Data Analysis and Manifold Learning

Radu Horaud
PERCEPTION team
INRIA Grenoble Rhône-Alpes
655, avenue de l'Europe
38330 Montbonnot, France
Radu.Horaud@inrialpes.fr

March 25, 2011

http://perception.inrialpes.fr/people/Horaud/Courses/DAML_2011.html

Total number of course hours: 24 (2h$\times$12)

Agenda: The course will start in January and will finish in April. There will be approximatively 2-3 lectures per month.

Who should attend: PhD candidates and Master students who are interested in applying machine learning methodologies to their research. The course is open to any researcher interested in manifold learning tools for data analysis.

Course Objective

The objective of this course is to familiarize PhD candidates and Master students in computer science, electrical engineering, and applied mathematics with data analysis methodologies that fall within the topic of *statistical machine learning*, (Hastie *et al.*, 2009) and which have their roots in algebraic graph theory and heat-diffusion in Riemannian geometry. Within the light of these powerful mathematical tools, the data will be analysed from the perspective of their *intrinsic* geometry.

There are many fields that can profit from these methods, such as:

- Visual computing (images, image collections, and videos)
- Graphical and animation methods (voxels, meshes, 3D scans),
- Robotics (2D and 3D point clouds, multi-sensory data, inverse kinematics)
- Audio signal processing (auditory scene analysis, speech recognition, audio-visual interpretation),
- Medical analysis (MRI and fMRI data), and
- Data mining (text, multimedia documents, social networks).

Very often, the task is to cluster the data in order to discover meaningful groupings, segmentations, or associations (unsupervised learning) or to classify the data based on prior training (supervised and semi-supervised learning).

Prerequisites. The course will only require basic knowledge in linear algebra, matrix analysis, probability theory, and statistics.

Brief Course Description

Modern data analysis and knowledge acquisition systems make use of various machine learning methods in order to understand the intrinsic structure of the data to be analyzed and to classify the data, cluster the data, or to infer abstract representations of the data (Hastie *et al.*, 2009). In the recent past, statistical machine learning has played a central role with emphasis on supervised or unsupervised methods. Very often, the data to be analyzed live in a high-dimensional space and one great challenge is to be able to map the data into a lower dimensional space such that standard statistical methods could be efficiently applied. Moreover, in may cases the data lie on a non-linear subspace (manifold) but neither the actual structure nor the dimension of the latter is known in advance.

In this course we will study *spectral dimensionality reduction* methods, i.e., methods that operate on either the covariance or the Gram matrices built over the input data (Bishop, 2006). We will briefly

review principal component analysis (PCA) and multidimensional scaling (MDS) and we will then turn our attention towards *graph-based methods*. We will formally introduce undirected weighted graphs as a convenient representation of the data and we will concentrate on the study of these graphs based on the algebraic (spectral) properties of their associated graph matrices, namely *spectral graph theory*. We will study in detail the associated non-linear dimension reduction algorithms such as Isomap, locally-linear embedding (LLE) and in particular Laplacian eigenmaps.

We will introduce a more general type of data embedding based on the *discrete heat-kernel*, which is the fundamental solution of the heat-diffusion equation on graphs. Within this framework, graphs are viewed as discretizations of Riemannian manifolds (Biyikoglu *et al.*, 2007). We will study in detail the properties of the *heat kernels* and of *heat matrices*. We will adopt a point of view that unifies spectral graph theory (Chung, 1997; Godsil and Royle, 2001), non-linear dimensionality reduction, and kernel methods for machine learning (Shawe-Taylor and Cristianini, 2004).

We will study in detail *spectral clustering* and *spectral graph matching* which will be illustrated using examples from computer vision: shape segmentation, shape registration, and video classification.

Course Organization and Material

The course will be split in 12 classes of 2 hours. A course website will be open with course material available: relevant papers and book chapters, course slides, matlab code and test data for the basic algorithms that will be discussed.

There will be no more than 2-3 classes per month in order to allow every participant to read the proposed material. Participants will be encouraged to propose case-studies associated with their own research work and to discuss possible solutions in terms of manifold learning.

References

Bishop, C. (2006). *Pattern Recognition and Machine Learning*. Springer.

Biyikoglu, T., Leydold, J., and Stadler, P. F. (2007). *Laplacian Eigenvectors of Graphs*. Springer.

Chung, F. (1997). *Spectral Graph Theory*. American Mathematical Society.

Godsil, C. and Royle, G. (2001). *Algebraic Graph Theory*. Springer.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.

Shawe-Taylor, J. and Cristianini, N. (2004). *Kernel Methods in Pattern Analysis*. Cambridge University Press.