Processing Biological Data

Prof. Dr. Volkhard Helms — Chair for Computational Biology Saarland University
Dr. Pratiti Bhadra — Tutor
Summer Semester 2020

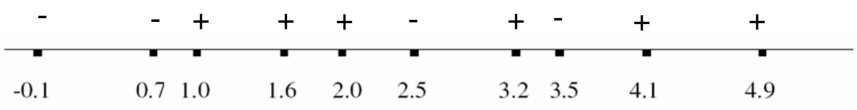
Assignment Sheet 4
Deep Learning
Due: 2nd July, 2020 10:15 am

Submission

1. Submit your source code and solution a single PDF attachment to pratiti.bhadra@bioinformatik.uni-saarland.de.
2. Use Python for coding and document your source code.
3. Subject of the email should be Assignment4-"your name"
4. Online tutorials
   (a) full course: https://nptel.ac.in/courses/106/106/106106184/
   (b) basic introduction: https://www.youtube.com/watch?v=O5xeyoRL95U
   (c) programming: https://machinelearningmastery.com/tutorial-first-neural-network-python-keras/

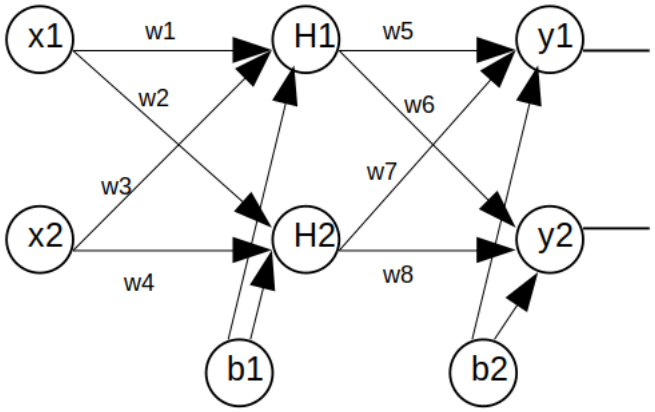
Please feel free to contact me for any clarifications either via email or you can reach me in building E2.1, Room 3.03 (preferably between 3 pm and 4 pm).

1. **Exercise 4.1: Basics of Deep learning and neural network** [50 points]
   (a) Briefly explain what is meant by over-fitting. How can one avoid over-fitting in deep neural networks? (5 point)
   (b) What is meant by "dropout" in relation to neural networks? Which of the following statement is true for dropout? (5 points)
      i. Dropout gives a way to approximate by combining many different architectures
      ii. Dropout can help preventing overfitting
      iii. Dropout prevents that the hidden layers co-adapt.
   (c) What is an activation function of neural networks? What is the purpose of the activation function in neural networks? What are the advantages of ReLu function over Sigmoid function? (5 points)
   (d) Cross validation: Carry out leave-one-out cross-validation (LOOCV) in a simple classification problem. Consider the following dataset with one real-valued input x (numbers on the line in the figure) and one binary output y (negative and positive sign). We are going to use k-NN with Euclidean distance to predict $\hat{y}$ for x.

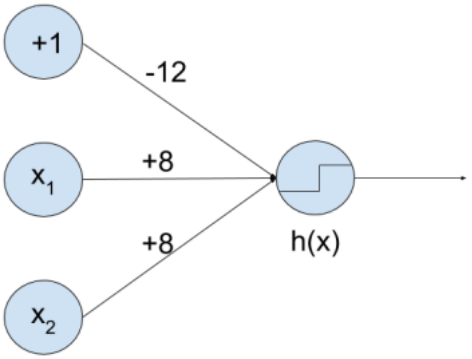


What is the LOOCV error of 1-NN on this dataset? Give your answer as the total number of misclassifications. (5 points)

(e) With respect to the figure shown at the end of this problem,determine the values at hidden and output layers if x1 = 0.05, x2 = 0.10, w1 = 0.15, w2 = 0.20, w3 = 0.25, w4 =0.30, w5 = 0.40, w6 = 0.45, w7 = 0.50, w8 = 0.55, b1= 0.35 and b2=0.60. Assume that the activation function is sigmoid (logistic function $f(x) = \frac{L}{1+e^{-x}}$). Determine the total error (the mean squared error) if the target (or actual) outputs are $y1^T = 0.01$ and $y2^T = 0.99$. What is the updated weight of w5 after one iteration of the back-propagation algorithm on this example? Assume that the learning rate is 0.5. (10+5+10 = 25 points)



(f) Consider the following neural networks which take two binary valued inputs x_1 , $x_2 \in \{0, 1\}$ and assume that the activation function is the threshold function (h(x) = 1 if x > 0; 0 otherwise). Which logical functions does the network compute? (5 points)



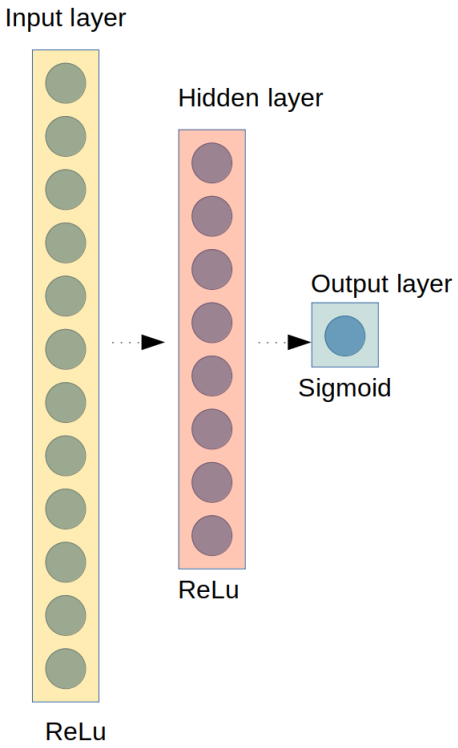
2. **Exercise 4.2: Programming: Multilayer perception classification [50 points]**
Download data files (red-wine.csv and white-wine.csv) from supplementary. Pleases look at different python packages (keras, scikit-learn, seaborn, matplotlib.pyplot, pandas and tensorflow etc.)

(a) The sulfates are one component of wine. Sulfate ions can cause people to have headaches. I'm wondering if this influences the quality of the wine. Please illustrate the relation or dependency between 'sulphates' and 'quality' using a figure or plot. Is there any difference between "red" and "white" wine? The figure should have axis-level and legend. (7 points)

(b) Describe the correlation matrix and its importance? Plot the correlation matrix of variables (or features) of the wine dataset. the correlation coefficient should be reflect on plot. Do

you get any important information from this plot which may help you to build an efficient classifier? (8 points)

(c) Build a neural network architecture as shown in the figure below to predict the class of wine (red = '1' and white = '0'). Use 5-fold cross validation. Present accuracy from each fold and evaluate the performance of your final model by accuracy, F1-score, recall, precision. Standardize the wine data (zero mean and unit variance) before classification. Set batch size and epoches to 10 and 150, receptively. Use "binary cross entropy loss" as loss function and ADAM optimizer. (25 points)



(d) What is your opinion on the wine dataset? Is there any way to improve the performance of your model using some data processing methods? If yes, then please provide the performance (accuracy,f1-score,recall,precision) of the improved model and explain the reason behind the improvement. (10 points)