* P(Carnivorous|N)\*P(Pork|N)\*P(N))/P(X') = 0.43

P(X`|U) = (P(Noodle|U) \* P(Carnivorous|U)\*P(Pork|U)\*P(U))/P(X') = 0.43

From above values, we can conclude that Healthy value for Noodles Breakfast, Carnivorous Lunch and Pork Dinner would be No or Unsure.

**2. (60 points) Python practice for Naïve Bayes classification**

**Use the Loans data, and run 5 Naïve Bayes techniques to find the best parameters and performance**

* Use Loans\_20K.csv data by using 10-fold cross validation
* Use Loans\_200K.csv data by using 75% as training, 25% as testing

Note:

* You need to change different/multiple parameters to find the best NB model.
* You can find data sets from “slide & data” on blackboard system

Submission

* The ipynb and saved html files
* A comparison of different parameters and accuracy values