**Northeastern University**

CS 5100 Foundations of Artificial Intelligence

**Homework#6 Summer 2021** [300 points]

***PART A: Theory and Algorithms*** *[100 points] \* See PART B Prog Assignment on Page 3*

Please - clearly write your **full name** on the first page. Submit a single PDF file.

Please provide brief but complete explanations, using diagrams where necessary, and suitably using your own words. While presenting calculations, explain the variables and.

Study ML resources provided in the lecture (ISLR, Witten etc.) and refer Chapter 18 of Russel AI textbook – selected sections only, plus ML and DL notes provided. Answer the below:

1. **From textbook p.763** [40 points]

18.1

18.3

18.17

18.19

1. Study the Home Credit Default Risk Kaggle competition and data sets. [10 points]

https://www.kaggle.com/c/home-credit-default-risk

Specify 4 machine learning questions relevant to this use case.

Against each ML question also list one or more algorithms which would help answer it.

1. Study the SVM Python coding/tutorials at the below links: [15 points]

https://stackabuse.com/implementing-svm-and-kernel-svm-with-pythons-scikit-learn/ https://towardsdatascience.com/support-vector-machine-python-example-d67d9b63f1c8

Then, use the below data set and apply SVM using Py libraries to classify the Mushrooms using the Iris and Mushroom data sets.

<https://archive.ics.uci.edu/ml/datasets/Iris>

<https://archive.ics.uci.edu/ml/datasets/Mushroom>

1. Using the ID3 Algorithm method, construct the full decision tree for the modified Contact Lenses dataset provided (ModLense.xls) where only Soft or None are the 2 possible class targets (classification). Show all the steps of the calculations with Information and Information Gain etc. and of the construction of your decision tree. [35 Points]

**Section 2**: Programming Assignment (PA)

1. **Decision Trees** [40 points]

Study Chapter 4 **Algorithms Sections 4.3 and 4.4 (Pages 99 to 115).**  Data Mining: Practical Machine Learning Tools and Techniques (2013) by Witten, I., Frank, E. and Hall, M. A. <http://www.cs.waikato.ac.nz/ml/weka/book.html>

Also study the ID3 algorithm method for D.T. building, provided to you in class.

Understand how you inspect a data set and build a Decision Tree. Then, answer the below questions.

1. Inspect the data set given in the AnimalData.xls. Then, draw manually (no code) a decision tree of three to five levels deep which classifies the animals into a mammal, bird, reptile, fish, amphibian, insect or invertebrate.   Justify your selection of the Root Node. Limit to a depth of 3 levels only, so your Tree calculations do not become too many.
2. For the same data set, **use Python** to construct a Decision Tree (as in #1) and a set of Classification Rules. Compare your answers with those from #1 and #2

For (C) above, use Python Orange package

https://orange.readthedocs.io/en/latest/reference/rst/Orange.classification.rules.html

* For your convenience, some Log to Base 2 logarithm values given below

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **x** | 1/2 | 1/3 | 1/4 | 3/4 | 1/5 | 2/5 | 3/5 | 1/6 | 5/6 | 1/7 | 2/7 | 3/7 | 4/7 | 1 |
| **log2(x)** | -1 | -1.5 | -2 | -0.4 | -2.3 | -1.3 | -0.7 | -2.5 | -0.2 | -2.8 | -1.8 | -1.2 | -0.8 | 0 |

Explain your results. Be sure to show NN diagram.

1. **Ridge and Lasso Regression for Regularization** [60 points]

In this assignment, you might need Jupyter notebook, as you will play around with the models. For this, you would need to install it. **Study and refer** to this implementation:

<http://www.science.smith.edu/~jcrouser/SDS293/labs/lab10-py.html>

**Install Jupyter notebook and ipynb**

with [conda](https://www.anaconda.com/download" \t "_blank):

conda install ipython jupyter

with pip:

# first, always upgrade pip!

pip install --upgrade pip

pip install --upgrade ipython jupyter

To open the ipynb files, you can open it using the command ipython notebook filename.ipynb from the directory it is downloaded on to. After running this command, you will be navigated to the Juypter notebook running on localhost:8888.

Zip file contents:

Train.csv

House-Price-Prediction-Ridge-Lasso.ipynb

The Jupyter notebook is trying to solve the below problem statement

### Problem Statement for the example provided is below:

A US-based housing company named Surprise Housing has decided to enter the Australian market. The company uses data analytics to purchase houses at a price below their actual values and flip them at a higher price. The company wants to know

* Which variables are significant in predicting the price of a house, and
* How well those variables describe the price of a house.

**Assignment Question:**

Based on the above example, you pick a data set and a problem of your choice and come up with some prediction scenario and use ridge and lasso regression to come up with the factors that influence your dataset the most.

A good place for Data sets (not needed for this PA; just a reference)

<https://www.kaggle.com/datasets>