

Assignment 1

Silvan Robert Adrian

November 26, 2018

Contents

1	Make Your Own	1
2	Illustration of Markov's and Chebychev's Inequalities	2
3	Tightness of Markov's Inequality	3
4	Digits Classification with Nearest Neighbours	3
5	Nearest Neighbours for Multiclass Classification	4
6	Linear Regression	4

1 Make Your Own

1. My X would take values of: Degree average (7-scale) $\in \{1, 2, 3, 4, 5, 6, 7\}$, average number of sleep hours $\in \{0, \dots, 24\}$, belong to DIKU $\in \{0, 1\}$, ECTS in maths $\in \mathbb{R}$
2. Seven grade scale, $y \in \{1, 2, 3, 4, 5, 6, 7\}$
3. I would use a least square lost function. This would take into account how close the predictor is from the actual label and penalize outliers.
 $\ell(Y', Y) = \sum_{n=1}^n (Y - Y')^2$
4. Three different datasets would be needed. A train set which would be used for calculating different prediction rules of the algorithm. Next, a train set which given the prediction rules will predict Y' , given a known Y , the loss function can be used for calculating the error. The best prediction rule will be chosen. Finally, in order to evaluate its performance, I would use a validation set, from where an unbiased estimator of the expected lost can be calculated using the loss function. This is an indicator of the performance.
5. Possible issues are: the algorithm does not generalized well, if the yearly samples are not independent among them (some years tend to have more students of a certain profile or the grading system is somewhat different), then perhaps a more complex model or loss function (for instance generalized least squares) would be helpful for improving the predictions. At the same time, it is possible that the information collected is not correlated with the final grade of the student, and that it works well during the training due to an excessive amount of features (overfitting). In this case, the whole model should be changed to a different set of features and larger amount of data.

2 Illustration of Markov's and Chebychev's Inequalities

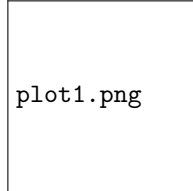


Figure 1: Representation of fraction of experiments above or equal to the threshold (X-axis). In blue there is the empirical fraction, while in green and red its upper bound was computed using Chebyshev's and Markov's inequalities respectively.

2. In Fig1 (blue line) we may observe the empirical frequency of binary 0,1 experiments which mean is bigger or equal than a certain threshold (X-axis).

3. As we are calculating averages of 20 discrete outcomes, there are only certain values that the mean can take, which in this case only include multiples of 0.05. Therefore, the probability of having a number bigger or equal than 0.51 is the same than 0.55.

4. Markov bound can be observed in red on Fig1. In order to calculate the Markov bound, the theoretical mean of the distribution of outcomes on 20 coin tosses was assumed to be equal the mean of a single Bernoulli trial (by central limit theorem) with value 0.5.

5. In green Chebyshev's bound might be seen in Fig1. Variance was calculated by central limit theorem ($0.5 * 0.5 * 20^{-1}$). The bound was limited to probability values minor or equal than 1.

6. The frequency of experiments averages over or equal a certain α decreases exponentially (blue line), nevertheless Markov bound decreases in a much slow rate (it is a less robust upper bound since is far from the estimated expectation). Chebyshev's for low thresholds do not perform well, it is highly loose. For higher thresholds, it becomes a better bound than Markov, thanks to the fact that it takes into account the amount of samples used (1^6).

7. The probability of observing the mean of an experiment over or equal 0.95 is the same than getting 19 out of 20 coins flipping on the same side. This can be calculated as a binomial distribution($p=0.5, n=20$).

$$P(x \geq 19) = P(X = 19) + P(X = 20) = 20 * 0.5^{19} * 0.5 + 0.5^{20} = 1.907349e - 05 + 9.536743e - 07 = 2.002716e - 05$$

For being equal or bigger than 1 is the same than equal to one, which can be model as a binomial($p=0.5, n=20$) of getting 20 out of 20 coins flipping to the same side.

$$P(X = 20) = 0.5^{20} = 9.536743e - 07$$

3 Tightness of Markov's Inequality

Given the equality: $\frac{E(x)}{\epsilon} = P(x \geq \epsilon)$

For a random variable $X \in \{0, \epsilon\}$

$$\frac{(0 \cdot P(0) + \epsilon \cdot P(\epsilon))}{\epsilon} = P(\epsilon) \text{ As there are only two numbers } P(x \geq \epsilon) = P(\epsilon)$$

$$\frac{\epsilon \cdot P(\epsilon)}{\epsilon} = P(\epsilon) ; P(\epsilon) = P(\epsilon)$$

4 Digits Classification with Nearest Neighbours

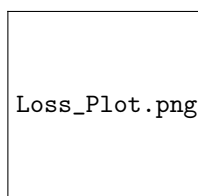


Figure 2: Representation of the loss of K-nn classifier for different Ks (X axis). In the legend the different classifications can be seen. The prefix Val makes reference to the usage of the validation dataset (Last 20% of train) while test uses the complete test dataset for calculating the loss. Please note that val 0-1 and test 0-1 overlap.

- How well does the validation error match the test error?

The figures show a similar pattern. Generally with a higher error in the test than in the validation dataset.

- How closely does the best value of K according to the validation error match the best value of K according to the test error?

They are really close, 5-6 and 0-1 totally agree, while 0-8 even though shows a really similar trend do not match their maximum minimum.

- How the validation and test errors behave as a function of K?

The error is correlated with the value of K. This correlation tends to increase the error with higher Ks, but there are some intermediate regions of low error.

- Does the best value of K depend on the difficulty of the task and how? (It is easier to tell apart "0" and "1" than "5" and "6"; the difficulty of separating "0" and "8" should be somewhere in between.)

0-1 is the easiest classification and is seen since the best value of K is almost any value for both validation and test. 0-8, seems to prefer low and intermediate k values, while the most difficult task requires low values of k (particularly 3). This can be due to the fact that there are much more fuzzy differences

between the numbers and if further away correlations are chosen many mislabels will be taken on the way.

5 Nearest Neighbours for Multiclass Classification

K Nearest Neighbors (K-NN) for Multiclass Classification with $Y = \{1, -1\}$

- 1: Input: Training labeled points $(x_1, y_1), \dots, (x_n, y_n)$ and a point x that has to be classified.
- 2: Calculate the distances $d_i = d(x_i, x)$.
- 3: Sort d_i in ascending order.
- 4: For each position calculate the cumulative mode of the labels.
- 5: The mode of a certain position (k) presents the classification given by K neighbors.

6 Linear Regression

1. See `code.zip`
2. W_1 or line slope: 9.4893 ; b or intercept: -10.4269
3. The plot is provided in Fig3.

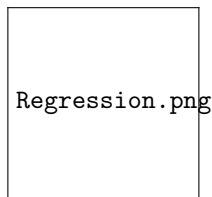


Figure 3: Linear regression of thousands Kelvin degrees vs energy. The scatter dots represent the data points. Two different linear models were adjusted by using Least Squares. In orange a simple affine linear model, while in blue, the data was transformed using X^3 before using least squares.

4. Error: 0.012434221615054074, variance= 1.2689295555555555 ; Quotient: 0.009798984948073176.
The quotient is much lower than 1. This means that the spread of data across the regression line is lower than the spread of data across its mean. If the regression line were constant and at the mean, this coefficient would be one. This would mean that there is no linear correlation between both variables. This coefficient is not expected to be bigger than 1, in non-linear models the best solution will be at least the mean and if the least squares is calculated from a predictor that corresponds to the mean, both the variance and loss are equal.
5. Independent b or intercept: -1.0663, W_1 or line slope: 1.4163. The error: 0.0005. The plot can be seen in Fig3.