Chapter 12

Learning metrics

Samuel Kaski, Janne Sinkkonen, Jaakko Peltonen

12.1 Introduction

Visualization and clustering of multivariate data are usually based on mutual distances of samples, measured by heuristic means such as the Euclidean distance of vectors of extracted features. Our recently developed methods remove this arbitrariness by learning to measure important differences. The effect is equivalent to changing the metric of the data space. The laborious implicit supervision by manually tailored feature extraction will to a large extent be replaced by an automatically learned metric, while discoveries can still be made with unsupervised learning methods within the constraints set by the new metric.

It is assumed that variation of the data is important only to the extent it causes variation in *auxiliary data* c which is available paired to the primary data $\mathbf{x} \in \mathbb{R}^n$. Such auxiliary data are available at least in the settings where supervised learning methods, regression and classification, are usually applied. The difference here is that the goal is to model and understand the *primary data* and learn what is relevant there, whereas in supervised learning the sole purpose is to predict the auxiliary data.

In other words, important variation in $\mathbf{x}$ is supposed to be revealed by variation in the conditional density $p(c|\mathbf{x})$. The distance d between two close-by data points $\mathbf{x}$ and $\mathbf{x} + d\mathbf{x}$ is defined as the difference between the corresponding distributions of c, measured by the Kullback-Leibler divergence D_{KL}, i.e.

$$d_L^2(\mathbf{x}, \mathbf{x} + d\mathbf{x}) \equiv D_{\mathrm{KL}}(p(c|\mathbf{x}) \| p(c|\mathbf{x} + d\mathbf{x})) = d\mathbf{x}^T \mathbf{J}(\mathbf{x}) d\mathbf{x} \,, \tag{12.1}$$

where $\mathbf{J}(\mathbf{x})$ is the Fisher information matrix. Bankruptcy risk is an example of auxiliary data that indicates importance in the analysis of the financial states of companies: the $c = 1$ if the company goes bankrupt and $c = 0$ if it stays alive. The Kullback-Leibler divergence is locally a metric, and locality can be relaxed by extending the metric. Proximity relations (i.e., loosely speaking, the topology) of the data space are preserved, but the arbitrariness of feature selection is removed by locally re-scaling the data space to make it reflect important variation in data.

The Fisher information matrix has earlier been used to construct metrics to spaces of probability models (see, e.g., [2]). The novelty here is that the information matrix is applied in the data space to construct a new metric there.

We call the idea of measuring distances in the data space by approximations of (12.1) the learning metrics principle. In practice, the idea can be applied in two ways. One can estimate $p(c|\mathbf{x})$ first and then plug the new metric, computed from the estimates, into a standard unsupervised method. Another possibility is to more directly insert the new metric into the cost function of a suitable method. Examples of these approaches are discussed in more detail below.

12.2 Learning metrics for Self-Organizing Maps

The principle

One way of approximating (12.1) is to first compute an estimator $\hat{p}(c|\mathbf{x})$ for the conditional density, and then compute the Fisher information matrix from the estimator. In practice the distance can be computed directly from

$$d_L^2(\mathbf{x}, \mathbf{x} + d\mathbf{x}) = E_{\hat{p}(c|\mathbf{x})} \left\{ \left(d\mathbf{x}^T \frac{\partial}{\partial \mathbf{x}} \log \hat{p}(c|\mathbf{x}) \right)^2 \right\} \; ; \tag{12.2}$$

the matrix $\mathbf{J}$ need not be formed explicitly.

We have applied this principle to the Self-Organizing Map (SOM) [4]. A SOM consists of a grid of computational units, with a model vector $\mathbf{m}_i$ associated to each unit i. After the SOM has been computed these model vectors follow the input data in an ordered fashion: model vectors of close-by units on the lattice remain close-by in the input space.

The SOM algorithm iterates two steps at discrete time steps t: winner selection and adaptation. In the new metric the index w of the unit closest to the current input $\mathbf{x}(t)$ is sought by

$$w(\mathbf{x}(t)) = \arg\min_i d_L^2(\mathbf{x}(t), \mathbf{m}_i(t)) \; . \tag{12.3}$$

Here the local distance approximation is also used for non-local distances, which has turned out to be reasonably accurate in many applications.

It can be shown that the adaptation step in the steepest descent direction of the new non-Euclidean metric, i.e. at the direction of the natural gradient [2], is equivalent to the familiar SOM learning rule,

$$\mathbf{m}_i(t+1) = \mathbf{m}_i(t) + h_{wi}(t)(\mathbf{x}(t) - \mathbf{m}_i(t)) \; . \tag{12.4}$$

Here $h_{wi}(t)$ is the neighbourhood function, a decreasing function of distance on the map lattice. The neighbourhood tightens slowly with the iterations.

In the case of the bankruptcy application described below, the computational complexity is approximately doubled compared to that of a SOM in the Euclidean metric.

A demonstration with a toy data is presented in Figure 12.1. In the learning metric the SOM finds the relevant horizontal dimensions of the data and does not waste resources on representing the irrelevant vertical dimension.

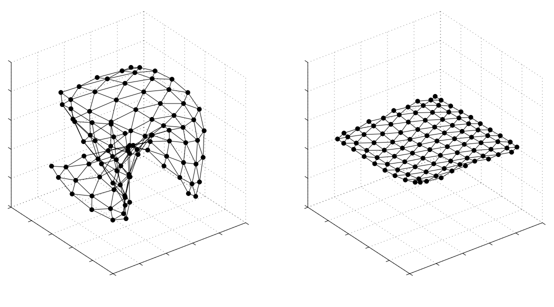
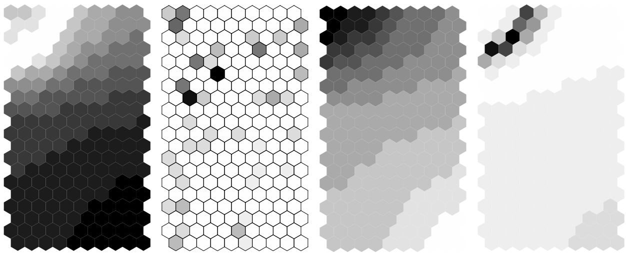


Figure 12.1: Projections of the model vectors of two SOMs representing the same data, in the Euclidean (left) and the learning metric (right). The primary data is evenly distributed within the unit cube. Only the horizontal dimension is relevant, indicated by the auxiliary data which is of constant distribution in the vertical dimension.



a **b** **c** **d**

Figure 12.2: Sample visualizations of bankruptcy data with a SOM in the learning metrics. **a** Posterior bankruptcy probability, **b** empirical ratio of healthy to bankrupt companies, **c** distribution of the values of a profitability indicator, and **d** relevance of the profitability indicator. The hexagons correspond to SOM units and light shades denote high values. Each company is located in one hexagon.

Bankruptcy Analysis

Most quantitative studies of bankruptcy have aimed at prediction, mainly by classification and probability estimation based on financial statements. A complementary approach pursued earlier in our laboratory [5, 6] is to analyze and visualize the effects of corporate behaviour on the bankruptcy risk. We have applied learning metrics, which enables us to guide the visualizations by the most important indicator of bankruptcy analysis: whether the company goes (has gone) bankrupt or not. The SOM should then emphasize the most interesting features of the financial statements, i.e. those that contribute locally to the bankruptcies. The method automatically discards irrelevant features of input data, and therefore makes it possible to incorporate potential-looking financial indicators more freely that is possible with a completely unsupervised method.

In this case study, the data consisted of financial statements from Finnish companies. Several financial indicators were extracted from the statements, and used as primary data. The auxiliary variable indicated whether the company went bankrupt within 3 years of the statement.

Gaussian mixture-based estimates for the conditional density of bankruptcy were computed and used to build the metric for SOMs. As expected, the SOMs in the new metric were significantly more accurate in modeling bankruptcy risk than SOMs in the Euclidean metric. Visually, the learning metric displays were comparable or better than Euclidean displays. Sample visualizations are shown in Figure 12.2. The learning metric makes it possible to additionally visualize the local *relevance* of each variable, one at a time, in different locations of the map display (Fig. 12.2**d**).

12.3 Discriminative clustering

A general goal of clustering is to minimize within-cluster distortion or variation, and to maximize between-cluster variation. We have applied the learning metrics principle (12.1) to clustering by replacing the distortion measure within each local cluster by a kind of within-cluster Kullback-Leibler divergence. This causes the clusters to be internally as homogeneous as possible in terms of the distribution of the auxiliary data $p(c|\mathbf{x})$. The other side of the coin is that between-cluster differences in $p(c|\mathbf{x})$ are maximized; this is the reason for coining the method *discriminative clustering.*

The clusters are defined in terms of the primary data, which enables clustering of future samples without any auxiliary data. The locality of the clusters in the primary data space makes them useful for exploratory analysis.

In vector spaces

For vectorial data, we have parameterized the clusters as softened Voronoi regions, either in the Euclidean metric or, in some applications, the inner product metric of the unit hypersphere. The principle of stochastic approximation leads to very simple on-line learning algorithms which can be interpreted as modifications of the traditional Hebb rule.

The cost function of discriminative clustering is in general the distortion

$$E = \int \sum_j y_j(\mathbf{x}) D_{\mathrm{KL}}(p(c|\mathbf{x}) \| \boldsymbol{\psi}_j)\, p(\mathbf{x})\, d\mathbf{x}\ , \tag{12.5}$$

where the $y_j(\mathbf{x})$ denote the softened Voronoi regions ($\sum_j y_j(\mathbf{x}) = 1$), and $\boldsymbol{\psi}_j$ is a region-wise prototype for the conditional distribution $p(c|\mathbf{x})$. The cost function measures the inhomogeneity of the clusters w.r.t. $p(c|\mathbf{x})$, equivalent to a conditional likelihood for finite data and hard clusters.

Our cluster memberships have been of the form $y_j(\mathbf{x}) = Z^{-1}(\mathbf{x}) \exp(-d^2(\mathbf{m}_j, \mathbf{x})/\sigma^2)$, where $d(\mathbf{m}_j, \mathbf{x})$ is the distance between $\mathbf{x}$ and the *location prototype* $\mathbf{m}_j$, measured in the original metric of the data space. Then the on-line algorithm for minimizing the distortion takes a particularly simple form. For soft Euclidean Voronoi regions, an observation $\mathbf{x}$ with auxiliary information c causes simple updates to cluster locations:

$$\mathbf{m}_j := \mathbf{m}_j + \alpha(\mathbf{x} - \mathbf{m}_j) \log \frac{\psi_{li}}{\psi_{ji}}\ . \tag{12.6}$$

The clusters j and l are chosen randomly, proportionally to the activations $y_j(\mathbf{x})$.

So far the algorithm has been applied to clustering of text documents, yeast gene expressions, and financial statements of companies [8, 9]. In each of these applications, clusters with interesting variation are found by guiding the algorithm with suitable relevance-inducing auxiliary data.

It can be shown that the clustering maximizes mutual information between the auxiliary data and the clusters interpreted as a random variable. Other mutual information maximizing clustering methods have been proposed earlier [1, 10]. Our method combines a novel on-line algorithm with the induction of a proximity-preserving quantization to a continuous data space.

Of texts

We have extended the scope of the method to text documents, under the commonly used simplifying assumption that the documents are 'bags of words', finite-length samples from a multinomial distribution. All knowledge of the document's topical content is encoded into the parameters of the distribution, and hence we cluster the distributions. Discriminative clustering applied to texts is then called Discriminative Distributional Clustering.

The clusters are now defined as Voronoi regions in the parameter space of the distributions $\mathbf{q}$, with Kullback-Leibler divergence from the prototypes as the distance measure. The Voronoi regions are softened for computational reasons by $y_j(\mathbf{q}; \mathbf{\Theta}) = Z^{-1}(\mathbf{q}) \exp(-\kappa D_{KL}(\mathbf{q}, \boldsymbol{\theta}_j))$. Here $Z(\mathbf{q})$ normalizes the memberships of $\mathbf{q}$, $\boldsymbol{\theta}_j$ is the distributional prototype of cluster j, and $\mathbf{\Theta}$ denotes all the parameters of the membership functions.

The method has been applied to scientific texts from the INSPEC database [7]. The guiding auxiliary data is composed of keywords chosen by the document authors. From these, the method is able to learn what is relevant in the full texts, which we measured with an expert categorization of the documents by informaticians.

Discriminative clustering improves the results of text clustering: The clusters are more closely related to relevant categories given by human experts, even though those categories were not used to train the models.

12.4 Discriminative clustering is vector quantization in learning metrics

We have shown [3] that at the limit of a large number of hard clusters (Voronoi regions), discriminative clustering performs a kind of vector quantization in learning metrics: The Euclidean distortion of normal vector quantization becomes replaced with the distortion computed in the new Fisher metric (12.1) defined by the conditional distributions $p(c|\mathbf{x})$ of the relevance-indicating auxiliary data. The Voronoi regions are, however, still defined in the original metric of the primary data space. This is due to the current parameterization of the clusters as the (soft) Voronoi regions of the original metric. In other words, the discriminative clustering algorithm is closely connected to the learning metrics principle.

References

[1] Suzanna Becker. Mutual information maximization: models of cortical self-organization. *Network: Computation in Neural Systems*, 7:7–31, 1996.

[2] Shun ichi Amari and Hiroshi Nagaoka. *Methods of Information Geometry*. American Mathematical Society and Oxford University Press, 2000.

[3] Samuel Kaski and Janne Sinkkonen. Principle of learning metrics for exploratory data analysis. Submitted to a journal.

[4] Samuel Kaski, Janne Sinkkonen, and Jaakko Peltonen. Bankruptcy analysis with self-organizing maps in learning metrics. *IEEE Transactions on Neural Networks*, 12:936–947, 2001.

[5] Kimmo Kiviluoto. Predicting bankruptcies with the self-organizing map. *Neurocomputing*, 21(1–3):191–201, 1998.

[6] Kimmo Kiviluoto and Pentti Bergius. Analyzing financial statements with the self-organizing map. In *Proceedings of WSOM'97, Workshop on Self-Organizing Maps, Espoo, Finland, June 4–6*, pages 362–367. Helsinki University of Technology, Neural Networks Research Centre, Espoo, Finland, 1997.

[7] Jaakko Peltonen, Janne Sinkkonen and Samuel Kaski. Discriminative clustering of text documents. Submitted to a conference.

[8] Janne Sinkkonen and Samuel Kaski. Clustering by similarity in an auxiliary space. In Kwong Sak Leung, Lai-Wan Chan, and Helen Meng, editors, *Proceedings of IDEAL 2000, Second International Conference on Intelligent Data Engineering and Automated Learning*, pages 3–8. Springer, Berlin, 2000.

[9] Janne Sinkkonen and Samuel Kaski. Clustering based on conditional distributions in an auxiliary space. *Neural Computation*, 14:217–239, 2002.

[10] Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. In *37th Annual Allerton Conference on Communication, Control, and Computing*, Urbana, Illinois, 1999.