Coursework Assignment

Deadline: 20th May 2019 at 4PM

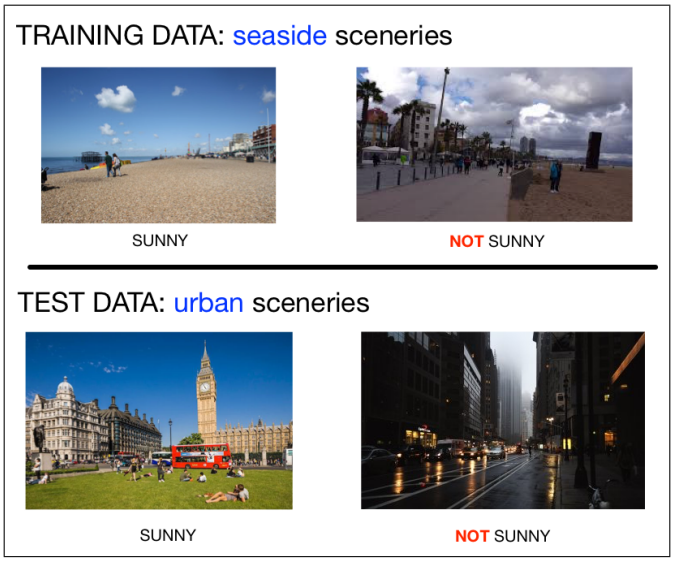
This assignment brief was first released on 24th March 2019

Summary: You have to participate in the **Kaggle competition** and have to submit a **2-page report** (using the provided **template**) and an **implementation code**.

As part of the practical assessment you are required to participate in the Kaggle in Class Competition "**Bring back the sun!**" `https://www.kaggle.com/t/7532bc3e0ea8465fa588379534fccd24` (**please use this URL link**). The assessment grade, which is worth 100% of the total grade, is separated into 2 components: the 2-page report and the source code. The code component will be weighted based on your performance in the Kaggle competition. No submission to the competition means 0.0 weight. **You have to make at least one submission to the Kaggle competition to get a non-zero weight!**

The competition is a binary classification problem (label 1 for sunny and label 0 for not sunny). You are provided with 259 labelled training data (146 of sunny scenes and 113 of not sunny scenes), and 2818 of test data, which are not labelled. The task is to develop a binary-class classifier that predicts the labels for the test data set. Each data instance is represented as a 4608 dimensional feature vector. This vector is a concatenation of 4096 dimensional deep Convolutional Neural Networks (CNNs) features extracted from the fc7 activation layer of CaffeNet [1] and 512 dimensional GIST [2] features (**this representation is given therefore you do not need to perform any feature extraction on images**).

Might be interesting (read: **important**) to note that our training data were collected in a seaside area, while test data were collected in an urban area.



[1] `http://caffe.berkeleyvision.org`

[2] `http://cvcl.mit.edu/Papers/IJCV01-Oliva-Torralba.pdf`

Additionally, you are also provided with three types of information that might be useful when building your classifier: a) additional 2331 labelled training data which is incomplete as it has missing feature values, b) confidence of the label annotation for each training data point (259 labelled training data and additional but incomplete 2331 labelled training data), and c) the proportion of positive (sunny) data points and the proportion of not sunny data points in the test set. You can choose ("*life is full of choices and consequences*") to incorporate or to ignore these additional data.

You can use any of your favourite classifiers. Some of the classifiers that we have discussed/will discuss in the class are: linear perceptron, multi-layer perceptron, random forest, support vector machine, and logistic regression. You are not required to code the classifier from scratch. Feel free to use some of machine learning toolboxes such as Weka [3] (in Java), scikit-learn [4] (in Python; **recommended**), shogun [5] (in C++), or stats [6] (in Matlab). We value your creativity in solving the binary classification problem with the **four twists**. You have to reason which classifier or combination of classifiers you use, how you handle issues specific to competition data set such as high dimensionality of the data (large number of features), how you do model selection (training-validation split or cross validation), and how you do further investigations to take into account the **four extra information**: 1) additional but incomplete labelled training data, 2) test label proportion, 3) the annotation confidence on labels, and 4) the domain adaptation problem (differing domains between training and test; seaside versus urban).

Details of Research Report

You are expected to write a 2-page report detailing your solution to the Kaggle competition problem. Please use the provided latex or word template. Your report should include the following components (you are allowed to combine descriptions #2 and #3 but make sure we can easily identify them).

1. APPROACH (Maximum mark: 10) You should present a high-level description and explanation of the the machine learning approach (e.g. support vector machine, logistic regression, or a combination thereof) you have adopted. Try to cover how the method works and notable assumptions on which the approach depends. Pay a close attention to characteristics of the data set, for example: high dimensionality.

2. METHODOLOGY (Maximum mark: 30) Describe how you did training and testing of the classifier of your choice. This should include model selection (Did you do model selection? what was it meant for? what were you selecting?) and feature pre-processing or feature selection if you chose to do it. Feature pre-processing could be in the form of [7]:

- Standardisation: to remove the mean and scale the variance for each feature, that is to make each feature having 0 mean and 1 standard deviation.
- Normalisation: to scale individual observation or data point to have a unit norm, be it L1 or L2 norm.

[3] http://www.cs.waikato.ac.nz/ml/weka/
[4] https://scikit-learn.org/stable/
[5] http://www.shogun-toolbox.org
[6] https://uk.mathworks.com/help/stats/index.html
[7] https://scikit-learn.org/stable/modules/preprocessing.html#standardization-or-mean-removal-and-variance-scaling

- Binarisation: to threshold numerical feature values to get boolean values.
- Scaling: to scale features to lie between minimum and maximum values, for example to lie in [0,1] or [-1,1].

Feature selection [8] methods are for example: filter methods such as univariate feature selection based on chi-squared statistics, wrapper methods such as recursive feature elimination, and L1 norm penalisation for sparse solutions. You are provided with two types of features: CNNs features and GIST features. Are they equally important?

Describe any of your **creative solutions** with respect to additional characteristics of the competition data set, such as how to incorporate the extra information about: **additional training data with many missing features, test label proportions, training label confidence, and domain adaptation problem** [9]. For the latter, this Python library might be useful: `https://pypi.org/project/libtlda/`.

Reference to appropriate literature should be included.

3. RESULTS AND DISCUSSION (Maximum mark: 30) The main thing is to present the results sensibly and clearly. Present the results of your model selection. There are different ways this can be done:

- Use table or plot to show how the choice of classifier hyper-parameters affect performance of the classifier using validation set (**refer to lectures on model selection**). Classifier hyper-parameters are for example, minimum number of instances required to split an internal node, maximum number of features to consider when looking for the best split, and choice of impurity measure in decision tree; hyper-parameters for each decision tree and number of trees in random forest; regularisation values in support vector machine and in logistic regression.
- Use graphs to show changing performance for different training sets (**learning curve; refer to lectures on model selection**), if you choose to do that.

If any, provide analysis on the usefulness of taking into account the provided additional incomplete training data, test label proportions, training label confidence, and addressing explicitly the domain adaptation problem.

You should also take the opportunity to discuss any ways you can think of to improve the work you have done. If you think that there are 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, if any have been learnt? Were your goals achieved? Is there anything you now think you should have done differently?

Details of Code

You must also submit your implementation codes. 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. As mentioned earlier, the code component will be weighted based on your performance in the Kaggle competition. No submission to the competition means 0.0 weight. **Maximum mark: 30**

[8] `https://en.wikipedia.org/wiki/Feature_selection`
[9] `https://en.wikipedia.org/wiki/Domain_adaptation`

Marking Criteria

70% – 100% **Excellent**

Shows very good understanding supported by evidence that the student has extrapolated from what was taught, **through extra study or creative thought** (e.g. incorporating additional training data with many missing features, test label proportions, training label confidence, and addressing domain adaptation problems). Work at the top end of this range is of exceptional quality. Report will be excellently structured, with proper references and proper discussion of existing relevant work. The report will be neatly presented, interesting and clear with a disinterested critique of what is good and bad about approach taken and thoughts about where to go next with such work.

60% – 69% **Good**

The work will be very competent in all respects. Work will evidence substantially correct and complete knowledge, though will not go beyond what was taught. Report should be well-structured and presented with proper referencing and some discussion/critical evaluation. Presentation will generally be of a high standard, with clear written style and some discussion of related work.

50% – 59% **Satisfactory**

Will be competent in most respects. There may be minor gaps in knowledge, but the work will show a reasonable understanding of fundamental concepts. Report will be generally well-structured and presented with references, though may lack depth, appropriate critical discussion or discussion of further developments, etc.

40% – 49% **Borderline**

The work will have some significant gaps in knowledge but will show some understanding of fundamental concepts. Report should cover the fundamentals but may not cover some aspects of the work in sufficient detail. The work may not be organized in the most logical way and presentation may not be always be appropriate. There will be little or no critical evaluation or discussion. References may be missing, etc.

30% – 39% **Fail**

The work will show inadequate knowledge of the subject. The work is seriously flawed, displaying major lack of understanding, irrelevance or incoherence. Report badly organized and incomplete, possibly containing irrelevant material. May have missing sections, no discussion, etc.

Below 30% **unacceptable (or not submitted)**

Work is either not submitted or, if submitted, so seriously flawed that it does not constitute a bona-fide report/script.