Chapter 4.26
Advanced Data Mining and Visualization Techniques with Probabilistic Principal Surfaces:
Applications to Astronomy and Genetics

Antonino Staiano
University of Napoli, "Parthenope", Italy

Lara De Vinco
Nexera S.c.p.A., Italy

Giuseppe Longo
University "Federico II" of Napoli Polo delle Scienze e della Tecnologia, Italy

Roberto Tagliaferri
University of Salerno, Italy

ABSTRACT

Probabilistic principal surfaces (PPS) is a non-linear latent variable model with very powerful visualization and classification capabilities that seem to be able to overcome most of the shortcomings of other neural tools. PPS builds a probability density function of a given set of patterns lying in a high-dimensional space that can be expressed in terms of a fixed number of latent variables lying in a latent Q-dimensional space. Usually, the Q-space is either two- or three-dimensional and thus, the density function can be used to visualize the data within it. The case in which Q = 3 allows to project the patterns on a spherical manifold that turns out to be optimal when dealing with sparse data. PPS may also be arranged in ensembles to tackle complex classification tasks. As template cases, we discuss the application of PPS to two real- world data sets from astronomy and genetics.

INTRODUCTION

The explosive growth in the quantity, quality, and accessibility of data that is currently experienced in all fields of science and human endeavor, has triggered the search for a new generation of computational theories and tools, collectively constituting the field of data mining, capable to assist humans in extracting useful information (knowledge) from huge amounts of distributed and heterogeneous data. This revolution has two main aspects: on the one hand in astronomy, as well as in high energy physics, genetics, social sciences, and in many other fields, traditional interactive data analysis and data visualization methods, have proved to be far inadequate to cope with data sets that are characterized by huge volumes and/or complexity (ten or hundreds of parameter or features per record, cf. Abello, Pardalos, & Resende, 2002, and references therein). In second place, the simultaneous analysis of hundreds of parameters may unveil previously unknown patterns that will lead to a deeper understanding of the underlying phenomena and trends.

Knowledge discovery in databases or KDD is therefore becoming of paramount importance not only in its traditional arena, but also as an auxiliary tool for almost all fields of research. In this chapter, after a short introduction on the latent variable models, we shall first focus on the visualization and classification capabilities of the spherical probabilistic principal surfaces and then on the possibility to build PPS ensembles. Finally, we shall discuss two applications in the fields of astronomy and genetics. All results have been obtained in the framework of the Astroneural collaboration: a joint project between the Department of Mathematics and Informatics of the University of Salerno and the Department of Physical Sciences of the University Federico II of Napoli. The main goal of the collaboration is to implement a user-friendly data-mining tool capable to deal with heterogeneous, high-dimensionality data sets. All software is implemented under the Matlab computing environment exploiting the LANS Pattern Recognition Matlab Toolbox (http://www.lans.ece.utexas.edu/~lans/lans/) and the Netlab Toolbox (Nabney, 2002).

LATENT VARIABLE MODELS

The goal of a latent variable model is to express the distribution $p(\mathbf{t})$ of the variable $\mathbf{t}=(t_1, \ldots, t_D)$ in terms of a smaller number of latent variables $\mathbf{x} = (x_1, \ldots, x_Q)$, where $Q < D$. To achieve this, the joint distribution $p(\mathbf{t},\mathbf{x})$ is decomposed into the product of the marginal distribution $p(\mathbf{x})$ of the latent variables and the conditional distribution $p(\mathbf{t}|\mathbf{x})$ of the data variables, given the latent variables (Bishop, 1999). Expressing the conditional distribution as a factorization over the data variables the joint distribution becomes:

$$p(\mathbf{t},\mathbf{x}) = p(\mathbf{x})\,p(\mathbf{t}\mid\mathbf{x}) = p(\mathbf{x})\prod_{d=1}^{D} p(t_d\mid\mathbf{x}). \quad (1)$$

The conditional distribution $p(\mathbf{t}|\mathbf{x})$ is then written in terms of a mapping from latent variables to data variables, so that $\mathbf{t}=\mathbf{y}(\mathbf{x};\mathbf{w})+\mathbf{u}$. $\mathbf{y}(\mathbf{x};\mathbf{w})$ is a function of the latent variable $\mathbf{x}$ with parameters $\mathbf{w}$, and $\mathbf{u}$ is an $\mathbf{x}$-independent noise process. If the components of $\mathbf{u}$ are uncorrelated, the conditional distribution for $\mathbf{t}$ will factorize as in (1). Geometrically, the function $\mathbf{y}(\mathbf{x};\mathbf{w})$ defines a manifold in data space given by the image of the latent space. The definition of the latent variable model is completed by specifying the distribution $p(\mathbf{u})$, the mapping $\mathbf{y}(\mathbf{x};\mathbf{w})$, and the marginal distribution $p(\mathbf{x})$. The type of mapping $\mathbf{y}(\mathbf{x};\mathbf{w})$ determines the specific latent variable model. The desired model for the distribution $p(\mathbf{t})$ of the data is then obtained by marginalizing over the latent variables:

$$p(\mathbf{t}) = \int p(\mathbf{t}\mid\mathbf{x})\,p(\mathbf{x})\,d\mathbf{x}.$$

Although this integration will, in general, be analytically intractable, there exist specific forms