Mobile Sensing and Data Analysis Assignment

Mobile and Ubiquitous Computing Module

Description:

"Big data analysis", "data mining" and "machine learning" are among the most sought for skills, both in academia, as well as in industry. Nowadays, huge amounts of data are available from the Web, online social networks, but also personalised sensing devices, such as smartphones. The data is meaningless until it is mined for useful information. Machine learning helps us build models of data behaviour so that interesting relationships can be identified.

In this assignment you will run a mobile sensing application SampleMe that will help you obtain your personal accelerometer data, together with information about your interruptibility. You will then convert that data to a form that is suitable for machine learning. Finally, you will build and test a model of your interruptibility.

Outcomes:

After you complete this assignment you will be able to:

- List and briefly describe the stages of data-driven research
- Transform raw sensor data to representative descriptive features
- Apply basic machine learning methods in WEKA

Instructions and Marking:

This assignment counts towards 5% of your final mark for the module. The credit within the assignment is further distributed as follows:

- **10% - Run SampleMe Android application from now on till Saturday 1.3.2014 on a personal Android phone.**
  - We will periodically check the number of data points in the database. If we see that your phone does not generate data points, we will contact you on the email you provided when you registered for the SampleMe application. In such case, please try to coordinate with your lecturer, and arrange a time to come to the Computer Science building, meet us, and have the application fixed.
  - Note that it is impossible to get the rest of the assignment credit if you fail to complete the data collection part.
  - You should find at least one more person with an Android phone (friend, family, etc.) who will run SampleMe in parallel to you. This is to ensure you get some data even if something goes wrong, for example, your phone is lost, accelerometer misbehaves, etc.
  - IMPORTANT: if you do not have an Android phone, please email sampleme.bham@gmail.com as soon as possible, and we will try to get you one.

- **40% - Extract features from your raw accelerometer data sensed by SampleMe**

- Access you accelerometer data from:
  - www.cs.bham.ac.uk/~pejovicv/SampleMe/data_access.php
    The link will be available ten days after you start your data collection.
- Write a program that:
  - Reads a text file with a number of accelerometer samples and the corresponding interruptibility label; the file will be provided by the above URL. Each sample consists of a list of comma separated consecutive accelerometer readings (approximately 200 of them).
  - Calculates the following for each accelerometer sample (i.e. for each 200 reading batch):
    - Mean intensity m = $\sum\left(\sqrt{\left(x_i^2+y_i^2+z_i^2\right)}\right)$
    - Intensity variance v = $\frac{1}{n}*\sum\left(\sqrt{\left(x_i^2+y_i^2+z_i^2\right)}-m\right)^2$
    - Intensity mean crossing rate MCR =
      $\frac{1}{N-1}*\sum_{i=2}^{N} I\left(\left(\sqrt{\left(x_i^2+y_i^2+z_i^2\right)}-m\right)\left(\sqrt{\left(x_{i-1}^2+y_{i-1}^2+z_{i-1}^2\right)}-m\right)<0\right)$
      where *I(x)* is the indicator function, and is equal to 1 if its argument (*x*) is true, and is equal to 0 if its argument is not true.
  - Prints out a WEKA-ready representation of the given data. WEKA requires the following fields "`@relation`" with a name of the relation you are trying to describe. You can pick any name. A list of attributes given with the keyword "`@attribute`" followed by the attribute name and type. Real number values, such as accelerometer mean, variance and MCR are "`numeric`" values in WEKA. Interruptibility can take two values "`yes`" and "`no`", therefore it is not numeric, but nominal and the two possible values for interruptibility should be given when you define the attribute. Finally, the actual values you calculated should be given after the "`@data`" label, one line per one sample. In summary, your output should look like this:

    ```
    @relation ANY_NAME
    @attribute meanAcc numeric
    @attribute varAcc numeric
    @attribute MCRAcc numeric
    @attribute interruptibility {'yes','no'}

    @data
    m1, v1, MCR1, inter1
    m2, v2, MCR2, inter2
    m3, v3, MCR3, inter3
    …
    ```
    … and so on depending on how many data points you have. Where values m, v and MCR are calculated by your programme, and the value `inter` is given for each data point.
  - Saves the results in a file called `interruptibility.arff`
  - You can write your code in either Python or Java. Your code has to run. No partial marks will be given based on the code that does not run.

- **40% - Build, train and test an interruptibility classifier in WEKA**
  - Download and install WEKA on your laptop http://www.cs.waikato.ac.nz/ml/weka/
    - If you don't have a laptop, let us know, it might be possible to get WEKA running on some of the lab machines.

- Get familiar with WEKA:
  - Short online tutorial: http://www.youtube.com/watch?v=m7kpIBGEdkI
  - Manuals and tutorials on the WEKA website
- Load your accelerometer – interruptibility data file (`interruptibility.arff`) in WEKA:
  - Open WEKA Explorer and open the file you created in the previous task. You should see the list of attributes on the left hand side of the window.
- Build a classifier for your interruptibility
  - Select the "classify" tab.
  - Select one of the classifiers. "NaiveBayes" under "bayes" is a good starting point.
  - Select the way you want to train and test your classifier. Available options include:
    - **training set** – where you test and train your classifier on the same data. Can you tell why this is not good?
    - Train on the loaded data, but test on a separate **supplied test set**.
    - **Cross-validate**, so that a small randomised subset of your data set is used for testing, while the majority is used for training. Then, the whole process is repeated N times and the results are averaged out.
    - **Percentage split** your data set so that a certain percentage of it is used for training, and the rest for testing.
  - Start the classifier
- Examine classification results
  - In the right hand side of the WEKA window you will see a summary of the classification results. Find the percentage of correctly classified instances. Find the confusion matrix that tells you how often your classifier predicted your interruptibility as "Yes" when in fact it was "No" and vice versa.
  - Copy the results that you find interesting to your report.
- Save the classifier to a file
  - Right click on the classifier name under "Results list" and select "Save model".
  - Check the resulting file. If you restart WEKA and reload your data set file, you should be able to reload the model. Just go to "Classify" tab, right click in the blank are of the "Results list" and load your model. Please verify that this works before submitting your assignment.
- Feel free to experiment with different classifiers.

- **10% - Write a short report about your experiences**
  - The idea of the report is to help you summarise and understand the process that you just completed. You performed mobile sensing-based data collection, feature extraction, machine learning-based classification. You should be aware of all the steps so that you can efficiently market your skills when you apply for PhD studies, a job in industry, etc.
  - The report should be 1-2 pages in length, single space, point 11 font. Try to include only the most relevant results to confine to the page limit, i.e. don't just copy-paste everything that WEKA has to output.
  - The report has to contain the following:
    - A brief explanation of all the steps you performed in the assignment.
    - Your thoughts on the feature extraction from raw data: why do we perform it? Why are mean, variance and MCR relevant features? What are some other alternatives for features? If you were to conduct a similar interruptibility study what data would you include?
    - Reflection on classification with WEKA. There is no strict format of this section, but some ideas of what you could write about are: have you explored functions of

WEKA that were not mentioned in the assignment? Are there any interesting features of the software that you would like to know more about? Which classifiers have you tried on your data? If you tried more than one classifier, briefly compare the results. Which training/test set method did you use and why?

- Please save the report in the pdf format.

Assignment Submission and Deadlines:

Your full data will be available on **1.3.2014** at
www.cs.bham.ac.uk/~pejovicv/SampleMe/data_access.php
You will need your username and password from the time of registration in order to access the data.

You need to submit:

- Your code for feature extraction. The code should take as an input a text file you get from the above link, and output `interruptibility.arff`
- Your WEKA classifier model file.
- Your progress report pdf file.

The submission will be through Canvas.

Deadline for the submission is **13.3.2014.**

Miscellaneous:

The University policies about academic integrity hold for this assignment as well. Please be aware of them, as any plagiarism and cheating will be heavily sanctioned:
https://intranet.birmingham.ac.uk/as/studentservices/conduct/plagiarism/guidance-students.aspx