Assignment 1 - Machine Learning, Fall 2018

Cristian Mitroi

27 November 2018

1

1.1

I would collect information about:

- age
- grades in different previous courses that are related:
  - programming
  - algorithms
  - data science
  - statistics
  - probability
- nr of bachelor diplomas
- nr of master diplomas

The sample space would be previous students from the ML course.

1.2

The label space would be the grades in 7-point Danish scale: -3, 0, 2, 4, 7, 10, 12. I would encode these as: 0, 1, 2, 3, 4, 5, 6

1.3

I would choose a loss function that captures the continuity of the grading scale (as a regression problem). Though I would make sure to round the resultant prediction

I would choose mean-squared error:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

1.4

I would hold out a part of the samples for validation and test. I would select the optimal parameters in terms of their performance on the validation set. Then I would evaluate my model's performance in terms of its accuracy on the test set.

1.5

It is possible that the model might underfit or overfit. In these cases I would attempt to gather more sample points (perhaps students from other years) or more features (years experience in programming, high school grades, etc.)

2

2.1

See plot

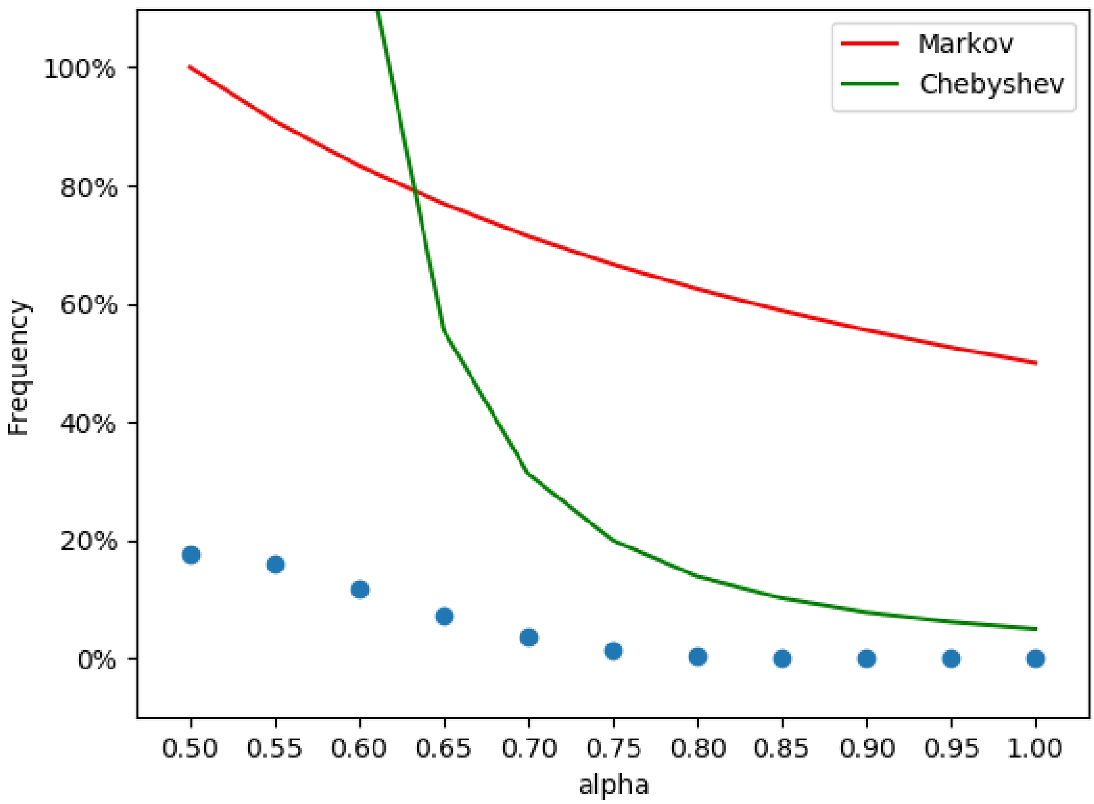

Figure 1: plot exercise

2.3

With $\alpha = 0.51$ we would have the right-hand side be $0.51 * 20$, which would not be an integer.

2.4

See figure "plot exercise" above.

2.5

See figure "plot exercise" above.

2.6

We can see that the Markov bound is very loose when compared with the actual empirical frequencies.

By comparison, the Chebyshev bound is tighter. But even in that case, it is quite loose.

2.7 For alpha = 1 and 0.95 calculate the exact probability

For $\alpha = 1$

The probability can't be larger than 1. Thus we have:

$P(\frac{1}{20}\sum_{i=1}^{20} X_i = 1)$

$P(\frac{1}{20}\sum_{i=1}^{20} X_i = \frac{20}{20})$

Thus we need all the coin tosses (20) to be 1 (heads, by convention).

The probability of one coin toss to be 1 is 1/2. The probability that all 20 of them are 1 is $(1/2)^{20}$.

For $\alpha = 0.95$

$P(\frac{1}{20}\sum_{i=1}^{20} X_i \geq \frac{19}{20})$

Thus we have need at least 19 coin tosses to be 1 (heads, by convention). The probability that 19 of them are 1 is $(1/2)^{19}$. To this we add the probability that all of them are 1 (see above). Thus the final probability is $(1/2)^{20} + (1/2)^{19}$.

3 Tightness of Markov's Inequality example

We have $P(x = 1) = P(x = 0) = 0.5$

Thus:

$E(X) = 0.5$

$P(x \geqslant 1) = \frac{E(X)}{1}$

$P(X \geqslant 1) = \frac{0.5}{1} = 0.5$

The only case is $P(X = 1)$. Which is indeed 0.5 *q.e.d.*

4 kNN

Discussion

We can see that in the case of 0 and 1 the relation between the error rate and the validation error stays the same as K changes. The error values themselves don't change either. This might be because comparing 0 and 1 is very easy.

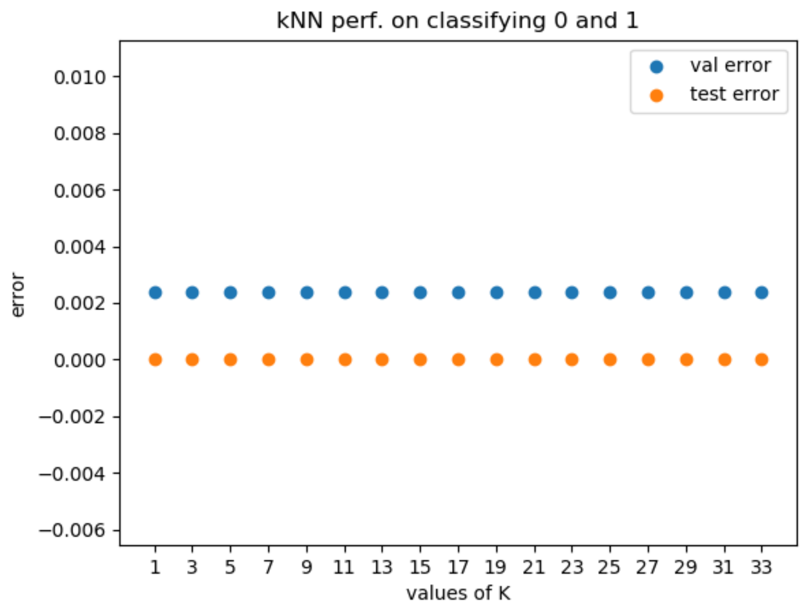

Figure 2: ex4-1

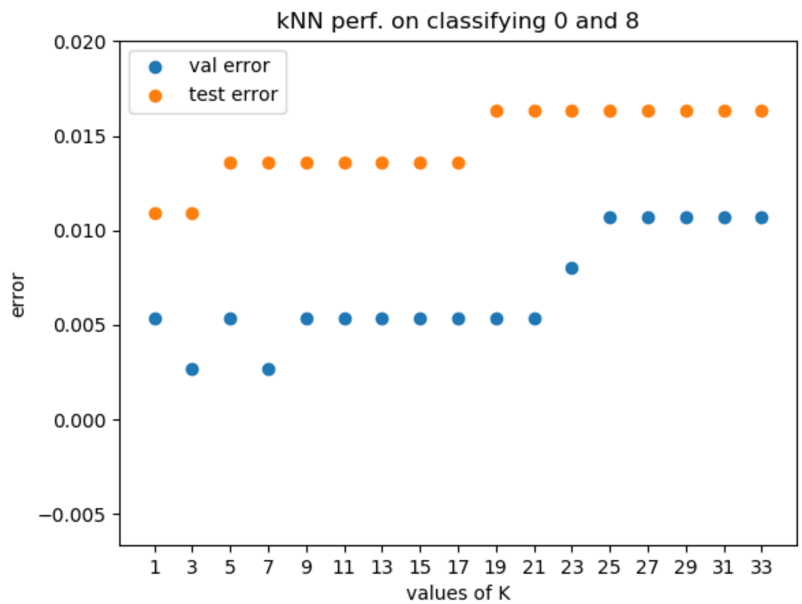

Figure 3: ex4-2

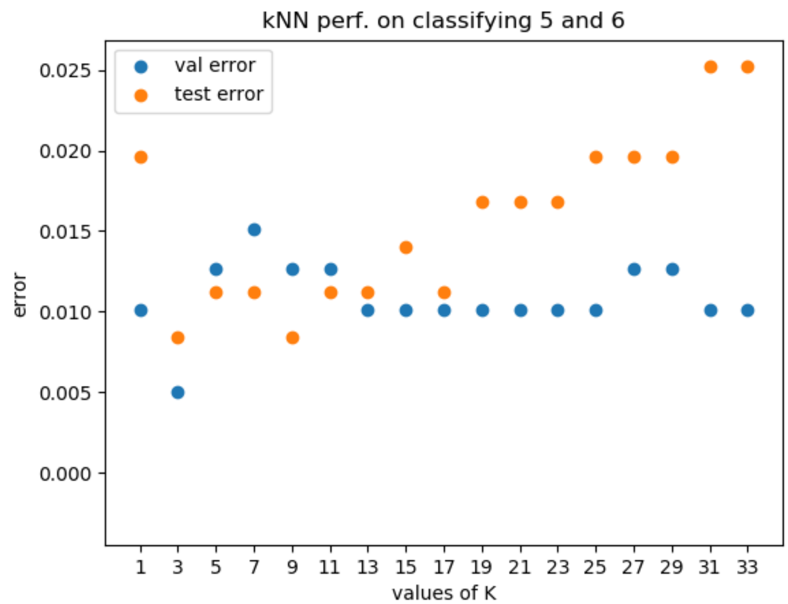

Figure 4: ex4-3

In the case of 0 and 8, it seems the optimal Ks are 1 and 3 for the test set. We can also see how the error increases the more we increase the value of K. This might be due to the increased complexity of distinguishing between 0 and 8.

In the case of 5 and 6 the test error is increasing faster than the validation error is, for the higher values of K. Also, there doesn't seem to be a clear relation between validation error and test error for the lower range of values for K.

5 kNN multiclass

For the two-class digit classification (0 vs 1, 0 vs 8, 5 vs 6) I simply extracted all the entries in the train, test and validation set that included those digits.

For multi-class classification I would simply use all the data. I would perform the predictions on the entire training, validation, and test sets.

6 Linear Regression

6.2

$$w = 9.4893, b = -10.4270$$

$$MSE = 0.0124$$

6.3

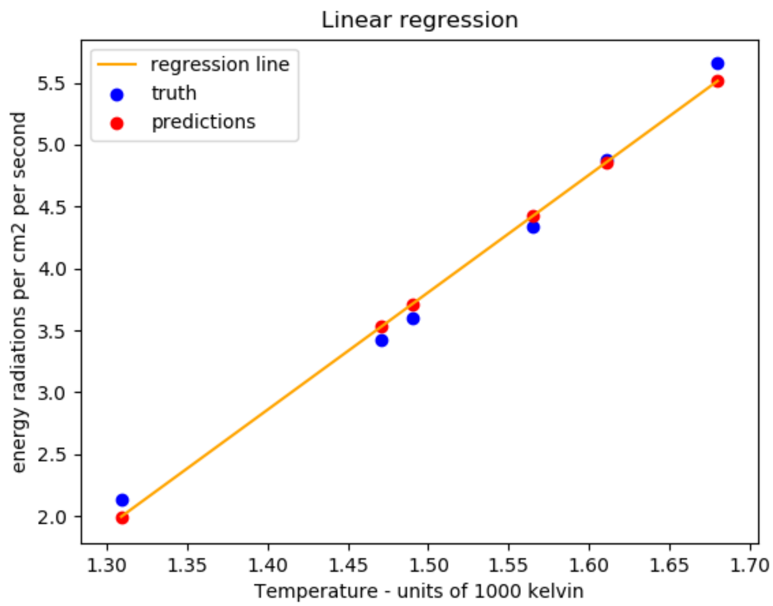

Figure 5: plot ex 6.3

6.4

$Var = 1.2689$

$\frac{Var}{MSE} = 0.0098$

$Variance$ defines the difference between each entry and the mean value.

MSE measures the difference between each prediction and the ground truth label.

$\frac{Var}{MSE}$ thus measures the accuracy of the regression model. If this is close to 1, it means that the model cannot predict better than the mean. If it's lower, or close to 0, it means the model is performing well.

6.5

$MSE = 0.0005$

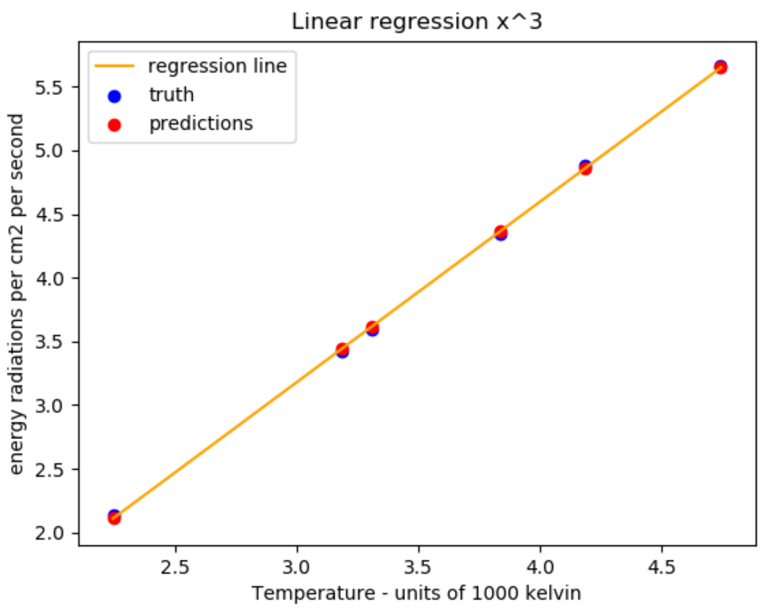

Figure 6: plot ex 6.5