CCE5225 Assignment II, Semester I, 2020-21.
Weighting: (30% of final unit mark)

Submission deadline (29th January 2021)

Introduction and Background:

In this assignment you will be investigating whether a Bayesian Chain Classifier (BCC) performs better in a in multi-label classification setting when compared to a Binary Relevance Model. In the experimental setup you will be using the SpatialVOC2k dataset. This dataset is typically used to develop models in predicting spatial relations among objects depicted in images. The data set has been prepared for you, including feature extraction and is readily split into the *development* and *test* portions.

In summary, you will first develop a baseline binary relevance logistic regression classifier. Next you will manually derive a Bayes network that relates the output classes, and finally you will update the LR classifier with the Bayes network. Although not necessary you can visit the SpatialVOC2k page for some background information:
https://github.com/muskata/SpatialVOC2K

In the zipped file you should find the development and test datasets (pickle files), the jupyter notebook with a dataset loader function, and a review paper on multi-label classification. In addition, you will need the Sci-Kit-Learn (LR implementation), Matplotlib and NumPy libraries.

Submission:

1. Submit your Python 3.x code and answers in an appropriately organised jupyter notebook (IPYNB), with clearly demarcated question numbers. Use markdown text, Mathjax and scanned diagrams or images as required.
2. In addition to the IPYNB file, submit a pdf version of the IPYNB with all cells executed and results displayed.

Graded Exercises and Questions:

Section I: Preparing the data.

1. Referring to the attached jupyter notebook, load and explore the dataset. More specifically:
    a. Compute the average number of output labels per example, separately for the train set and the test set.
    b. Flatten the output labels to a single vector and compute the distribution of the output labels separately for the train set and the test set.
    c. Compute the distribution of the composite output labels separately for the train set and the test set. (This will be a long tailed distribution)
    d. Compute the co-occurrence probability distributions, i.e. the probability of an output label given the occurrence of another output label. The answer should be a matrix of size 17x17.

[1, 2, 2, 5 marks]

2. Prepare the input (**X,** size = #examples x 54) and output (**Y,** size = #examples x 17) matrices (the geometrical features matrix is given in the dataset as a list of numpy 1D arrays. More specifically:
    a. Object labels encoded into one-hot binary vectors (each vector is of length 20 and two of these are required to represent the trajector and the landmark) concatenated with respective geometrical features.
    b. Multi-label output vectors. You can use the one-hot encoder in (sklearn. preprocessing) to generate a sum of the one-hot-vectors, (one for each preposition in the multi-label set). (see example in the attached jupyter notebook)

[5, 5 marks]

3. Develop functions to compute the multi-label accuracy metrics. Test the functions with appropriate examples. (See the companion paper for the formulae of the various metrics). More specifically, develop the following functions:
    a. Intersection over union
    b. Precision
    c. Recall
    d. Precision per output label
    e. Recall per output label

[2, 2, 2, 2, 2 marks]

Section II: Building and Evaluating the models.

4. Develop the binary relevance model using Logistic Regression models, i.e. an independent logistic regression model for each output label that outputs the probability that a preposition is acceptable for a given spatial configuration:
    a. Split the development set to carry out hyper-parameter optimisation and retrain the final models on the full development set.
    b. Generate the results on the test set for all metrics listed in Q3 above.

[20, 10 marks]

5. Using the co-occurrence matrix as well as intuition, develop the structure of a Bayesian Network (BN) for the BCC.

[20 marks]

6. Retrain the logistic regression models to include the BN (from Q5) and generate the results on the test set for all metrics listed in Q3 above.

[10 marks]

7. Compare the results from Q4 and Q6 and conclude on whether the BCC performs better than the binary relevance model.

[10 marks]

End of document