Machine Learning Homework 3

Subigya Paudel, Maulik Chhetri, Mahiem Agrawal

February 2021

1 Exercise 1. (Modelling inputs / outputs)

1.1 Briefly describe the data set and all involved variables in your own words. If some information is missing on the UCI Repository site, do your own search for these details

1. **Stat log (Shuttle) Data set** This data set consists of data points each of which has 9 numerical attributes. The last attribute is the class to which the data point is assigned to, and this value ranges from 1 to 7 and signifies the one out of the seven classes which the data point belongs to which are

   (a) Rad Flow
   (b) Fpv close
   (c) Fpv Open
   (d) High
   (e) Bypass
   (f) Bpv Close
   (g) Bpv Open

2. **Computer Hardware Data set** This data consist of 10 attributes where 9 are the input parameters and 1 is the output parameter. two of the attributes are non-numeric data and the rest are numeric continuous data. The non-numeric data are vendor name and model name. The continuous numeric attributes are as follows:

   (a) vendor name: 30 (adviser, amdahl,apollo, basf, bti, burroughs, c.r.d, cambex, cdc, dec, dg, formation, four-phase, gould, honeywell, hp, ibm, ipl, magnuson, microdata, nas, ncr, nixdorf, perkin-elmer, prime, siemens, sperry, sratus, wang)

   (b) Model Name: many unique symbols

(c) MYCT: machine cycle time in nanoseconds (integer)

(d) MMIN: minimum main memory in kilobytes (integer)

(e) MMAX: maximum main memory in kilobytes (integer)

(f) CACH: cache memory in kilobytes (integer)

(g) CHMIN: minimum channels in units (integer)

(h) CHMAX: maximum channels in units (integer)

(i) PRP: published relative performance (integer)

(j) ERP: estimated relative performance from the original article (integer)

1.2 Model the data set via input and output random variables / vector

1. The statlog (shuttle) data set can be modelled by the following set:

   Input Vector:

   $$X : \Omega \to R^8$$

   Output Vector:

   $$G : \Omega \to \{1, 2, 3, 4, 5, 6, 7\}$$

   So the data set can be described by the following set

   $$\{(x_i, y_i)\}_{i=1}^{58000}$$

2. Computer Hardware Data Modelling:

   Input Vector:

   $$X : \Omega \to R^9$$

   Output Vector:

   $$Y : \Omega \to R$$

   So the data set can be described by the following set

   $$\{(x_i, y_i)\}_{i=1}^{209}$$

   We take the 9 models, excluding the model name since it contains undefined number of model names unlike the vendor name, which had fixed types of values,30, and can be associated with $1 \dots 30$

1.3 Formulate a question that can be solved using machine learning on this data set and give the type of machine learning that will allow to answer the question.

1. **Shuttle Dataset**
   The data in the data set can be used to classify the data to the seven classes based on the input parameters. Since there are labels/outputs assigned to the training, the application of the the data in this data set is an example of supervised and classification learning.

2. **Computer Hardware Dataset**
   Predict the estimated relative performance of any random hardware given the aforementioned attributes.
   Regression will allow to answer the question. Since the ERP value is continuous, we require a predictor function $f(x)$, where $x \in R^8$

2 Exercise 2. SPAM e-mail representation

2.1 Description of the variables in the spam email

The input vector for this task would consist of 57 attributes.

1. Attribute [0 to 48]: Number from [0-100] which would signify the percent of each of the 48 keywords in that specific email.

2. Attribute [49 to 54]: Percentage of characters in the e-mail that match CHAR, i.e. 100 * (number of CHAR occurrences) / total characters in e-mail

3. Attribute 55: The average length of uninterrupted sequences of capital letters.

4. Attribute 56: Length of longest uninterrupted sequence of capital letters

5. Attribute 57: Total number of capital letters in the e-mail

The following key words has been used:
make, address, all, 3d, our, over, remove, internet, order, mail, receive, will, people, report, addresses, free, business, email, you, credit, your, font, 000, money, hp, hpl, george, 650, lab, labs, telnet, 857, data, 415, 85, technology, 1999 , parts, pm, direct, cs, meeting, original, project, re, edu, table, conference,
The following charcaters have been used:
; ([! $ #

Mathematically;

$$X : \Omega \to R^{57}$$

The output is a simple classification of whether an email is spam or not.We denote 0 as "spam" ans 1 as "not spam"

2.2 Alternatives

An alternative that can be used instead of this dataset is the SMS Spam Collection Data Set.
It can be found in https://archive.ics.uci.edu/ml/datasets/sms+spam+collection.