University of Waterloo

ECE 657A: Data and Knowledge Modeling and Analysis

Winter 2024

Assignment 3: Data Cleaning and Dimensionality Reduction

Due: Apr 8th, 2024, 11:59pm

Overview

Assignment Type: Done in groups of up to three students.

Hand in: One report (PDF) or python notebook per group, via the LEARN DropBox. Also submit the code / scripts needed to reproduce your work. (If you are submitting by PDF, if you don't know LATEX should try to use it, it's good practice and it will make the project report easier)

Objective: To gain experience on the use of Neural Networks and Fuzzy Inference Systems.

Problem 1:

Using a feedforward back-propagation neural network that contains a single hidden layer (with a variable number of hidden nodes each having an activation function of the logistic form), investigate the outcome of the neural network for the following mappings:

- $f(x) = x * \sin(6\pi x) * \exp(-x^2)$ where $x \in [-1, 1]$
- $f(x) = \exp(-x^2) * \arctan(x) * \sin 4\pi x$ where $x \in [-2, 2]$

For each function, create two sets of input/output data, one for training and the other for testing (these will be random values within the interval of the variable x). The dataset you create should 1000 for each function. You can choose the ratio as 70% of the data for training and 30% of the data for testing.

a) Investigate the behavior of the training and testing errors of the output when the number of hidden nodes is fixed (e.g., 3 hidden nodes). You can vary the number for the sample data from 20 samples to 200 samples.

b) Investigate the effect of the number of hidden nodes on the training and testing error output for a given fixed number of training samples (e.g., 100 training samples). Try to find the optimal number of hidden nodes to lead to the smallest training and testing error.

c) Make qualitative and quantitative deductions considering these simulations. Produce plots showing the training and testing errors for every simulation.

Problem 2:

We need to train a MLP network for obtaining the output of the following two-to-one mapping function.

$$f(x_1, x_2) = \sin(2\pi x_1) * \cos(0.5\pi x_2) * \exp(-x_1^2)$$

a) Set up two sets of data, one for network training and the other for testing (70% for training and 30% for testing). The total number of input-output data is 500 and is obtained by randomly varying the input variables (x_1, x_2) within the interval [-1,1] by [-4 4].

b) First, fix the number of hidden neurons to 4 (double of the number of input nodes) and analyze the performance of the obtained network (training and testing output errors). Use one hidden layer for the exercise.

c) Analyze the performance of the network with more and then with fewer hidden nodes (2, 6, 8, 12, 20). Find the best number of hidden neurons leading to the least training and testing network error and discuss.

Problem 3:

Suppose that the state of "fast speed" of a machine is denoted by the fuzzy set F with membership function $\mu_F(v)$. Then the state of "very fast speed", where the *linguistic hedge* "very" has been incorporated, may be represented by $\mu_F(v\text{-}v_o)$ with $v_o > 0$. Also, the state "presumably fast speed", where the linguistic hedge "presumably" has been incorporated, may be represented by $\mu_F^2(v)$.

(a) Discuss the appropriateness of the use of these membership functions to represent the respective linguistic hedges.

(b) In particular, if

$$F = \{\frac{0.1}{10}, \frac{0.3}{20}, \frac{0.6}{30}, \frac{0.8}{40}, \frac{1.0}{50}, \frac{0.7}{60}, \frac{0.5}{70}, \frac{0.3}{80}, \frac{0.1}{90}\}$$

in the discrete universe $V = \{0, 10, 20, \ldots, 190, 200\}$ rev/s and $v_o = 50$ rev/s, determine the membership functions of "very fast speed" and "presumably fast speed".

(c) Suppose that power p is given by the relation (crisp)

$$p = v^2$$

For the given fast speed (fuzzy) F, determine the corresponding membership function for power. Compare/contrast this with "presumably fast speed".

Note: for c, use presumably fast from b, and the comparison can be linguistic.

Problem 4:

Sketch the membership function $\mu_A(x) = e^{-\lambda(x-a)^n}$ for $\lambda = 2$, $n = 2$, and $a = 3$ for the support set $S = [0,6]$. On this sketch separately show the shaded areas that represent the following fuzziness measures:

(a) $M_1 = \int_S f(x)dx$ where $\begin{aligned} f(x) &= \mu_A(x) && \text{for } \mu_A(x) \le 0.5 \\ &= 1 - \mu_A(x) && \text{for } \mu_A(x) > 0.5 \end{aligned}$

(b) $M_2 = \int_S \left| \mu_A(x) - \mu_{A_{1/2}}(x) \right| dx$

where $\mu_{A_{1/2}}$ is the $\alpha-$cut of $\mu_A(x)$ for $\alpha = 1/2$

(c) $M_3 = \int_S \left| \mu_A(x) - \mu_{\bar{A}}(x) \right| dx$

where $\bar{A}$ is the complement of the fuzzy set A. Evaluate the values of M_1, M_2 and M_3 for the given membership function.

i. Establish relationships between M_1, M_2 and M_3.
ii. Indicate how these measures can be used to represent the degree of fuzziness of a membership function.
iii. Compare your results with the case $\lambda = 1$, $a = 3$, and $n = 2$ for the same support set, by showing the corresponding fuzziness measures on a sketch of the new membership function.

Problem 5:

Show that max [0, x+ y - 1] is a t-norm. Also, determine the corresponding t-conorm (i.e., s-norm). Hint: Show that the non-decreasing, commutative, and associative properties, and the boundary conditions are satisfied.

Problem 6:

(a) Consider the membership function $\mu_A(x) = e^{-\lambda|x-a|^n}$, for a fuzzy set A.
Interpret the meaning of the parameters a, λ and n. In particular, discuss how (1) fuzziness and (2) a fuzzy adjective or fuzzy modifier such as "very" or "somewhat" of a fuzzy state may be represented using these parameters.

(b) Using the general membership function expression used in part (a), give an analytical representation for temperature inside a living room, that has the three fuzzy states "cold, comfortable, and hot". You must give appropriate numerical values for the parameters of the analytical expression.