ML : Assignment 2016-2017 – Image Advert Classification

Image Classification : Advert Images in Web Pages

Background: The automatic detection of advert images in web pages is an interesting problem that allows such images to be automatically blocked from given users due to their age, location (e.g. school campus), connection (e.g. via iphone) or personal choice (e.g. advert blocking "plug-ins for modern web browsers to save bandwidth). In addition, it allows users (and service administrators) to be selective in the allocation of resources such as storage and bandwidth utilization in relation to such images.

It also allows some (less scrupulous) internet access providers to intercept and substitute web page adverts on the pages users browse over their connection with adverts for their own products and services. This is a powerful marketing tool for such network operators that they can then use to assist in covering the costs of providing the connection (e.g. free hotel wi-fi providers).

In this practical assignment we will investigate the use of machine learning techniques for the automatic classification of a given web page image as being an advert image or not.

Task Specification – Image Classification based on URL, text and size attributes

In this task you will investigate the use of different machine learning techniques to solve this classification problem. From the range of approaches that we have studied in lectures **select three machine learning classification approaches** (i.e. machine learning algorithms) that would be suitable for this problem. Your task is to experiment with these approaches using the provided data set and report the results. *N.B. The majority of assignment marks relate to good experimentation and statistical presentation of the results (see Lecture 1), not just writing program code.*

A set of data is provided on the Machine Learning Blackboard page.

“This dataset represents a set of possible advertisements on Internet pages. The features encode the geometry of the image as well as phrases occuring in the URL, the image's URL and alt text, the anchor text, and words occuring near the anchor text. The task is to predict whether an image is an advertisement ("ad") or not ("nonad").” - ad_cranfield.name description file

You are provided with the source code for reading in this data for use with the machine learning algorithms available in OpenCV. Ths file is called reader.cpp and can be found on the Machine Learning Blackboard page.

You should use this as the basis for implementing your chosen machine learning approaches using the OpenCV Machine Learning Library. The provided data is not split into training and testing sets and is not randomly shuffled in any way. You may find the *randomize.cc* and *selectlines.cc* examples available from in the Tools folder on Blackboard.

Additional Program Specifications

Additionally, to facilitate easy testing, your program(s) **must** meet the following **functional requirements:**

● Your program(s) must accept a training and testing data set files at the command line in the form:

```
yourprogramname filename.train filename.test
```

● Your program(s) must compile and work with OpenCV 2.4.11 on the lab PCs.

Marks

The marks for this practical will be awarded as follows:

- **Working programs** for the classification of the provided data using 3 different machine learning approaches from the course *20%*
- *Clear, well documented program source code 5%*
- **Experimental Procedure Followed** 45% (data preparation, performance measurement, parameter search and selection, comparison of techniques, good experimental practice ...)
- **Report (detailing experiments and results)**
    - ○ Discussion/details of approaches chosen and experimental procedure 10%
    - ○ Evidence of the performance of your chosen approaches on the data 15%
    - ○ Conclusions from the experimentation 5% **Total 100%** *N.B. The marks in this assignment are generally awarded for good experimentation and good experimental procedure that support the results, not for achieving the best classification rates.*

Submission You must submit the following:

- ● Full program **source code** for your program(s) used in the above task and copies of any training or test data sets you constructed from the original data file.
- ● **Working executable(s)** meeting the above ***"additional program specifications"*** for testing.
- ● **(Short) Report (max. 750 words!)** detailing your chosen machine

learning approaches and the results/conclusions of your experimentation with these approaches on the provided data set. Provide any tables, graphs and charts (as many as you feel necessary) to summarise and support the performance of your chosen approaches on this data. Submit this as a PDF or (*if you really, really must!)* a Microsoft Word document. ***Make it clear in the initial comments of your source code how to run your executable. Your executable must run on one of the teaching lab based PCs – ensure compatibility before submission.***

Plagiarism : You must not plagiarise your work.

You may use program source code from the provided course examples, the OpenCV library itself or any other source BUT this usage must be acknowledged in the comments of your submitted file. Automated software tools will be used to detect cases of source code plagiarism in this practical exercise (taking into account the provided common source code examples from the course and the assignment itself).

You should have been made aware of the Cranfield University policy on plagiarism. Anyone unclear on this must consult the course lecturer prior to submission of this practical. University Plagiarism Guidance: http://www.cranfield.ac.uk/library/cranfield/support/page41148.html

To submit your work create a directory named by your capitalised Firstname and Surname (e.g. *KermitFrog*). Place all required files in this directory. Zip (not rar/7z/tgz etc.) this entire directory structure and submit it via the Blackboard portal.

Submission Deadline (ALL STUDENTS) :

9:30am (UK time) – 21st April 2017 *(late submissions will be penalised)*