Databases and Data Mining

Assignment 4

23-11 2009

Due: Friday 21-12 2009
Grading: This assignment will be graded from 0 to 10.
Notes: Please read carefully:

- Groups of 1-3 students are allowed.
- Use C, C++ (MS Visual C++, or gcc), Python, or JAVA together with an HMM Toolkit of your choice (GHMM [4] is recommended). Also MatLab code is allowed, if your MatLab code installs and runs under the available LIACS license on the Linux student machines.
- Put the complete code with **short** and **clear** instructions on how to compile and execute it in a single directory called ***"<your student number(s)><your last name(s)>_4"***. Please add all the data files you used with clear instructions.
- Write down your report for this assignment in a *.pdf* file with the following name ***"<your student number(s)><your last name(s)>_4.pdf"***, e.g., ***"012345jansen_4.pdf"*** and put it in the same directory as the code**.** (A *.doc* file is also allowed.)
- Compress the complete directory into a single zip file called ***"<your student number(s)><your last name(s)>_4.zip".***
- Send this *.zip* file as an attachment of an e-mail with subject **"DBDM_4"** to erwin@liacs.nl.
- Grading will be based on the originality, and quality, of your code, the quality of your analysis and results, and the argumentation, validity, and clarity of your **final report**. Do not forget to clearly state the references you used for your work!

Introduction

Hidden Markov Models (HMMs) have been used in many application areas to model very different types of data [2, 3]. Especially, in the field of continuous speech recognition remarkable successes have been obtained by applying HMMs. A very nice tutorial on HMMs applied to the field of speech recognition is by L.R. Rabiner [1].
HMMs have also been applied to stock market analysis [5,6]. Understandably, people are very interested in ways to automatically predict the world market composite and other indexes based on the index price data seen thus far. **The goal of this assignment** is to automatically predict (as accurate as possible) the **FTSE 100 Index** prices based on historical composite index price using HMMs.

Datasets

Obtain your datasets from ***Google Finance*** (http://www.google.com/finance). See also http://computerprogramming.suite101.com/article.cfm/an_introduction_to_google_finance for an introduction on how to obtain historical data using ***Google Finance***.

Carefully describe and explain your experimental set up and the data sets you have chosen to use.

Statistical Report

Part of your report should be a detailed statistical report of the data sets you used. Explain your analysis and findings. Describe how you used these results in your solution.

HMM Modeling

It is encouraged that for designing and modeling your HMM for predicting the **FTSE 100 Index** prices, you use the General Hidden Markov Model library (GHMM) [4]. GHMM is a freely available LGPL-ed C library implementing efficient data structures and algorithms for basic and extended HMMs with a Python interface. Of course you are also allowed to use another HMM Toolkit of your choice (e.g. HTK, HMM under Matlab, etc.).
Please give very clear instructions on how you modeled, trained and tested your HMM! In principle your fellow students should be able to check your work and results very easily using these instructions.

HMM Training & Testing

Describe and define the data sets you used for training and testing your HMM, respectively. Report your results in a table. Make a zip-file with the data sets and the code for executing your experiments and obtaining your results.

HMM Validation

The performance of your HMM will be validated using **FTSE 100 index** data that will become available after 23-11 2009. Give instructions on downloading the necessary data and executing a (single) script for this validation.

References

1. L.R. Rabiner, A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, Proceedings of the IEEE, Vol. 77, No. 2, pp 257-286, February 1989.
2. A. Krogh, I. Saira Mian, D. Haussler, A Hidden Markov Model that finds genes in E. coli DNA, Nucleic Acids Research, Vol. 22, ,pp. 4768-4778, 1994.
3. http://www.genome.wisc.edu/sequencing/o157.htm
4. http://ghmm.sourceforge.net/ and http://ghmm.org/
5. R. Hassan, A combination of hidden Markov model and fuzzy model for stock market forecasting, Neurocomputing archive, Vol. 72 , Issue 16-18, pp 3439-3446, October 2009.
6. M.R. Hassan, B. Nath, Stock market forecasting using hidden Markov model: a new approach, in: Proceedings of the Fifth International Conference on Intelligent Systems Design and Applications, 2005, pp. 192-196.