You should submit **precisely 3 files**:

1. **Two page report** (only references may appear on a third page) in pdf format using the word template provided.

2. **Code** (a version that will run; **do not submit any data**).

3. **Class predictions for the test data.** This must be in the same format as the file sample_valid_predictions.csv and have the name **predictions.csv**.

The Brief

For this assignment, you will carry out a binary classification task, and write a report on this. Please read this brief carefully and also the Marking Criteria and Requirements below.

The data comes from photos, and your task is to come up with a machine learning method for classifying the photos according to whether or not they are 'memorable'. The data you are given for each photo consists of 4608 features. 4096 of these were extracted from a deep Convolutional Neural Network (CNN) [1], and the remaining 512 are gist features [2]. (You are given all these features as a 1-dimensional array, so you will not be performing any feature extraction on raw images.)

There are two files of training data. The first contains 600 samples with all the data present. The second contains 2800 samples, which have some missing data, as indicated by a NaN (not a number). The training data have class labels, 1 for memorable, and 0 for not memorable. In addition, there is also a confidence label for each sample. The class labels were assigned based on decisions from 3 people viewing the photos. When they all agreed, the class label could be considered certain, and a confidence of 1 was written down. If they didn't all agree, then the classification decided on by the majority was assigned, but with a confidence of only 0.66.

There is one file of test data, containing 2000 samples. You must obtain predictions for the class labels of these. (Note that, as with the second training set, the samples in the test data set contain missing features.)

Your job is to obtain the best predictions you can, and to justify your methods. You should reason for which classifier or combination of classifiers you use, how you do model selection (training-validation split or cross validation), and how you handle the specific issues with these data (large number of features, missing data, the presence of confidence labels for the classes of the training data). We value creative approaches!

You may make use of any classifier, such as: single-layer perceptron, multi-layer perceptron, SVM, random forest, logistic regression. You are not required to code classifiers from scratch, and you can use any machine learning toolbox you like, such as scikit-learn (https://scikit-learn.org/stable/ (Links to an external site.)).

Marking criteria and requirements

Your report must contain the following sections.

1. Approach (10 marks)

Present a high-level description and explanation of the machine learning approach (e.g. multi-layer perceptron, logistic regression, or a combination thereof) you have used. You should cover how the method works and key assumptions on which the approach depends. Pay close attention to characteristics of the data set, for example: high dimensionality.

2. Methods (30 marks)

Describe in detail what you did and include references to appropriate literature.

- How did you train and test your classifier(s)?

- How did you do model selection?

- Did you rescale the data? (See https://scikit-learn.org/stable/modules/preprocessing.html (Links to an external site.))

- Did you do feature selection? You are provided with two types of features: CNN features and gist features. Are they equally important? (See https://en.wikipedia.org/wiki/Feature_selection (Links to an external site.))

- How did you deal with missing data? And did you do something with the confidence labels?

3. Results and discussion (30 marks)

Use graphs and/or tables to illustrate the results of your model selection (see week 4 lectures). For example,

- Show how the choice of classifier hyper-parameters affect performance of the classifier, using a validation set.

- Show changing performance for different training sets. How useful, relatively, are the incomplete training data? How useful is it taking account of the training label confidence?

In the discussion:

- If you think that there might be ways of getting better performance, then explain how.

- If you feel that you could have done a better job of evaluation, then explain how.

- What lessons have been learned?

4. Coding (20 marks) and accuracy of your predictions (10 marks)

Please make sure we will be able to run your code as is. High quality codes with a good structure and comments will be marked favorably.

We will compute the accuracy of your class predictions for the test data (as percentage correct), and give you a mark out of 10 for this. Those of you with the most accurate predictions will score 10/10. Those of you with the least accurate predictions will score 5/10. If you don't submit a class predictions file in the correct format, then you score 0/10.

Footnotes:

[1] Extracted from the fc7 activation layer of CaffeNet http://caffe.berkeleyvision.org (Links to an external site.)

[2] http://people.csail.mit.edu/torralba/courses/6.870/papers/IJCV01-Oliva-Torralba.pdf