Project 1 – k-NN Classification

*Joseph Park - x172917 Date: 4/10/20*

Objective

The assignment is an illustration of classification based on the k nearest neighbors model using particle collision data from a particle accelerator to identify particles.

k-NN Prediction Model

Step 1: Download the data set from Kaggle, look for missing data and look at summary, size and shape.

Step 2: Since the data set is too large to process, randomly sample. (Also, based on summary, as shown below, run process with and without normalization and compare results.)

Step 3: Partition data into training vs testing and target vs classified data.

Step 4: Train and classify data.

Step 5: Create a loop to optimize classification metric of choice, which is accuracy (for particles classification), with respect to k.

Step 6: Plot accuracy of the k’s and write confusion matrix for the k with max accuracy.

Step 7: Normalize the random sample and repeat steps 3-5 and compare results.

Step 8: Reset seed to what it was right before (pseudo)randomly sampling the 5000 particles (because I used same variable names), and repeat steps 2-5, this time with z-score standardization.

p theta beta

## Min. :0.2101 Min. :0.09318 Min. :0.5501

## 1st Qu.:0.6688 1st Qu.:0.28107 1st Qu.:0.8374

## Median :1.0701 Median :0.40215 Median :0.9460

## Mean :1.2352 Mean :0.48677 Mean :0.8995

## 3rd Qu.:1.6564 3rd Qu.:0.60799 3rd Qu.:0.9854

## Max. :4.7210 Max. :2.19003 Max. :1.4903

## nphe ein eout

## Min. : 0.000 Min. :0.00000 Min. :0.00000

## 1st Qu.: 0.000 1st Qu.:0.00000 1st Qu.:0.00000

## Median : 0.000 Median :0.03484 Median :0.04450

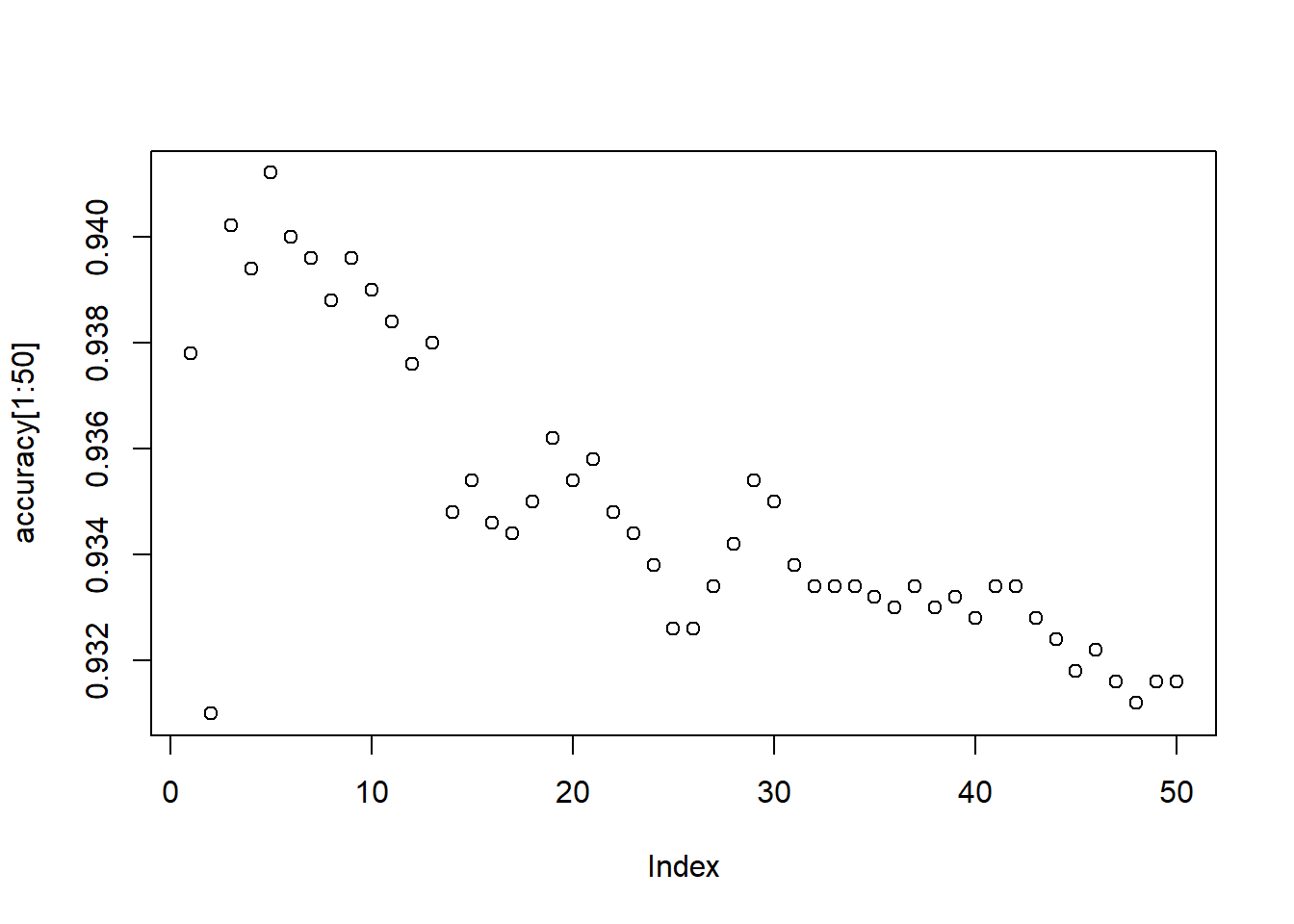
## Mean : 3.563 Mean :0.05664 Mean :0.07125

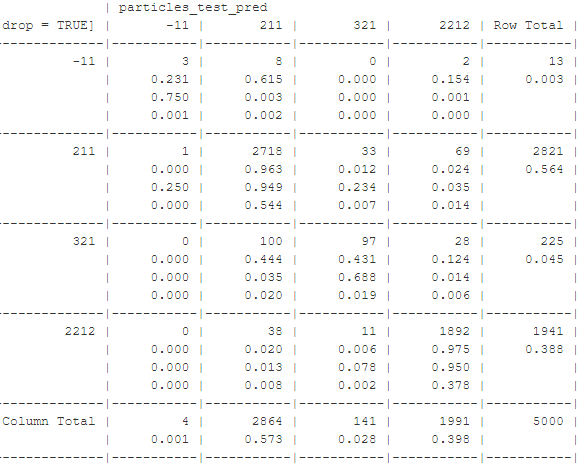
## 3rd Qu.: 0.000 3rd Qu.:0.06812 3rd Qu.:0.09468

## Max. :347.000 Max. :0.85484 Max. :0.95280

Results

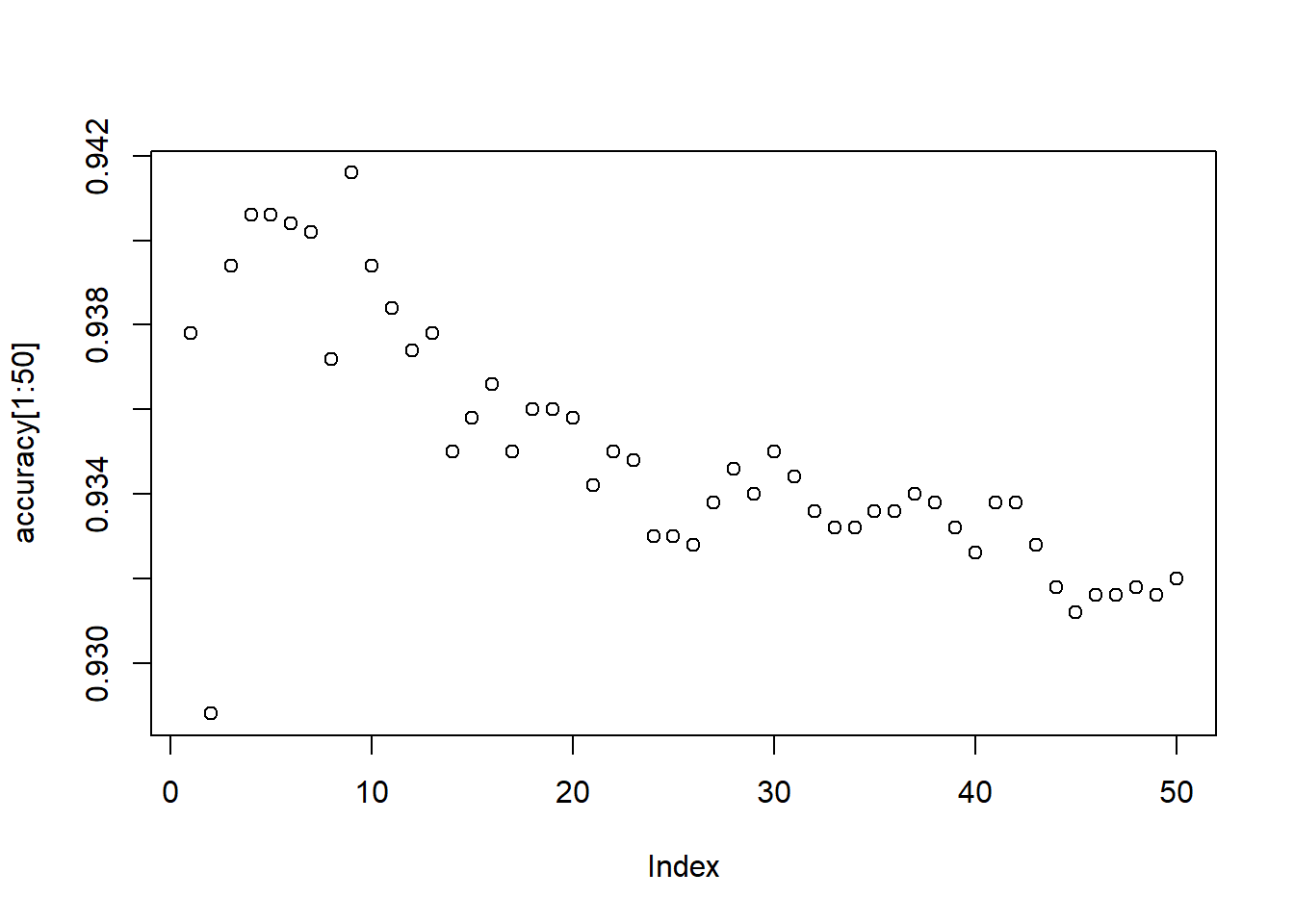
Without standardization

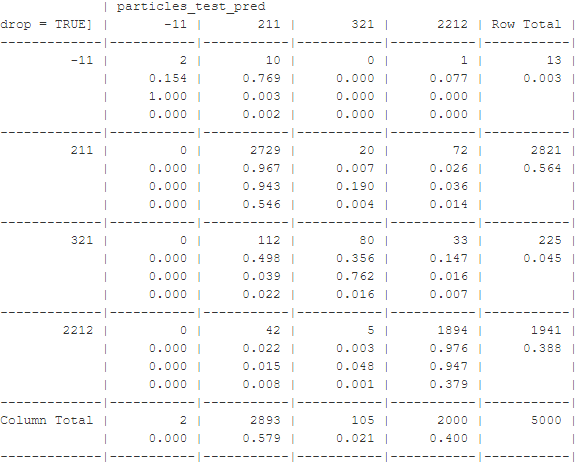




^Columns are predicted values and rows are actual values

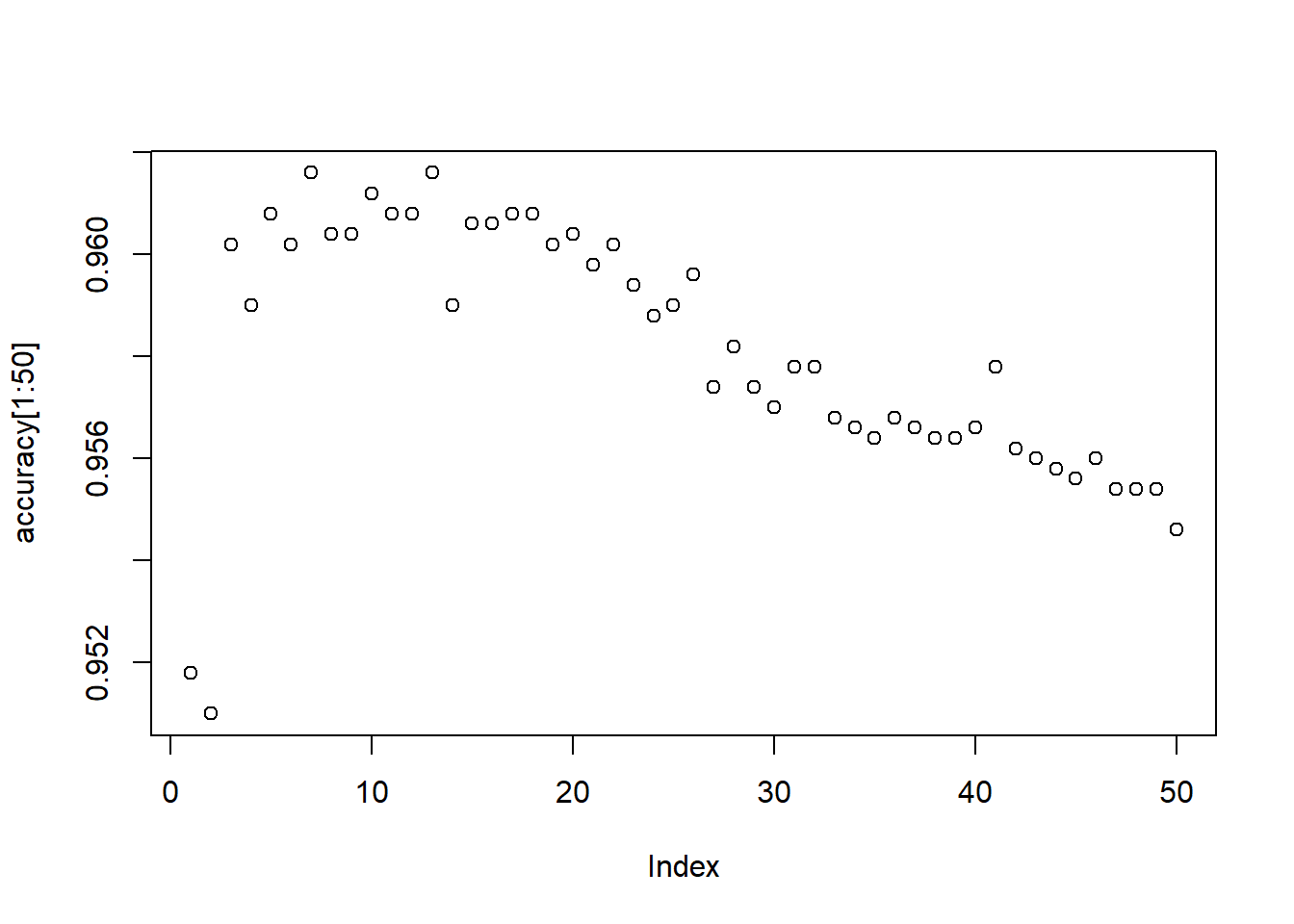
"The best accuracy is 0.9412"

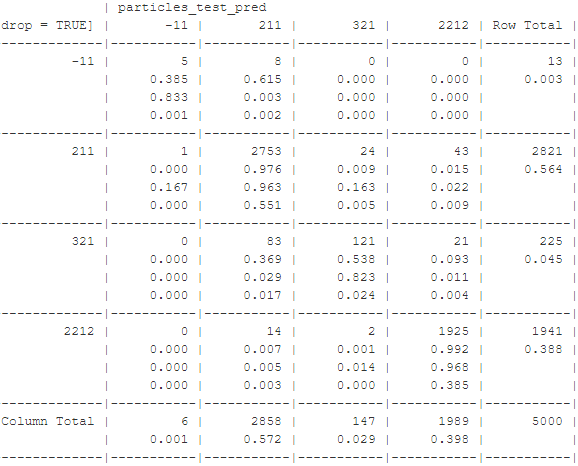
Now with normalization:



"The best accuracy is 0.9416"

Now with z-score standardization:





"The best accuracy is 0.9616"

Interpretation of the Results

Note 1: If particle -11 was of particular interest, a supercomputer (and thus ability to process a sample size significantly more 5000) would be of use, as 13 data points is not reliable.

Note 2: The accuracy is very good for all k’s and has a downward sloping behavior after about k = 5.

Note 3: The order(s) of magnitude difference in actual numbers of particles appear to have the classification overpowered by the numbers of two particles (211 and 2212) that are more prevalent.

Note 4: To elaborate on Note 3, particle 321 is greatly misclassified as these two and disproportionately more to the one that is more prevalent: 100/28 > 2821/1941. Particle -11 is even more misclassified and with similar behavior of disproportion.

Note 5: If we cared in particular about particle -11 or particle 321, we could optimize with respect to metric of choice for whichever particle(s).

Note 6: It seems I was on to something with my assessment of the summary for the following reasons.

Note 7: Accuracy with and without normalization are equal to the fourth decimal place (despite there being **6** features.)

Note 8: Percentage increase of correct classification of particle -11 and particle 321 were significantly greater than the percentage decrease of the other two particles (meaning, if we weighted equally the correct classification of each particle, then classification without standardization would be significantly better.)

Note 9: z-score standardization outperforms all around.