Homework 3 – Data Mining I – John Oehninger

Exercise 1

1b) We could make a plot of the error rate and K for a certain range. Then one could easily choose the k that presents smallest error rates. Or we could just add an error rate column to our data frame and then sort by error rate in ascending order. This would present the best k at the top of the data frame, second best k in row 2 and so on. But this does not help us in the case that our model is overfit.

To account for this problem we could apply cross-validation to strength our model, according to the lecture this is the way to go when optimizing the parameter, but I assume we could also apply bootstrapping to optimize.

Further research of this problem has shown me that often an odd number is chosen for k if there are 2 features in the data. Also, the square root of n is used to find a good k.

1c) The training step in k-NN is of time complexity O(1). This is known as lazy learning. Which is really amazing. The space complexity of k-NN is of O(nd), where n is the number of data points and d is the number of features in the data set. In the traning phase, space complexity would also be of O(nd), where n is the number of data points and d the number of features.

1d) No, the time complexity does not increases.

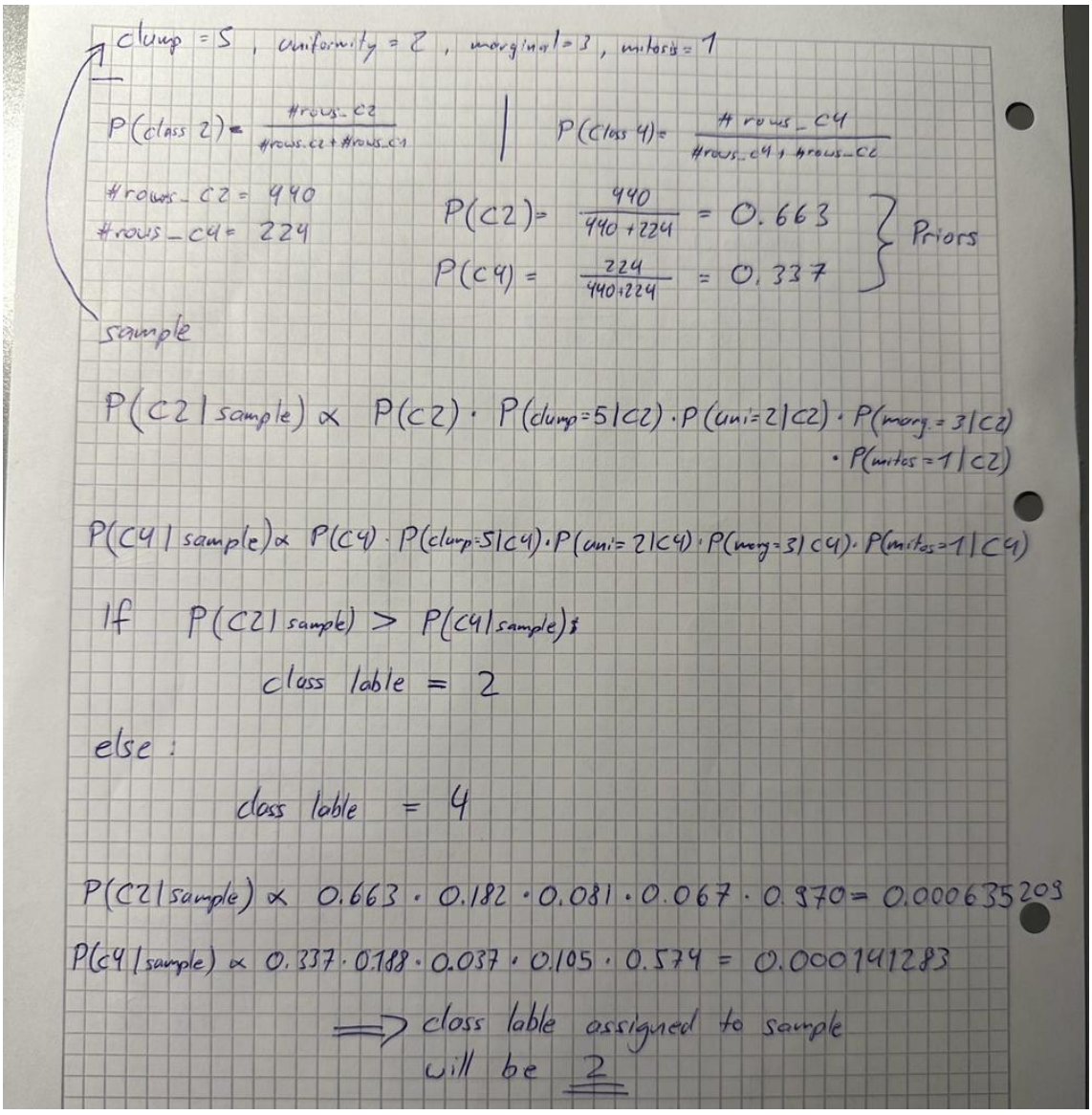
1e) Yes, k-NNdefinitely also works with other metrics, some better than others, but definitely.

I am convinced that k-NN should also work with semimetrics. Some semimetrics can be good distance measures, but since they usualy don't fulfill the triangle inequality they are sometimes not taken into consideration. Dynamic time warping (DTW) could be used well in time series classification.

1f) It is possible to use k-NN for regression. One can approximate the association between independent variables and the dependent variable by averaging the observations in the same area, but the area size must be chosen wisely or could even be chosen using cross-validation.

Exercise 2

2a)

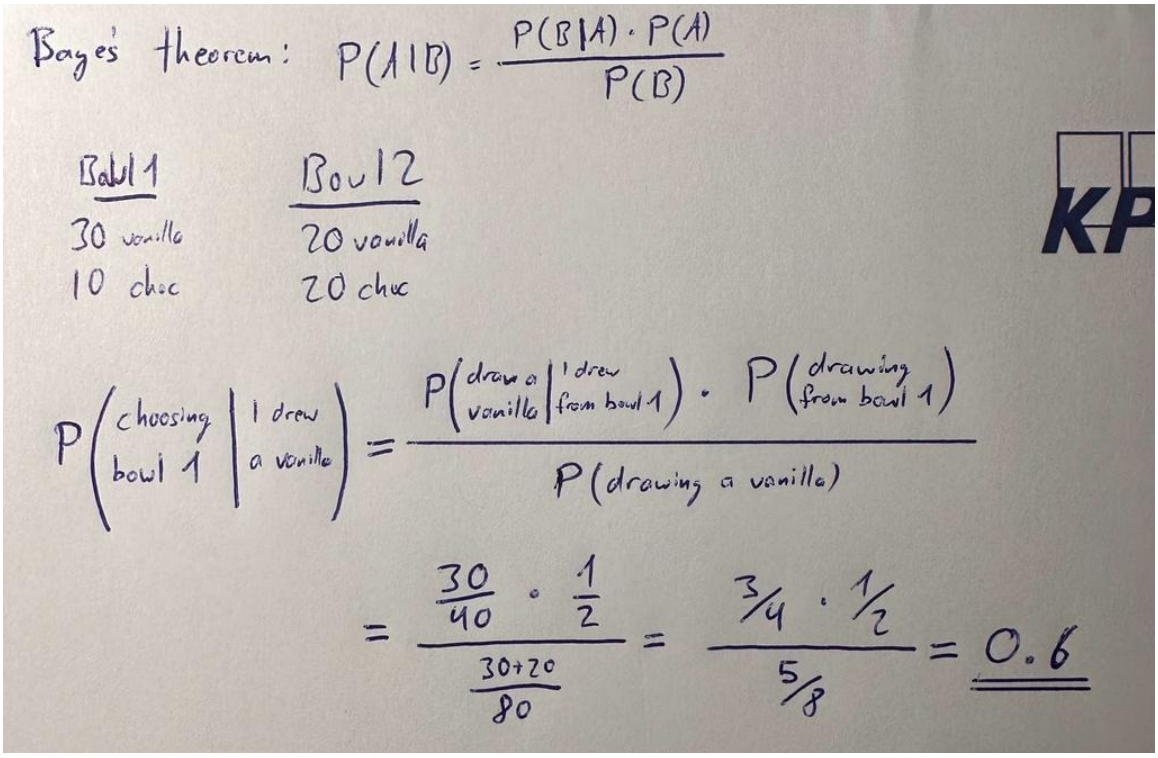


2b) To compute the probabilities one must account for the missing values (nan) by subtracting the number of nan values from the number of samples and then devide the counts by that number.

2c) To account for this problem one could add one count (or any count) to each (word), which would result in never getting a 0 probability in such a scenario. This works very well, because it doesn't change our prior probabilities.

Exercise 3

3a)



3b)

Task 1:

The maximal posterior probability is obtained at N = 60, where the posterior probability is 0.005905417875729855.

Task 2:

The expected value of the posterior distribution is: 333