COMENIUS UNIVERSITY IN BRATISLAVA
FACULTY OF MATHEMATICS, PHYSICS AND INFORMATICS
DEPARTMENT OF APPLIED INFORMATICS

Neural Networks

Project 2b – Visualizing Data

April 15, 2025

OVERVIEW

Task: Implement a self-organizing map (wiki) with rectangular topology and use it to visualize the *Seeds data set* from the UCI Machine Learning Repository in two dimensions.
[The data and additional information can be found on the original homepage. Alternatively: a simple text file having 210 lines (samples) and 8 columns (7 features + class).] *Do not* use the class column for training.
Deadline: May 11, 23:59
Late submissions are penalized by -2 points each day. **It is not possible to submit a project more than 5 days after the deadline.**

MODEL

To achieve better results, it is important to choose good model to train the network. In addition to exercises, implement L_2 and L_1 metric defining the neighbourhood in the grid space. Also implement the continuous update, where each neuron is always updated (even though some with a very small weight).
Briefly investigate the effect of these hyperparameters and choose the best performing model, for which report the following:

REPORT

- Include: how quantization error[1] decreases throughout training **[plot]**

[1] Quantization error = average distance of data point x_i to its best matching neuron c_j:
$E = 1/n \sum_i \min_j \|x_i - c_j\|$

  - also note that the quantization error decreases with the number of neurons. However, it makes little sense to have more neurons than data. Choose the grid size reasonably.

- how the average adjustment[2] of neuron positions changes during training **[plot]**
- which neurons are activated by which classes of input **[diagram]**
  1. single rectangular graph showing class membership for relevant neurons
  2. *sanity check*: the classes should not collide much in a properly trained SOM
- how the value of each of the seven attributes changes across the map **[heatmap]**
  - one rectangular heatmap for each attribute
- plotting the U-matrix, distances between adjacent neurons **[heatmap]**

BONUS

Examine whether a self-organizing map can be used as a successful classifier.

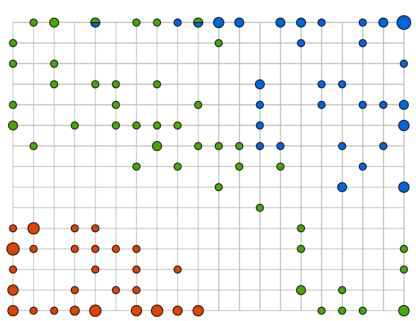
1. split the dataset (210 samples) into a training (150 samples) and a testing set (60 samples)
   - make sure that the classes are equally represented in both of them
2. train the map on the features (but not classes) of the training set
3. assign a class to each neuron of the map
   - most prevalent class of inputs corresponding to that neuron
4. test: for each test input, find the best-matching neuron and output its class

Investigate how to select the map parameters to maximize the testing accuracy. Report classification accuracy and the confusion matrix.

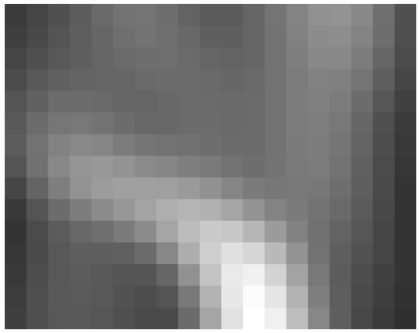
[2] Average amount of adjustment of a neurons at a time t - let $\Delta_{c_j}(t) = c_j(t) - c_j(t-1)$:
$A(t) = 1/k \sum_j \|\Delta_{c_j}(t)\|$

EXAMPLE

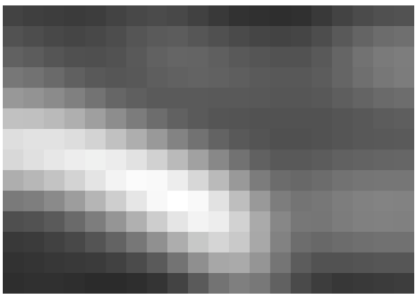
Diagrams from a self-organizing map of size 20x15 (not perfectly) trained on the Iris dataset:



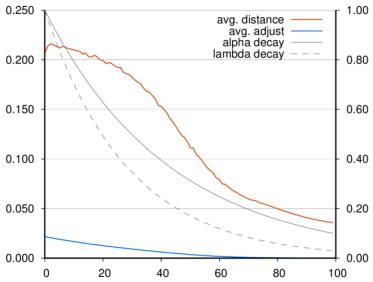
counts



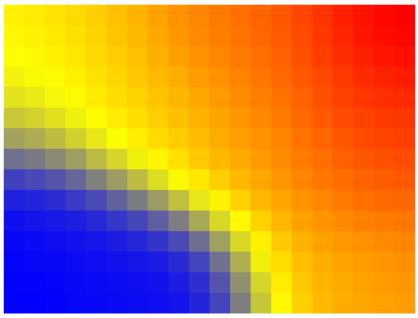
distancesHoriz



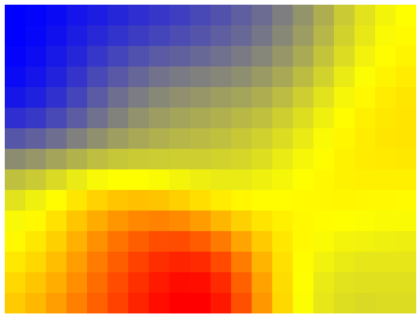
distancesVert



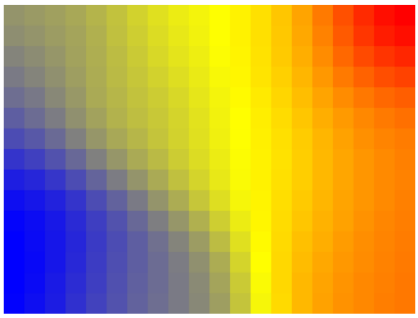
report



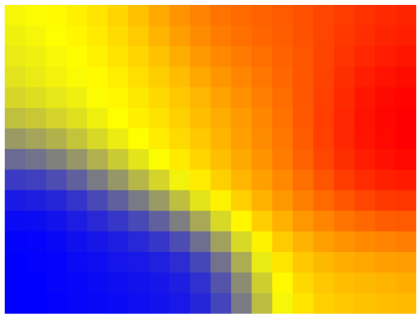
coord-2



coord-1



coord-0



coord-3