Machine Learning A

2023-2024

Home Assignment 2

Christian Igel **Yevgeny Seldin**
Department of Computer Science
University of Copenhagen

The deadline for this assignment is **14 September 2023, 18:00**. You must submit your *individual* solution electronically via the Absalon home page.

A solution consists of:

- A PDF file with detailed answers to the questions, which may include graphs and tables if needed. Do *not* include your full source code in the PDF file, only selected lines if you are asked to do so.
- A .zip file with all your solution source code with comments about the major steps involved in each question (see below). Source code must be submitted in the original file format, not as PDF. The programming language of the course is Python.
- IMPORTANT: Do NOT zip the PDF file, since zipped files cannot be opened in *SpeedGrader*. Zipped PDF submissions will not be graded.
- Your PDF report should be self-sufficient. I.e., it should be possible to grade it without opening the .zip file. We do not guarantee opening the .zip file when grading.
- Your code should be structured such that there is one main file (or one main file per question) that we can run to reproduce all the results presented in your report. This main file can, if you like, call other files with functions, classes, etc.
- Handwritten solutions will not be accepted. Please use the provided latex template to write your report.

1 Illustration of Markov's, Chebyshev's, and Hoeffding's Inequalities (24 points)

2.a Make 1,000,000 repetitions of the experiment of drawing 20 i.i.d. Bernoulli random variables $X_1, \dots, X_{20}$ (20 coins) with bias $\frac{1}{2}$ and answer the following questions.

1. Plot the empirical frequency of observing $\frac{1}{20}\sum_{i=1}^{20} X_i \geq \alpha$ for $\alpha \in \{0.5, 0.55, 0.6, \dots, 0.95, 1\}$.
2. Explain why the above granularity of α is sufficient. I.e., why, for example, taking $\alpha = 0.51$ will not provide any extra information about the experiment.
3. In the same figure plot the Markov's bound[1] on $\mathbb{P}\big(\frac{1}{20}\sum_{i=1}^{20} X_i \geq \alpha\big)$.
4. In the same figure plot the Chebyshev's bound[2] on $\mathbb{P}\big(\frac{1}{20}\sum_{i=1}^{20} X_i \geq \alpha\big)$. (You may have a problem calculating the bound for some values of α. In that case and whenever the bound exceeds 1, replace it with the trivial bound of 1, because we know that probabilities are always bounded by 1.)
5. In the same figure plot the Hoeffding's bound[3] on $\mathbb{P}\big(\frac{1}{20}\sum_{i=1}^{20} X_i \geq \alpha\big)$.
6. Compare the four plots.
7. For $\alpha = 1$ and $\alpha = 0.95$ calculate the exact probability $\mathbb{P}\big(\frac{1}{20}\sum_{i=1}^{20} X_i \geq \alpha\big)$. (No need to add this one to the plot.)

2.b Repeat the question (points 1-7 above) with $X_1, \dots, X_{20}$ with bias 0.1 (i.e., $\mathbb{E}[X_1] = 0.1$) and $\alpha \in \{0.1, 0.15, \dots, 1\}$.

2.c Discuss the results.

Do not forget to put axis labels and a legend in your plot!

[1] Markov's bound is the right hand side of Markov's inequality.

[2] Chebyshev's bound is the right hand side of Chebyshev's inequality.

[3] Hoeffding's bound is the right hand side of Hoeffding's inequality.

2 The Role of Independence (13 points)

Design an example of identically distributed, but *dependent* Bernoulli random variables $X_1, \dots, X_n$ (i.e., $X_i \in \{0, 1\}$), such that

$$\mathbb{P}\left(\left|\mu - \frac{1}{n}\sum_{i=1}^{n} X_i\right| \geq \frac{1}{2}\right) = 1,$$

where $\mu = \mathbb{E}\left[X_i\right]$.

Note that in this case $\frac{1}{n}\sum_{i=1}^{n} X_i$ does not converge to μ as n goes to infinity. The example shows that independence is crucial for convergence of empirical means to the expected values.

3 Tightness of Markov's Inequality (Optional question not for submission, 0 points)

In the previous question you have seen that Markov's inequality may be quite loose. In this question we will show that in some situations it is actually tight. Let ε^* be fixed. Design an example of a random variable X for which

$$\mathbb{P}(X \geq \varepsilon^*) = \frac{\mathbb{E}\left[X\right]}{\varepsilon^*}.$$

Prove that the above equality holds for your random variable.

Guidance: "Design a random variable" means design a distribution by which the random variable is distributed. To design a distribution you should say what values the random variable can take and with what probabilities. You should construct an example, where the random variable can take strictly more than one value, because otherwise the example is trivial. But two values are actually sufficient to make a valid non-trivial example.

4 The effect of scale (range) and normalization of random variables in Hoeffding's inequality (13 points)

Prove that Corollary 2.5 in Yevgeny's lecture notes (simplified Hoeffding's inequality for random variables in the $[0, 1]$ interval) follows from Theorem 2.3 (general Hoeffding's inequality). [Showing this for one of the two inequalities is sufficient.]

5 Linear Regression (50 points)

Consider the data $S = \{(x_1, y_1), \dots, (x_N, y_n)\}$ in the file `PCB.dt`[4], which contains the concentrations of polychlorinated biphenyl (PCB) residues in lake trouts from the Cayuga Lake, NY, as reported by Bache et al. (1972). "The ages of the fish were accurately known, because the fish are annually stocked as yearlings and distinctly marked as to year class. Each whole fish was mechanically chopped, ground, and thoroughly mixed, and 5-gram samples taken. The samples were treated and PCB residues in parts per million (ppm) were estimated using column chromatography" (Bates and Watts, 1988).

Each line in `PCB.dt` is one training pattern. The first number is the input (x), the age of the fish in years, and the second is the corresponding output / target / label (y), the PCB concentration (in ppm).

Tasks

1. Implement the linear regression algorithm as described in the lecture. For vector and matrix operations, such as computing the inverse of a matrix, you can use high-level (library) functions (e.g., from NumPy). Show the implementation of the linear regression algorithm in your report (you need not show boilerplate code, loading of teh data , etc.).

2. The task is to build a non-linear model $h : \mathbb{R} \to \mathbb{R}$ of the data in `PCB.dt`. The model should be of the form

   $$h(x) = \exp(ax + b) \tag{1}$$

   with parameters $a, b \in \mathbb{R}$.

   The parameters can be learned using linear regression in the following way. Before applying linear regression, transform the output data (the y values) by applying the natural logarithm.

   Then build an affine linear model using the transformed targets. By doing so, you effectively learn the (non-linear) model (1).

   That is, you have to perform the following steps:

   - Construct the data set $S' = \{(x_1, \ln y_1), \dots, (x_N, \ln y_n)\}$.
   - Fit a model $h'(x) = ax + b$ to S'.
   - Obtain the final model as $h(x) = \exp(h'(x))$

   Build the model, report the two parameters of the model as well as the mean-squared-error of the model h computed over the training data set S.

[4] Download from `https://github.com/christian-igel/ML/tree/main/data`

3. In this particular example, linear regression without the non-linear transformation gives a lower error. Do not get confused by this: In the above procedure we train for a low error "on logarithmic scale", and minimizing this error may not minimize the error on the original scale, which is the error you should consider in this exercise. The idea is that the data looks more linear on logarithmic scale, this is why a model of the form (1) (a standard *allometric* equation, Huxley and Teissier, 1936) has been considered in literature.

   Provide an example showing that

   $$\arg\min_{a,b} \sum_{i=1}^{N} (y_i - h(x_i))^2 \neq \arg\min_{a,b} \sum_{i=1}^{N} (\ln y_i - (ax_i + b))^2 \ . \tag{2}$$

   Note two important things: First, the transformation is non-linear. Second, we are considering a transformation of the output space that changes the error/objective function – the learning goal. In task 6, we change the input representation and compare approaches using same objective function.

   Provide an instructive example with two data points. An easy way is to assume that the output is independent of the input and a data set $S = \{(x_1, y_1), (x_2, y_2)\}$ with $x_1 = x_2$ and $y_1 \neq y_2$.

4. Plot the data and the model output. The plot must have proper axis labels and a legend. In *all* plots in this assignment, plot the logarithm of the PCB concentration versus the (not transformed) age.

5. Compute the coefficient of determination R^2 as

   $$1 - \frac{\sum_{i=1}^{N} (y_i - h(x_i))^2}{\sum_{i=1}^{N} (y_i - \overline{y}))^2} \ , \tag{3}$$

   where $\overline{y}$ denotes the mean of the training labels. Discuss this quantity. What does it mean if R^2 is 1 and especially if R^2 is 0? Can R^2 be negative?

6. Now let us build a non-linear model

   $$h(x) = \exp(a\sqrt{x} + b) \tag{4}$$

   with $a, b \in \mathbb{R}$. This is the same as before plus additionally applying the non-linear input transformation $x \mapsto \sqrt{x}$ to the input data.

   One can view this as the inputs x being mapped to a feature space $\mathcal{Z}$ by the *feature map* $\phi(x) = \sqrt{x}$.

   Build the model and report the mean-squared-error. Then plot the target (on logarithmic scale) and the model output *over the original inputs* (linear scale). That is, the unit of the x-axis should be years.

Compute R^2 for the new model and the transformed labels. Discuss the result in comparison to the previous model.

Deliverables: Source code (main parts also in the report); plot of data and model output; mean-squared error, parameters of regression model, discussion of R^2; counterexample showing the difference when fitting in non-linearly transformed output space; mean-squared error of model with transformed inputs, plot of data and the model given by the second model (note axis scaling in years), comparison of R^2 values of the two different models

References

C. A. Bache, J. W. Serum, W. D. Youngs, and D. J. Lisk. Polychlorinated biphenyl residues: Accumulation in cayuga lake trout with age. *Science*, 177 (4055):1191–1192, 1972.

D. M. Bates and D. G. Watts. *Nonlinear Regression Analysis and Its Applications*, volume 2. Wiley New York, 1988.

Julian S. Huxley and Georges Teissier. Terminology of relative growth. *Nature*, 137(3471):780–781, 1936.