(1)How does the value of 𝑘affect the fitted model?

In my opinion, KNN is one of the simplest idea of classification. And base on those groups, this regression indicate that nearest samples should share similar properties. Thus, in order to take all k-Nearest Neighbor in to consideration, averaging can be used to predict the result.

Therefore, the larger the k is, the prediction take more sample into consideration. However, it doesn’t means the more sample it use, the better the prediction result is.

In extreme case, if data are random data samples, using KNN is unreasonable.

If k is over large, it will confuse the distinction between the two group, making the classification imprecise. It just like assuming whether you are hungry is only relates to how much you have eaten within three hours. If you also consider the food intake four hours ago, the answer definitely different.

If k is too small, that means we don’t consider all the affections, which also lead to the inaccurate result. Still using the example above. If we only consider total intake within two hours, the result will also change.

I think the performance of k depends on the properties of data sample. Their relation with Nearest Neighbors decides which k is fittest.

(2) If n is the number of observations in the training set, what can you say about a k-NN regression model that uses k = n?

K=n denotes we consider all samples within data set to make prediction at each data point. Thus, prediction value remains unchanged when data point change.

Under the eating prediction explanation, we consider 24 data point (24 hours in a day). Using KNN regression and k=n, thus, our prediction result of how much we’d like to eat at every data point is equal to the average amount of total intake within a certain day.

(3) Are some of the calculated R^2 values negative? If so, what does this indicate?

Negative R^2 means that the prediction result is worse than average of all values. Maybe in reality, our schedule is to eat 30% at 7am, 40% at 12 am and 30% at 6am. But if prediction result is eat 100% and 12 pm in the evening, this model must worse than assuming eating 1/24 of total intake at each point.

(4) What does an R^2 score of 0 mean？

According to the definition of R^2, when a model’s R^2=0, this model is as good as mean value model (k=n under KNN regression) under this criteria of measure the good or bad of model. I want to mention that “as good as” doesn’t means each prediction point is equal to the prediction result of mean value model. It just means the sum of each point’s square of bias are equal to mean value model’s.

(5) Do the training and test $R^2$ plots exhibit different trends? Describe.

They have almost same trends. That because also sample point are in different sheet, they are come from the same system, thus, working and organizing in same way. As long as a data set is split into test set and train set randomly, both of them are from a same population. Although we may can’t figure out how this system work exactly, but our models won’t have scores that have too much differences under each data set.

(6) Explain how the value of k influences the training and test R^2 values.

K denotes how much nearest samples will influence the prediction result. The larger the k is, the more smooth prediction curve will be. This can be interpreted as prediction results will more analogues with sample values when using more values from data set.