Introduction to Machine Learning

Lab 2

許木羽 / 111000177

1. **Why do we take log when implementing the Bayesian Classifier?**

   As if we are taking log, we get from equation 1 to equation 2:

$$P(C|X) = \frac{P(X|C) \cdot P(C)}{P(X)} \qquad \log(P(C|X)) = \log(P(X|C)) + \log(P(C)) - \log(P(X))$$

   Equation 1 Equation 2

   Notice that log(P(X)) would be a constant that won't change no matter what the condition is. And as we want to pick the highest probability, we can assume all probability will be added by log(P(X)) and the highest probability remain the same. Another thing is for computing efficiency, make it slightly faster.

2. **Different between Naïve Bayesian and Gaussian Naïve Bayesian Classifier**

   Naïve Bayesian usually have the input of discrete value, usually as Boolean table. Gaussian Naïve Bayesian has a real number. The idea of Gaussian Naïve Bayesian to determine output true/false is by seeing the data distribution by mean and variance. Likelihood is calculated from expected value from data distribution before and take the highest one.

3. **Difficulty Encountered**

   Compared to Lab 1, this lab is much easier to do as there is template that makes 80%-90% done. I have a bad understanding of how to improve the F1 scores correctly, as I implement weight and outlier by myself. I use z-value and weighting by multiplying each likelihood log value, since if the weight is inside the log is not improving that much. I'm not confident enough to change real data value to discrete value since it takes away the function of Gaussian Naïve Bayes Classifier function. We could do that and just turn our model to Naïve Bayesian, but we don't really know how to make the value to the best discrete value possible, especially this topic is not in my field so I can't make a good assumption. That's my thought!

4. **Summarize**

   Naïve Bayesian is a classifier for discrete value, and Gaussian Naïve Bayesian for real data. The way we classify something is by using probability with given data. The higher probability is, the more likely we pick the answer. We are using logarithmic for the equation as it makes us easier to calculate and the computing time is faster.