

Lab 04 - Ori Meiraz and Cfir Hadar

Question 1

We've got accuracy of 1.0, because if $k = 1$, the model will always predict the label of the point for itself, because the point is always the nearest to itself.

Question 2

According to our results, the value for K is 19 in the whole data, but 7 in the sample_data.

Question 3

The variance was the lowest, given $n=2$.

Checked on "in1" file, we've printed the average and variance of the accuracies, and here're our results:

- When n increases -
 - The average increases until a peak, then comes down.
 - The variance increases.
- The peak of the average was at 20 fold, **but** the average of 10 was slightly lower, while it's variance was an order of magnitude lower.
- Therefore, **we'd recommend the 10 folds**, because even though the average of the 20 folds is better, the variance of it, is larger thus making our calculation of the average less reliable. In other words, were we to calculate the average of 20 again (with a different random seed), it might be lower than of that of 10 folds.

Question 4

From the results we can infer a couple of notes -

- The accuracy of $K=7$ is slightly better than that of $K=5$.
- Also, the MinMaxNormalizer yields a slightly better accuracy.
 - The most significant improvement of accuracy was from Dummy to MinMax ($K=7$), and even it was only by 1.24%.
 - So, we concluded that no normalization is required, because it's not cost-effective.