Mixture Models for Clustering and Dimension Reduction

Verbeek, J.J.

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
Many current information processing systems have to process huge amounts of data. Often both the number of measurements as well as the number of variables in each measurement—the dimensionality of the data—is very large. Such high dimensional data is acquired in various forms: images containing thousands of pixels, documents that are represented by frequency counts of several thousands of words in a dictionary, or sound represented by measuring the energy in hundreds of frequency bands. Because so many variables are measured, a wide variety of tasks can be performed on the basis of these sorts of data. For example, images can be used to recognize hand-written digits and characters, but also to recognize faces of different people.

Information processing systems are used to perform several types of tasks, such as classification, regression and data visualization. In classification, the goal is to predict the class of new objects on the basis of a set of supervised examples. For example, in a digit recognition task the supervised examples are images of digits together with a label indicating which digit is depicted in the image. The supervised examples are used to find a function that maps new images to a prediction of the class label. Regression is similar to classification, but the goal is here to predict a continuous number rather than a discrete class label. For example, a challenging regression application would be to predict the age of a person on the basis of an image of his face. In data visualization the goal is to produce an insightful graphical display of data. For example, the results of image database queries can be presented using a two dimensional visualization, such that similar images are displayed near to each other.

In many applications where high dimensional data is used the diversity in the data considered is often limited. For example, only a limited variety of images is processed by a system that recognizes people from an image of their face. Images depicting digits or cars are not processed by such a system, or perhaps only to conclude that it does not depict a face. In general, the high dimensional data that is processed for specific tasks can be described in a more compact manner. Often, the data can either be divided into several groups or clusters, or the data can be represented using fewer numbers: the dimensionality can be reduced.

It turns out that it is not only possible to find such compact data representations, but that it is a prerequisite to successfully learn classification and regression functions from
high dimensional examples. Also for visualization of high dimensional data a more compact representation is needed, since the number of variables that can be graphically displayed is inherently limited. In this thesis we present the results of our research on methods for clustering and dimension reduction. Most of the methods we consider are based on the estimation of probabilistic mixture densities: densities that are a weighted average of several simple component densities. A wide variety of complex density functions can be obtained by combining simple component densities in a mixture.

**Layout of this thesis.** In Chapter 1 we give a general introduction and motivate the need for clustering and dimension reduction methods. We continue in Chapter 2 with a review of different types of existing clustering and dimension reduction methods.

In Chapter 3 we introduce mixture densities and the expectation-maximization (EM) algorithm to estimate their parameters. Although the EM algorithm has many attractive properties, it is not guaranteed to return optimal parameter estimates. We present greedy EM parameter estimation algorithms which start with a one-component mixture and then iteratively add a component to the mixture and re-estimate the parameters of the current mixture. Experimentally, we demonstrate that our algorithms avoid many of the sub-optimal estimates returned by the EM algorithm. Finally, we present an approach to accelerate mixture densities estimation from many data points. We apply this approach to both the standard EM algorithm and our greedy EM algorithm.

In Chapter 4 we present a non-linear dimension reduction method that uses a constrained EM algorithm for parameter estimation. Our approach is similar to Kohonen's self-organizing map, but in contrast to the self-organizing map, our parameter estimation algorithm is guaranteed to converge and optimizes a well-defined objective function. In addition, our method allows data with missing values to be used for parameter estimation and it is readily applied to data that is not specified by real numbers but for example by discrete variables. We present the results of several experiments to demonstrate our method and to compare it with Kohonen's self-organizing map.

In Chapter 5 we consider an approach for non-linear dimension reduction which is based on a combination of clustering and linear dimension reduction. This approach forms one global non-linear low dimensional data representation by combining multiple, locally valid, linear low dimensional representations. We derive an improvement of the original parameter estimation algorithm, which requires less computation and leads to better parameter estimates. We experimentally compare this approach to several other dimension reduction methods. We also apply this approach to a setting where high dimensional 'outputs' have to be predicted from high dimensional 'inputs'. Experimentally, we show that the considered non-linear approach leads to better predictions than a similar approach which also combines several local linear representations, but does not combine them into one global non-linear representation.

In Chapter 6 we summarize our conclusions and discuss directions for further research.