Home Assignment 1
Machine Learning A

bkx591
Toke Emil Heldbo Reines

September 12, 2022

1 Make Your Own (10 points)

1. **What profile information would you collect and what would be the sample space $\mathcal{X}$?**

   To minimize bias based on sensitive information, I would not include sex, ethnicity or religion. My sample space would include previous grades on a fixed set of classes respectively - since not all students have identical academic history - study programme, age, hair-style, and shoe size. Alternatively, if a student has not participated in a previous course, a distinct value could be chosen to distinguish between grades and "not completed". Study programme, grades, and hair-style values would be one-hot encoded.

   - $grade\_courses \in [\mathbb{Z}]$
   - $study\_programme_{cs,math,phy,...} \in [0/1, ...]$
   - $age \in \mathbb{Z}$
   - $hair\_style \in \{bald, long, short, alternative\}$
   - $shoe\_size \in \mathbb{R}$

2. **What would be the label space $\mathcal{Y}$?**

   The label space could be one of three labels. A boolean value of whether a student will pass or not, a normalized representation of the 7-scale grade system (1-7) or the 7-grade system itself.

3. **How would you define the loss function $\ell(y', y)$**

For the boolean prediction, I would define the loss function as an averaged hit-or-miss, that is a real number between 0 and 1. For the two other label spaces, the loss function could be some mean numerical or step-wise distance to the truth.

4. **Assuming that you want to apply K-Nearest-Neighbors, how would you define the distance measure $d(x, x')$?**

   An euclidean distance would be a suiting measure.

5. **How would you evaluate the performance of your algorithm?**

   By the definition(s) of my loss function(s), using the boolean loss function I would like a value close to 1. For the two other functions, I would like a value as close to 0 as possible.

6. **Assuming that you have achieved excellent performance and decided to deploy the algorithm, would you expect any issues coming up? How could you alleviate them?**

   My model would not necessarily have any causal insights or explainable reasoning as why it predicted what it did. In other words, having a model with this much impact and responsibility, comes with a lot of ethical questions. I mentioned that I would not include sensitive information, but there might still be a bias or hidden link between shoe size and grades or pass/no-pass.
   Another issue that comes to mind is the size of our sample space and also how the features are distributed - we might see outliers in the data.

2 Digits Classification with K Nearest Neighbors (45 points)

We notice immediately, the decreasing maximum value of the validation error as the value of n increases. From this, my intuition is, that the model performs better the more data we include in our validation set. The specific relation between the validation error, the size of our validation set and the value of K tells us something about the overlap/isolation and spread-outńess of our clusters (there probably is a word for that). With a small validation set, we have a large variation in validation error between small and large values of K, which could indicate, that the probability of whether an 28x28 image is a 5 or a 6, is close to even.

The smaller value of n, the higher variance or fluctuation of validation error.

The larger a validation set, the more evenly distributed the 5 and 6 digits will be. We can see the gradually smaller variance from n=10 and n=80 in 2, as the validation set grows in size.

My conclusion as to what influence K has on the prediction accuracy, I infer from the 4 plots a tendency, that the lower the value of K, the greater prediction accuracy. Though at $n = 80, i = 4/5$ we see a plateau, and maybe even an increase in prediction accuracy at around $K = 30$.

As for the corrupted datasets, we see a graduate approach to randomness. Given our label space of 2 (digits 5 and 6), a random guess would on average be 50% correct, which we see stabilizing at around $K = 40$ with the heavily corrupted dataset.
For the uncorrupted dataset, we see no immediate increase in prediction quality when K increases, but for even the lightly corrupted dataset, this quality increases up to an optimal value of $K = 10$, after which it decreases again.

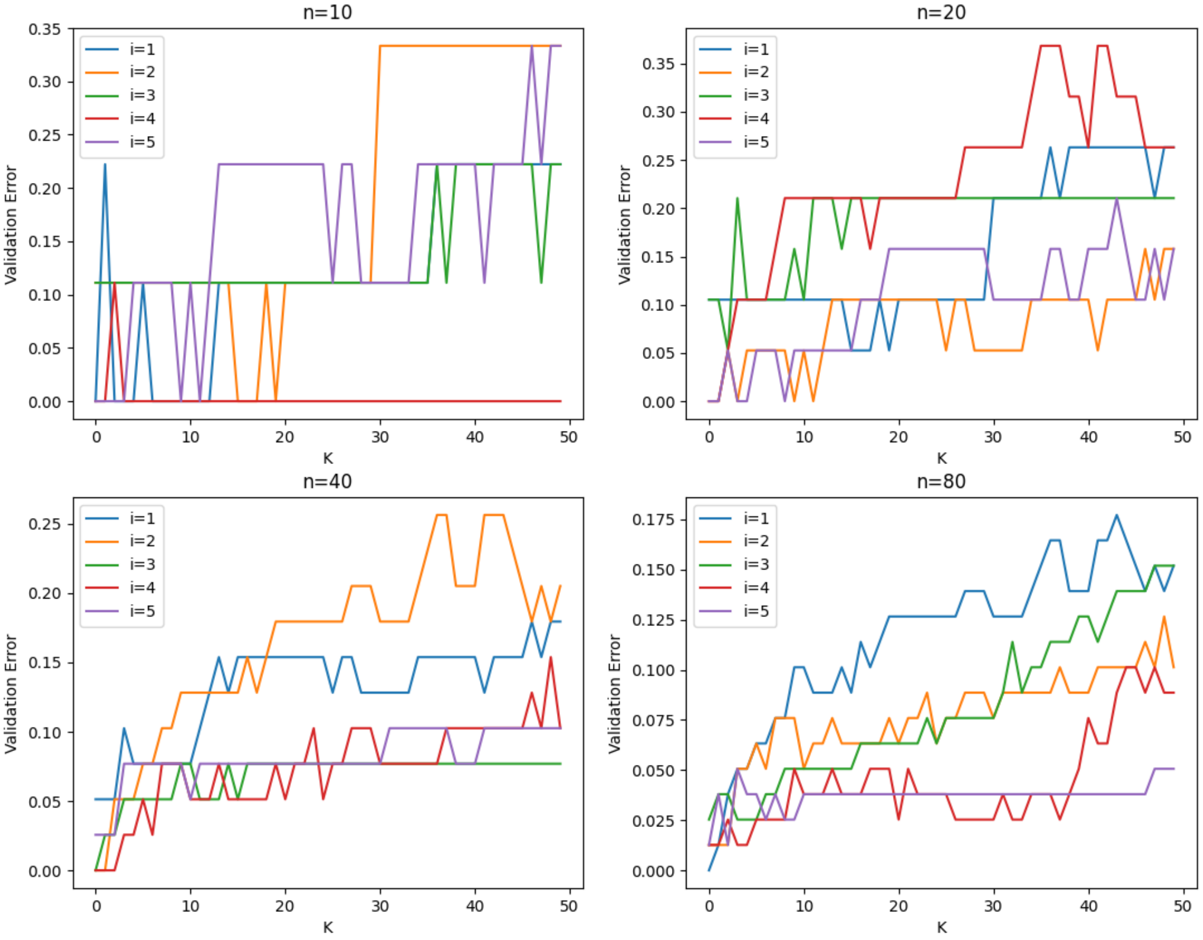

Figure 1: Validation error on K-NN using an increasingly larger validation set of size $n * i$ evaluated on K sizes of [0-50]

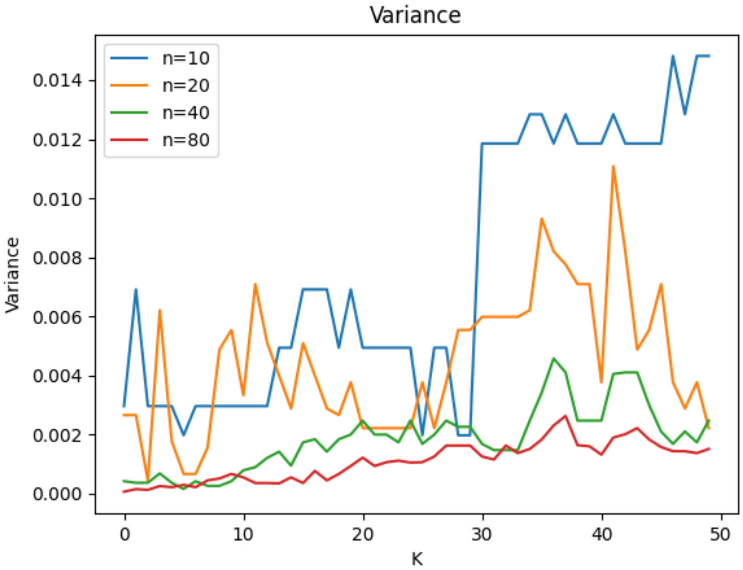

Figure 2: KNN Variance of validation error.

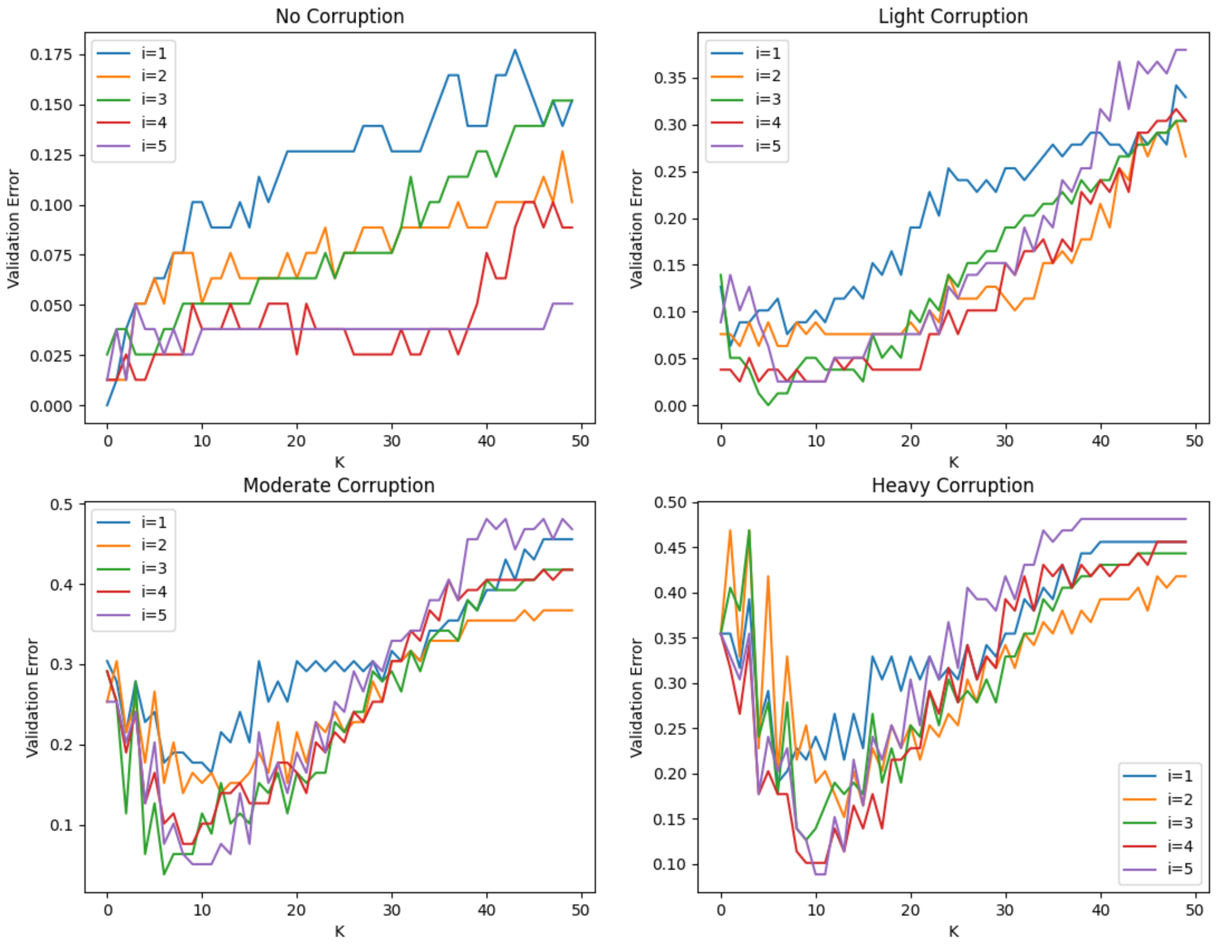

Figure 3: Validation error on K-NN using an increasingly larger validation set of size $n * i$ evaluated on K sizes of [0-50]

3 Linear Regression (45 points)

For this task, I have fitted two models, a non-linear model, and a linear model. I have two corresponding visualisations, both on the logarithmic scale on PCB values to better compare and visualize the fit of the square rooted feature space compared the regular.

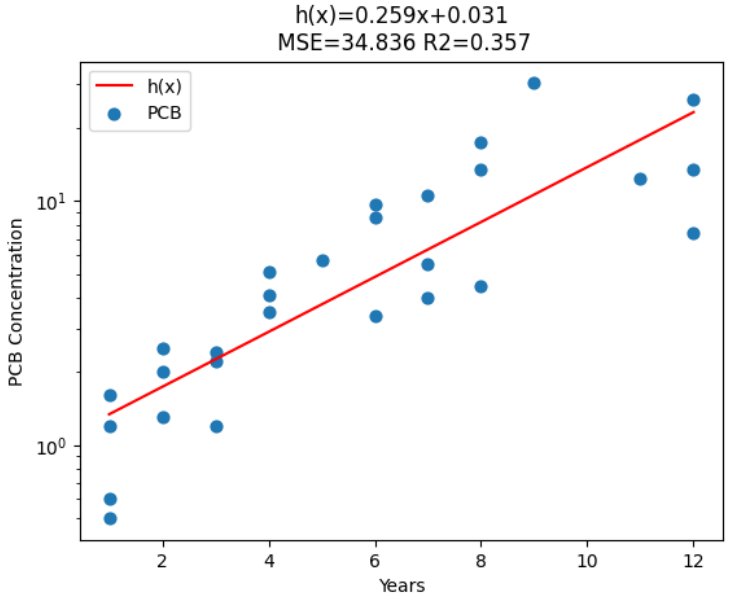

Figure 4: Linear model, $h(x)$, in raw label space.

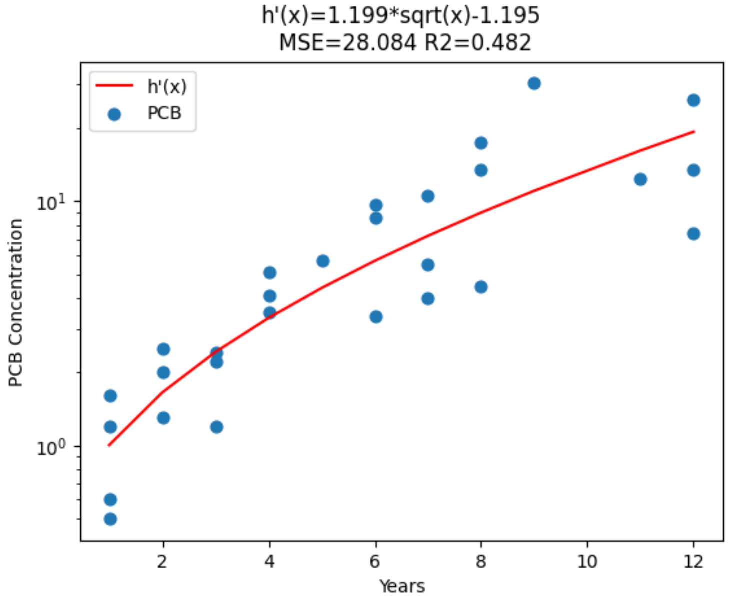

Figure 5: Non-linear model, $h'(x)$, in square rooted label space.

3.1 Discussion of R2

R^2 tells us good our model is at predicting. The MSE does as well, but in a space, not necessarily relatable to anything, as it is dependent on the actual values of our label space (PCB). R^2 is a ratio and therefore a real number between 0 and 1. It is not bound to the range of our label's values, which makes it easier to interpret. In this task, we have constructed two models, and using R^2 to compare them, gives us a relatable metric as to which model performs the best. On that matter, our second model is a non-linear model which achieves a better R^2, which indicates, that a linear-model might not be the best fit for our data. The R^2 value is 13% points better in the non-linear model than in the linear model - we can relate to that, where as the 6 MSE points are not relatable.