

ECS 171: Homework Set 2

Instructor: Ilias Tagkopoulos

TAs: Ameen Eetemadi, Nicholas Joodi, Cheng-En Tan

{eetemadi, npjoodi, cetan}@ucdavis.edu

Homework is due on October 31, 2017

General Instructions: The homework should be submitted electronically through Canvas. Each submission should be a zip file that includes the following: (a) a report in pdf format ("report_HW2.pdf") that includes your answers to all questions, plots, figures and any instructions to run your code, (b) the python code files. Please note: (a) do not include any other files, for instance files that we have provided such as datasets, (b) each function should be written in a separate file, with the appropriate remarks in the code so it is generally understandable (what it does, how it does it), (c) do not use any toolbox unless is it explicitly allowed in the homework description. Shared/copied code from any source is not allowed, as it is considered plagiarism. There is a 20% penalty per day for late homeworks.

1 WHERE DID THE BAKER GO? [100PT]

In this exercise, you will build a classifier that can find the localization site of a protein in yeast, based on 8 attributes (features). You will use the “Yeast” dataset (“yeast.data” file; 1484 proteins, 8 features, 10 different classes; no missing data) that is available in the UCI Machine Learning Repository:

<https://archive.ics.uci.edu/ml/machine-learning-databases/yeast/>

Perform and report (code and results) the following:

1. Construct a 3-layer artificial neural network (ANN) and specifically a feed-forward multi-layer perceptron to perform multi-class classification. The hidden layer should have 3 nodes. Split your data into a random set of 70% of the samples as the training set and the rest 30% as the testing set. For training, use stochastic gradient descent with back-propagation. Please note that you will never train with the testing set; the ANN will only take into account the training set for updating the weights. For the most popular class "CYT", provide 2 plots: (I) weight values per iteration for the last layer (3 weights and bias), (II) training and test error per iteration. [30pt]
2. Now re-train the ANN with all your data (all 1484 samples). What is your training error? Provide the final activation function formula for class "CYT" after training. [10pt]
3. For the ANN that you have built (3 layers, 1 hidden layer, 3 hidden nodes) calculate the first round of weight update with back-propagation with paper and pencil for all weights but for only the first sample. Confirm that the numbers that you calculated are the same with those produced by the code and provide both your calculations and the code output. Provide both calculations made by hand (scanned image is fine) and corresponding output from the program that shows that both are in agreement. [30pt]
4. Increase the number of hidden layers from 1 to 2 and then to 3. Then increase the number of hidden nodes per layer from 3 to 6, then to 9 and finally to 12. Create a 3x4 matrix with the number of hidden layers as rows and the number of hidden nodes per layer as columns, with each element (cell) of the matrix representing the testing set error for that specific combination of layers/nodes. What is the optimal configuration? What you find the relationship between these attributes (number of layers, number of nodes) and the generalization error (i.e. error in testing data) to be? [25pt]
5. Which class does the following sample belong to?[5pt]
Unknown Sample 0.49 0.51 0.52 0.23 0.55 0.03 0.52 0.39
6. Can you come up with a quantitative measure of uncertainty for each classification? What is the uncertainty for the unknown sample of the previous question? Justify your assumptions and method [5pt bonus]

2 PROJECT

The following project tasks will need to be completed. A Joke Submitter ID was/will be emailed out to each student. This identifier will be needed as you complete the below tasks.

- **Phase I - Deadline: October 23rd**

- Every student will need to submit the following Google forms (only one submission per Google form):

- * Jokes Google form
- * User Profile

- **Phase II - Deadline: October 24rd (Data team only)**

- The Data Management team will need to populate a Google form with jokes
- The managers from the Data Management teams will be given collaborator access to each form. These managers can then give access to any additional member from their team. The access will give an individual the ability to update a form, exports the data, etc.
- **NOTE: collaborators of all forms are prohibited from turning on the flag to collect the email address in the response. This requirement is to protect an individual's privacy**
- Please use the following url to create the ratings Google form: Ratings Survey
- Each joke will be one question. Use the example joke as a format for how the jokes will be loaded. Please add a unique ID to each joke

- **Phase III - Deadline: October 30th**

- This phase cannot be started until Phase II is complete
- Every student needs to complete the following Google form Ratings Survey

- **Phase IV - Deadline: November 2nd (Data team only)**

- The data team presents the quality controlled final Dataset and how to access it at the project presentation

GOOD LUCK!

