

Project 2 on Machine Learning, deadline November 13 (Midnight)

Data Analysis and Machine Learning FYS-STK3155/FYS4155

Department of Physics, University of Oslo, Norway

Nov 4, 2020

Classification and Regression, from linear and logistic regression to neural networks

The main aim of this project is to study both classification and regression problems by developing our own feed-forward neural network (FFNN) code. We can reuse the regression algorithms studied in project 1. We will also include logistic regression for classification problems and write our own FFNN code for studying both regression and classification problems. The codes developed in project 1, including bootstrap and/or cross-validation as well as the computation of the mean-squared error and/or the R^2 or the accuracy score (classification problems) functions can also be utilized in the present analysis.

The data sets that we propose here are (the default sets)

- Regression (fitting a continuous function). In this part you will need to bring back your results from project 1 and compare these with what you get from your Neural Network code to be developed here. The data sets could be
 1. Either the Franke function or the terrain data from project 1, or data sets your propose.
- Classification. Here you will also need to develop a Logistic regression code that you will use to compare with the Neural Network code. The data set we propose are the so-called [MNIST](#) data set of images representing hand-written numbers from zero to nine. These are discussed intensively in the lecture notes on neural networks, see for example the slides from [week 41](#)

However, if you would like to study other data sets, feel free to propose other sets. What we listed here are mere suggestions from our side. If you opt for

another data set, consider using a set which has been studied in the scientific literature. This makes it easier for you to compare and analyze your results. Comparing with existing results from the scientific literature is also an essential element of the scientific discussion.

In particular, when developing your own Neural Network and Logistic Regression codes for classification problems, the so-called Wisconsin Cancer data (which is a binary problem, benign or malignant tumors) may be studied. You can find more information about this at the [Scikit-Learn site](#) or at the [University of California at Irvine](#).

We will start with a regression problem and we will reuse our codes from project 1 starting with writing our own Stochastic Gradient Descent (SGD) code.

Part a): Write your own Stochastic Gradient Descent code, first step.

In order to get started, we will now replace in our standard ordinary least squares (OLS) and Ridge regression codes (from project 1) the matrix inversion algorithm with our own SGD code. You can choose whether you want to add the momentum SGD optionality or other SGD variants such as RMSprop or ADAGRAD. The lecture notes from [week 40](#) contain more details

Perform an analysis of the results for OLS and Ridge regression as function of the chosen learning rates, the number of mini-batches and epochs as well as algorithm for scaling the learning rate. You can also compare your own results with those that can be obtained using for example **Scikit-Learn**'s various SGD options. Discuss your results. For Ridge regression you need now to study the results as functions of the hyper-parameter λ and the learning rate γ . Discuss your results.

You will need your SGD code for the setup of the Neural Network and Logistic Regression codes.

Part b): Writing your own Neural Network code. Your aim now, and this is the central part of this project, is to write your own Feed Forward Neural Network code implementing the back propagation algorithm discussed in the lecture slides from [week 41](#).

We will focus on a regression problem first and study either the Franke function or terrain data (or both or other data sets) from project 1. Discuss again your choice of cost function.

Write an FFNN code for regression with a flexible number of hidden layers and nodes using the Sigmoid function as activation function for the hidden layers. Initialize the weights using a normal distribution. How would you initialize the biases? And which activation function would you select for the final output layer?

Train your network and compare the results with those from your OLS and Ridge Regression codes from project 1. You should test your results against a similar code using **Scikit-Learn** (see the examples in the above lecture notes from week 41) or **tensorflow/keras**.

Comment your results and give a critical discussion of the results obtained with the Linear Regression code and your own Neural Network code. Compare the results with those from project 1. Make an analysis of the regularization parameters and the learning rates employed to find the optimal MSE and R^2 scores.

A useful reference on the back propagation algorithm is [Nielsen's book](#). It is an excellent read.

Part c): Testing different activation functions. You should now also test different activation functions for the hidden layers. Try out the Sigmoid, the RELU and the Leaky RELU functions and discuss your results. You may also study the way you initialize your weights and biases.

Part d): Classification analysis using neural networks. With a well-written code it should now be easy to change the activation function for the output layer.

Here we will change the cost function for our neural network code developed in parts b) and c) in order to perform a classification analysis.

We will here study the MNIST data set of hand-written numbers as discussed in the lecture notes from [week 41](#). Use the **Softmax** function as activation function. Your code should however also be able to use a binary activation function as well.

To measure the performance of our classification problem we use the so-called *accuracy* score. The accuracy is as you would expect just the number of correctly guessed targets t_i divided by the total number of targets, that is

$$\text{Accuracy} = \frac{\sum_{i=1}^n I(t_i = y_i)}{n},$$

where I is the indicator function, 1 if $t_i = y_i$ and 0 otherwise if we have a binary classification problem. Here t_i represents the target and y_i the outputs of your FFNN code and n is simply the number of targets t_i .

Discuss your results and give a critical analysis of the various parameters, including hyper-parameters like the learning rates and the regularization parameter λ (as you did in Ridge Regression), various activation functions, number of hidden layers and nodes and activation functions.

As stated in the introduction, it can also be useful to study other datasets. In particular, the so-called Wisconsin Cancer data (which is a binary problem, benign or malignant tumors) may be studied. You find more information about this at the [Scikit-Learn site](#) or at the [University of California at Irvine](#).

Again, we strongly recommend that you compare your own neural Network code for classification and pertinent results against a similar code using **Scikit-Learn** or **tensorflow/keras** or **pytorch**.

Part e): Write your Logistic Regression code, final step. Finally, we want to compare the FFNN code we have developed with Logistic regression,

that is we wish to compare our neural network classification results with the results we can obtain with another method.

Define your cost function and the design matrix before you start writing your code. Write thereafter a Logistic regression code using your SGD algorithm. Study the results as functions of the chosen learning rates. Add also an l_2 regularization parameter λ . Compare your results with those from your FFNN code as well as those obtained using **Scikit-Learn**'s logistic regression functionality.

The weblink here <https://medium.com/ai-in-plain-english/comparison-between-logistic-regression-and-ffnn-using-the-mnist-data-set> logistic regression and FFNN using the MNIST data set. You may find several useful hints and ideas from this article.

Part f) Critical evaluation of the various algorithms. After all these glorious calculations, you should now summarize the various algorithms and come with a critical evaluation of their pros and cons. Which algorithm works best for the regression case and which is best for the classification case. These codes can also be part of your final project 3, but now applied to other data sets.

Background literature

1. The text of Michael Nielsen is highly recommended, see [Nielsen's book](#). It is an excellent read.
2. The textbook of [Trevor Hastie, Robert Tibshirani, Jerome H. Friedman, The Elements of Statistical Learning, Springer](#), chapters 3 and 7 are the most relevant ones for the analysis here.
3. [Mehta et al, arXiv 1803.08823, A high-bias, low-variance introduction to Machine Learning for physicists, ArXiv:1803.08823](#).

Introduction to numerical projects

Here follows a brief recipe and recommendation on how to write a report for each project.

- Give a short description of the nature of the problem and the eventual numerical methods you have used.
- Describe the algorithm you have used and/or developed. Here you may find it convenient to use pseudocoding. In many cases you can describe the algorithm in the program itself.
- Include the source code of your program. Comment your program properly.
- If possible, try to find analytic solutions, or known limits in order to test your program when developing the code.

- Include your results either in figure form or in a table. Remember to label your results. All tables and figures should have relevant captions and labels on the axes.
- Try to evaluate the reliability and numerical stability/precision of your results. If possible, include a qualitative and/or quantitative discussion of the numerical stability, eventual loss of precision etc.
- Try to give an interpretation of your results in your answers to the problems.
- Critique: if possible include your comments and reflections about the exercise, whether you felt you learnt something, ideas for improvements and other thoughts you've made when solving the exercise. We wish to keep this course at the interactive level and your comments can help us improve it.
- Try to establish a practice where you log your work at the computerlab. You may find such a logbook very handy at later stages in your work, especially when you don't properly remember what a previous test version of your program did. Here you could also record the time spent on solving the exercise, various algorithms you may have tested or other topics which you feel worthy of mentioning.

Format for electronic delivery of report and programs

The preferred format for the report is a PDF file. You can also use DOC or postscript formats or as an ipython notebook file. As programming language we prefer that you choose between C/C++, Fortran2008 or Python. The following prescription should be followed when preparing the report:

- Use Canvas to hand in your projects, log in at <https://www.uio.no/english/services/it/education/canvas/> with your normal UiO username and password.
- Upload **only** the report file or the link to your GitHub/GitLab or similar type of repos! For the source code file(s) you have developed please provide us with your link to your GitHub/GitLab or similar domain. The report file should include all of your discussions and a list of the codes you have developed. Do not include library files which are available at the course homepage, unless you have made specific changes to them.
- In your GitHub/GitLab or similar repository, please include a folder which contains selected results. These can be in the form of output from your code for a selected set of runs and input parameters.

Finally, we encourage you to collaborate. Optimal working groups consist of 2-3 students. You can then hand in a common report.

Software and needed installations

If you have Python installed (we recommend Python3) and you feel pretty familiar with installing different packages, we recommend that you install the following Python packages via **pip** as

1. pip install numpy scipy matplotlib ipython scikit-learn tensorflow sympy pandas pillow

For Python3, replace **pip** with **pip3**.

See below for a discussion of **tensorflow** and **scikit-learn**.

For OSX users we recommend also, after having installed Xcode, to install **brew**. Brew allows for a seamless installation of additional software via for example

1. brew install python3

For Linux users, with its variety of distributions like for example the widely popular Ubuntu distribution you can use **pip** as well and simply install Python as

1. sudo apt-get install python3 (or python for python2.7)

etc etc.

If you don't want to install various Python packages with their dependencies separately, we recommend two widely used distributions which set up all relevant dependencies for Python, namely

1. [Anaconda](#) Anaconda is an open source distribution of the Python and R programming languages for large-scale data processing, predictive analytics, and scientific computing, that aims to simplify package management and deployment. Package versions are managed by the package management system **conda**
2. [Enthought canopy](#) is a Python distribution for scientific and analytic computing distribution and analysis environment, available for free and under a commercial license.

Popular software packages written in Python for ML are

- [Scikit-learn](#),
- [Tensorflow](#),
- [PyTorch](#) and
- [Keras](#).

These are all freely available at their respective GitHub sites. They encompass communities of developers in the thousands or more. And the number of code developers and contributors keeps increasing.