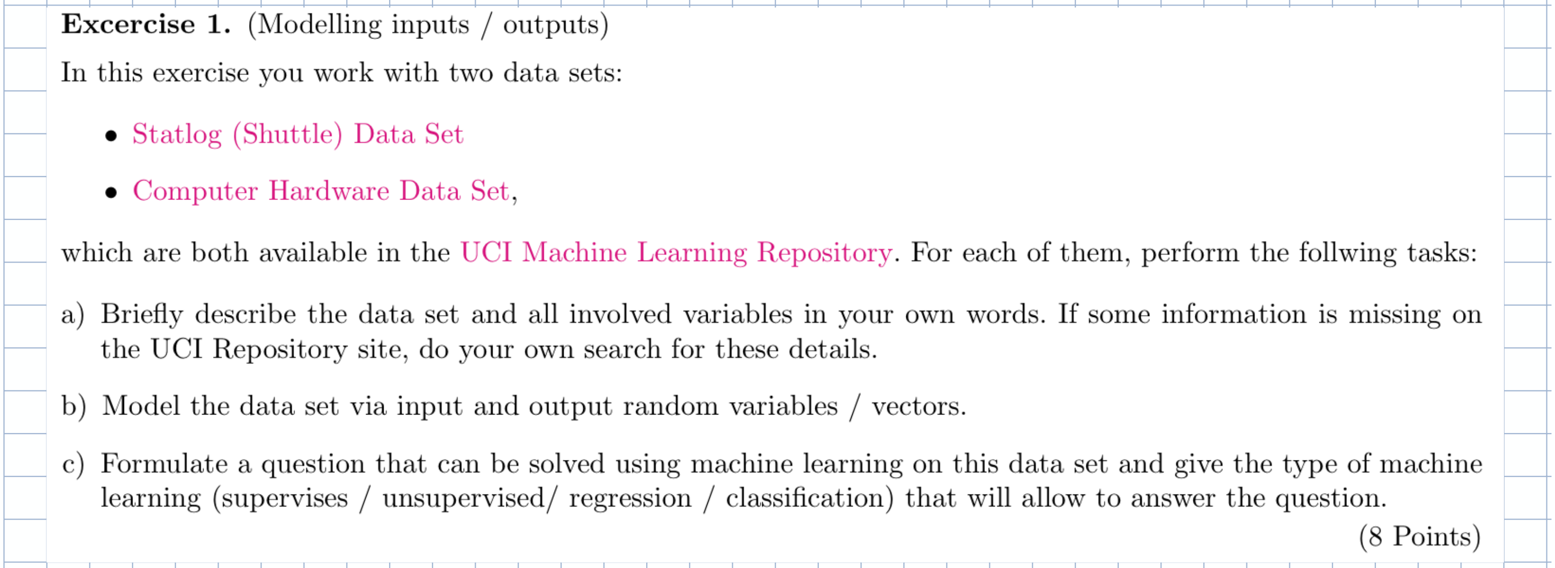
ASSIGNMENT 3 SOLUTION

Done By: Nayan Man Singh Pradhan



Excercise 1. (Modelling inputs / outputs)

In this exercise you work with two data sets:

- Statlog (Shuttle) Data Set
- Computer Hardware Data Set,

which are both available in the UCI Machine Learning Repository. For each of them, perform the follwing tasks:

a) 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.

b) Model the data set via input and output random variables / vectors.

c) Formulate a question that can be solved using machine learning on this data set and give the type of machine learning (supervises / unsupervised/ regression / classification) that will allow to answer the question.

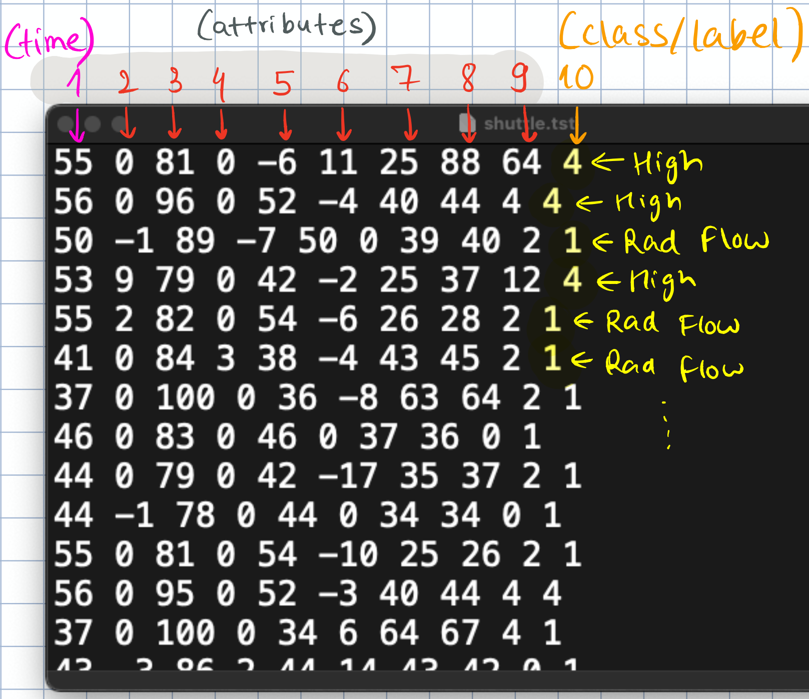
(8 Points)

- Statlog (shuttle) Data Set

a) The Statlog (shuttle) Data Set is a multivariate data set consisting of 9 attributes (columns) and 58,000 instances (rows) of integer values provided by Jason Catlett from NASA's shuttle datasets. According to the Data Set documentation, the shuttle dataset contains 9 attributes, all of which are numerical values. The first attribute is "time", and the last column is the class / label. There are in total 7 classes / labels. They are:

1. Rad Flow
2. Fpv Close
3. Fpv Open
4. High
5. Bypass
6. Bpv Close
7. Bpv Open

The documentation specifies that approximately 80% of the data belongs to class 1. I have attached a labelled screenshot of the downloaded data file for clarification.



```
55 0 81 0 -6 11 25 88 64 4
56 0 96 0 52 -4 40 44 4 4
50 -1 89 -7 50 0 39 40 2 1
53 9 79 0 42 -2 25 37 12 4
55 2 82 0 54 -6 26 28 2 1
41 0 84 3 38 -4 43 45 2 1
37 0 100 0 36 -8 63 64 2 1
46 0 83 0 46 0 37 36 0 1
44 0 79 0 42 -17 35 37 2 1
44 -1 78 0 44 0 34 34 0 1
55 0 81 0 54 -10 25 26 2 1
56 0 95 0 52 -3 40 44 4 4
37 0 100 0 34 6 64 67 4 1
43 -3 86 2 44 14 43 42 0 1
```

b) Input :

$$X : \Omega \rightarrow \mathbb{R}^9$$

Output

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

c) Supervised Classification Task :

Given 9 features/attributes of the Statlog shuttle, predict the class (rad flow, fpv close, fpv open, ...) of the shuttle.

- Computer Hardware Data Set

a) The Computer Hardware Data Set is a multivariate data set consisting of 9 attributes (columns) and 209 instances (rows) of integer values provided by Phillip Ein-Dor and Jacob Feldmesser. According to the Data Set documentation, out of the 9 attributes, 2 are non-predictive, 6 are predictive and 1 is goal field. The last column is the estimated relative performance estimated by the authors using linear regression method. There are no missing attributes. A description of the attributes :

1. Vendor name : Total of 30 vendor names : adviser, amdahl, apollo,...

2. Model name : Many unique symbols

3. MYCT : Machine Cycle time in nanoseconds (integer)

4. MMIN : Minimum main memory in kilobytes (integer)

5. MMAX : Maximum main memory in kilobytes (integer)

6. CACH : Cache memory in kilobytes (integer)

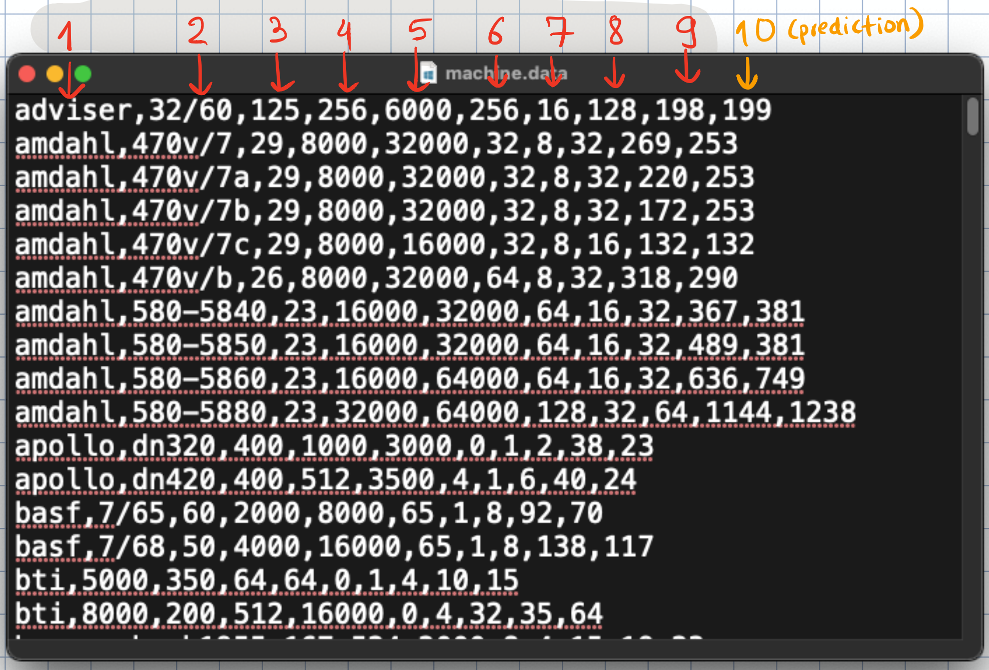
7. CHMIN : Minimum channels in units (integer)

8. CHMAX : Maximum channels in units (integer)

9. PRP : Published relative performance (integer)

10. ERP : Estimated relative performance by the authors using linear regression (integer)

(attributes)



```
adviser,32/60,125,256,6000,256,16,128,198,199
amdahl,470v/7,29,8000,32000,32,8,32,269,253
amdahl,470v/7a,29,8000,32000,32,8,32,220,253
amdahl,470v/7b,29,8000,32000,32,8,32,172,253
amdahl,470v/7c,29,8000,16000,32,8,16,132,132
amdahl,470v/b,26,8000,32000,64,8,32,318,290
amdahl,580-5840,23,16000,32000,64,16,32,367,381
amdahl,580-5850,23,16000,32000,64,16,32,489,381
amdahl,580-5860,23,16000,64000,64,16,32,636,749
amdahl,580-5880,23,32000,64000,128,32,64,1144,1238
apollo,dn320,400,1000,3000,0,1,2,38,23
apollo,dn420,400,512,3500,4,1,6,40,24
basf,7/65,60,2000,8000,65,1,8,92,70
basf,7/68,50,4000,16000,65,1,8,138,117
bti,5000,350,64,64,0,1,4,10,15
bti,8000,200,512,16000,0,4,32,35,64
```

b) Input:

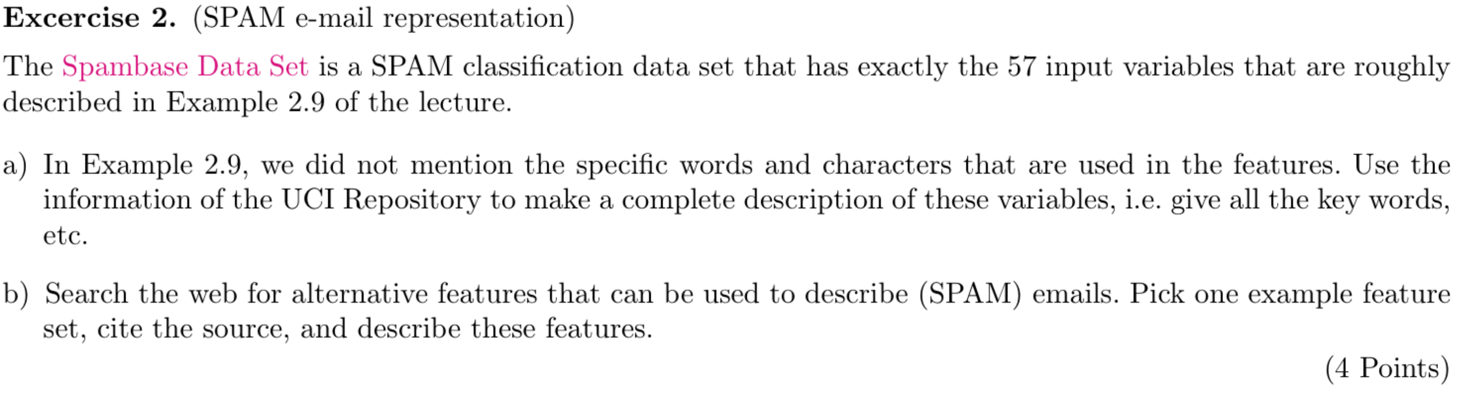
$$X : \Omega \rightarrow \mathbb{R}^9$$

Output:

$$Y : \Omega \rightarrow \mathbb{R}^K_{\geq 0}$$

c) Supervised Regression Task:

Given 9 features/attributes of the hardware of a computer, predict the estimated relative performance of the computer.



Excercise 2. (SPAM e-mail representation)

The Spambase Data Set is a SPAM classification data set that has exactly the 57 input variables that are roughly described in Example 2.9 of the lecture.

a) In Example 2.9, we did not mention the specific words and characters that are used in the features. Use the information of the UCI Repository to make a complete description of these variables, i.e. give all the key words, etc.

b) Search the web for alternative features that can be used to describe (SPAM) emails. Pick one example feature set, cite the source, and describe these features.

(4 Points)

a)
The provided Spambase Data Set has exactly 57 input variables. A complete description of the variables (key words) used in order to classify the emails are as following:

1. Frequency of word: "make"
2. Frequency of word: "address"
3. Frequency of word: "all"
4. Frequency of word: "3d"
5. Frequency of word: "our"
6. Frequency of word: "over"
7. Frequency of word: "remove"
8. Frequency of word: "internet"
9. Frequency of word: "order"
10. Frequency of word: "mail"
11. Frequency of word: "receive"
12. Frequency of word: "will"
13. Frequency of word: "people"
14. Frequency of word: "report"
15. Frequency of word: "addresses"
16. Frequency of word: "free"
17. Frequency of word: "business"
18. Frequency of word: "email"
19. Frequency of word: "you"
20. Frequency of word: "credit"
21. Frequency of word: "your"
22. Frequency of word: "font"
23. Frequency of word: "000"
24. Frequency of word: "money"
25. Frequency of word: "hp"
26. Frequency of word: "hpl"
27. Frequency of word: "george"
28. Frequency of word: "650"
29. Frequency of word: "lab"
30. Frequency of word: "labs"
31. Frequency of word: "telnet"
32. Frequency of word: "857"
33. Frequency of word: "data"
34. Frequency of word: "415"
35. Frequency of word: "85"
36. Frequency of word: "technology"
37. Frequency of word: "1999"

38. Frequency of word: "parts"
39. Frequency of word: "pm"
40. Frequency of word: "direct"
41. Frequency of word: "cs"
42. Frequency of word: "meeting"
43. Frequency of word: "original"
44. Frequency of word: "project"
45. Frequency of word: "re"
46. Frequency of word: "edu"
47. Frequency of word: "table"
48. Frequency of word: "conference"
49. Frequency of character: ";"
50. Frequency of character: "("
51. Frequency of character: "["
52. Frequency of character: "!"
53. Frequency of character: "$"
54. Frequency of character: "#"
55. Average length of uninterrupted sequences of capital letters
56. Length of longest uninterrupted sequence of capital letters
57. Total number of capital letters in the email
58. denotes whether email is spam (1) or not spam (0)

b)
Source: https://www.emerald.com/insight/content/doi/10.1108/EL-07-2019-0181/full/pdf?title=a-feature-centric-spam-email-detection-model-using-diverse-supervised-machine-learning-algorithms

The feature set:

- Number of words in email: The total number of words in the email
- Number of URLs: The total number of URLs in the email
- Number of repetitive words: Total number of words that have been repeated
- Number of unique words: Total number of words that have not been repeated
- Number of attachments: Total number of attachments in the email
- Number of co-occurring words: Total number of same words occurring together
- Number of capitalized words: Total number of capitalized words
- Number of nouns and pronouns: Total number of nouns and pronouns
- Contains emotional symbols: Total number of emotional symbols
- Number of question marks: Total number of question marks
- Number of spam words in the lexicon: Total number of spam words from the dictionary of the user
- Features based on the user's profile name: Total number of features based on the user's profile name
- Sentiment score of positive words: Computed sentiment score of positive words
- Sentiment score of negative words: Computed sentiment score of negative words
- Emotional symbols: Total number of emotional symbols
- Combined sentiment score: Total combined sentimental score
- Similarity score between the title and content of an email: Score based on similarity between title and actual content of the mail