University of Pittsburgh
CS 3750 Advanced Machine Learning
Professor Milos Hauskrecht

Handout 1
October 9, 2003

Midterm research project

Due: Monday, November 3, 2003

Midterm project

The objective of this project is to investigate extensions and refinements of the PCA model. The project is a group project and you will work in groups of 2-3 students. The readings that should get you started on the topic are available at:
$http://www.cs.pitt.edu/\sim milos/courses/cs3750/mid-project.html$

The product of this research should be (1) a written report and (2) a set of programs used in experiments. You will need to submit both to accomplish the project. The report should be self-explanatory and describe in sufficient detail the theory and relations behind the models. The best way to approach the write-up is to think that it is a scientific paper you want to present to the Machine Learning community. The programs can be written in C, C++, Matlab or Java, it is your choice. However, I expect you to describe in detail the platform and the implementation so I can run your programs and reproduce your experiments. Please do not expect me to provide any solutions or consulting on the programming issues.

Part A. Principal component analysis (PCA)

Principal Component Analysis (PCA) is a widely used method for reducing the number of dimensions of a data set. The PCA finds projections of a high dimensional data into a lower dimensional subspace such that the variance retained in the projected data is maximized. Equivalently the PCA gives uncorrelated projected distributions and minimizes the least square reconstruction error.

PCA has a variety of uses. Most often it is used for feature reduction, but it often used also in data visualization. As discussed earlier in the course the PCA is often used in the link analysis and where it can help to identify distinct node subcommunities in the link structure. The principal eigenvector of the connectivity matrix is used to define the hub and authority scores in Kleinberg's HITS. A very good summary of the research on link analysis for WWW can be found in the paper by Borodin et al (2001).

Tasks

1. Give a summary of PCA methods and their applications. Please use mathematical formulae, not just text to describe the method.

2. Use the PCA to analyze at least two datasets (of your choice) that represent:

   - A link structure (web, citations, document relations). Some datasets representing the link structure can be found at:
     $http://www-2.cs.cmu.edu/\sim jkubica/code/linkds.html$.
     Datasets used in Borodin's paper are available at:
     $http://www.cs.toronto.edu/\sim tsap/experiments/www10-experiments/$.
     Other possible sources of link structures is the Cora citation network used in experiments in some of the papers on the reading list.
   - Microarray gene expression data. Microarray datasets can be found, e.g. at:
     $http://smd.stanford.edu/$ under the published data option. A nice overview of the role of PCA and SVD in the analysis of gene expression data can be found at:
     $http://smd.stanford.edu/help/svd.shtml\#preparation$. You do not have to understand the detail of the biology, the point here is to apply the PCA to a dataset with high dimensional data points.

3. Analyze the results obtained via PCA.

Part B. Probabilistic PCA

The quadratic reconstruction cost optimized in the PCA is closely related to the ML estimate of the parameters of the covariance matrix of the multivariate Gaussian. This leads to a formulation of the PCA in terms of a statistical latent factor model. We refer to such a model as to **probabilistic PCA (PPCA)**. Probabilistic PCA (PPCA) is an extension of the traditional PCA, and was proposed independently by Roweis (1998) and Tipping and Bishop (1999). In traditional PCA, directions "outside" the subspace are simply discarded. In PPCA these directions are assumed to contain the same Gaussian noise. The advantage of PPCA over traditional PCA is that it defines a proper probabilistic model that can work on missing data and its parameters can be found using the EM algorithm.

Note: The PPCA model can easily be extended to mixture models (Tipping and Bishop), trained using the EM algorithm. Bishop has also proposed Bayesian methods based on ARD (automatic relevance determination) to automatically determine the number of dimensions to retain.

Tasks

1. Give a concise summary of the PCA and PPCA models and their relations. Summarize the main advantages and disadvanatges of the two models. Please use mathematical formulae as often as necessary, not just the text.
2. Implement a probabilistic PCA model with k latent factors. Your model should be such that k is a parameter and can change. Implement the EM procedure for finding the parameters of the model. An efficient PPCA procedure has been proposed by Roweis so you definitely want to take a closer look at his paper.
3. Apply the PPCA model to find the principal components on the datasets selected in part A. Compare the results obtained through a regular PCA and PPCA in terms of efficiency and the end result.

Part C. PHITS and related decompositions

The spectral analysis approaches based on PCA and SVD has been a very popular in information retrieval and link analysis (see Berry et al 1999 and Borodin et al 2001). Recently a number of researchers has pointed out that the Gaussian assumption underlying the interpretation of these models is not always an accurate represention of a more realistic structures for analysis of documents and networks. To this point (Hofmann 1999a, Hofmann 1999b) has proposed the probabilistic latent semantic analysis model, also called 'aspect' model, which performs a form of probabilistic non-negative matrix decomposition. Cohn and Chang (2000) refined this approach to model citation networks and used it to build PHITS – a probabilistic version of the Kleinberg's HITS algorithm which is based on PCA. In further development Cohn and Hofmann (2001) proposed a combined framework for modeling documents and document networks (such as world wide web or citation networks).

Tasks

1. Give a summary of probabilistic latent variable approaches based on aspect model and offer an alternative to the PCA analysis.
2. Implement the latent variable model with decompositions based on multinomial probabilities. Please follow the PHITS approach based on Cohn and Chang (2000)
3. Implement the EM algorithm for learning the model from data. You may use 'tempering' as proposed by Hofmann to avoid the problem of local optima.
4. Compare and analyze the performance of PCA and PHITS model on the link structure dataset (not microarray data) selected in Part A.