Discriminative Clustering: Optimal Contingency Tables by Learning Metrics

Janne Sinkkonen, Samuel Kaski, and Janne Nikkilä

Helsinki University of Technology, Neural Networks Research Centre,
P.O. Box 9800, FIN-02015 HUT, Finland,
{Janne.Sinkkonen,Samuel.Kaski,Janne.Nikkila}@hut.fi,
http://www.cis.hut.fi/projects/mi

Abstract. The *learning metrics principle* describes a way to derive metrics to the data space from paired data. Variation of the primary data is assumed relevant only to the extent it causes changes in the auxiliary data. *Discriminative clustering* finds clusters of primary data that are homogeneous in the auxiliary data. In this paper, discriminative clustering using a mutual information criterion is shown to be asymptotically equivalent to vector quantization in learning metrics. We also present a new, finite-data variant of discriminative clustering and show that it builds contingency tables that detect optimally statistical dependency between the clusters and the auxiliary data. A finite-data algorithm is demonstrated to outperform the older mutual information maximizing variant.

1 Introduction

The metric of the data space determines the goodness of the results of unsupervised learning: clustering, nonlinear projection methods, and density estimation. The metric, in turn, is determined by feature extraction, variable selection, transformation, and preprocessing of the data. The principle of learning metrics aims at automating part of the process of metric selection, by learning the metric from data.

It is assumed that the data comes in pairs $(\mathbf{x}, c)$: during learning, the primary data vectors $\mathbf{x} \in \mathbb{R}^n$ are paired with auxiliary data c which in this paper are discrete classes. Important variation in $\mathbf{x}$ is supposed to be revealed by variation in the 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 to be the difference between the corresponding distributions of c, measured by the Kullback-Leibler divergence D_{KL}. It is well known (see e.g. [3]) that the divergence is locally equal to the quadratic form with the Fisher information matrix $\mathbf{J}$, i.e.

$$d_L^2(\mathbf{x}, \mathbf{x} + d\mathbf{x}) \equiv D_{\mathrm{KL}}(p(c|\mathbf{x}) \| p(c|\mathbf{x} + d\mathbf{x})) = d\mathbf{x}^T \mathbf{J}(\mathbf{x}) d\mathbf{x} \ . \tag{1}$$

The Fisher information matrix has classically appeared in the context of constructing metrics for probabilistic model families. A novelty here is that the data vector $\mathbf{x}$ is considered as the *parameters* of the Fisher information matrix, the aim being to construct a new metric into the *data space*.

The Kullback-Leibler divergence defines a metric locally, and the metric can in principle be extended to an information metric or Fisher metric to the whole data space. We call the idea of measuring distances in the data space by approximations of (1) the learning metrics principle [1, 2].

The principle is presumable useful for tasks in which there is suitable auxiliary data available, but instead of merely predicting the values of auxiliary data the goal is to analyze, explore, or mine the primary data. Charting companies based on financial indicators is one example; there the bankruptcy risk (whether the company goes bankrupt or not) is natural auxiliary data.

Learning metrics is similar to supervised learning in that the user has to choose proper auxiliary data. The difference is that in supervised learning the sole purpose is to predict the auxiliary data, whereas in learning metrics the metric is supervised while the rest of the analysis can be unsupervised, given the metric.

In this paper we analyze clustering in learning metrics, or *discriminative clustering* (earlier also called semisupervised clustering) [2]. In general, a goal of clustering is to minimize within-cluster distortion or variation, and to maximize between-cluster variation. We apply the learning metrics by measuring distortions within each cluster by a kind of within-cluster Kullback-Leibler divergence. This causes the clusters to be internally as homogeneous as possible in conditional distributions $p(c|\mathbf{x})$ of the auxiliary variable. The mutual differences between the distributions $p(c|\mathbf{x})$ of the clusters are then automatically maximized, giving a reason to call the method discriminative.

We have earlier derived and analyzed discriminative clustering with information-theoretic methods, assuming infinite amount of data. In this paper we will derive a finite-data variant and theoretical context for it, in the limit of "hard" clusters (vector quantization). It is not possible to use gradient-based algorithms for hard clusters, and hence we derive optimization algorithms for a smooth variant for which standard fast optimization procedures are then applicable.

2 Discriminative Clustering is Asymptotically Vector Quantization in Fisher Metrics

2.1 Discriminative Clustering

We will first introduce the cost function of discriminative clustering by applying the learning metrics principle to the classic vector quantization or K-means clustering.

In vector quantization the goal is to find a set of prototypes or codebook vectors $\mathbf{m}_j$ that minimizes the average distortion E caused when the data are represented by the prototypes:

$$E = \sum_j \int_{V_j} D(\mathbf{x}, \mathbf{m}_j)\ p(\mathbf{x})\ d\mathbf{x}\ , \tag{2}$$

where $D(\mathbf{x}, \mathbf{m}_j)$ is the distortion caused by representing $\mathbf{x}$ by $\mathbf{m}_j$, and V_j is the Voronoi region of the cell j. The Voronoi region V_j consists of all points that are closer to $\mathbf{m}_j$ than to any other model, that is, $\mathbf{x} \in V_j$ if

$$D(\mathbf{x}, \mathbf{m}_j) \leq D(\mathbf{x}, \mathbf{m}_k) \tag{3}$$

for all k.

The learning metrics principle is applied to (2) by introducing a set of distributional prototypes $\boldsymbol{\psi}_j$, one for each partition j, and by measuring distortions of representing the *distributions* $p(c|\mathbf{x})$ by the prototypes $\boldsymbol{\psi}_j$. The average distortion is

$$E_{KL} = \sum_j \int_{V_j} D_{KL}(p(c|\mathbf{x}), \boldsymbol{\psi}_j)\ p(\mathbf{x})\ d\mathbf{x}\ , \tag{4}$$

where distortion between distributions has been measured by the Kullback-Leibler divergence. Note that the Voronoi regions V_j are still kept local in the primary data space by defining them with respect to the Euclidean distortion (3).

The cost (4) is minimized with respect to both sets of prototypes, $\mathbf{m}_j$ and $\boldsymbol{\psi}_j$. The optimization is discussed further in Section 5.

It can be shown that minimizing (4) maximizes the mutual information between the auxiliary data and the clusters, considered as a random variable [2]. This holds even for the soft variant discussed in Section 5.

2.2 Asymptotic Connection to Learning Metrics

In this section we aim to clarify the motivation behind discriminative clustering, by deriving a connection between it and the learning metrics principle of using (1) as the distance measure. The connection is only theoretical in that it holds only for the asymptotic limit of a large number of clusters, whereas in practice the number of clusters will be small.

The asymptotic connection can be derived under some simplifying assumptions. It is assumed that almost all Voronoi regions become increasingly local when their number increases. (In singular cases, the data samples are identified with their equivalence classes having zero mutual distance.) There are always some non-compact and therefore inevitably non-local Voronoi regions at the borders of the data manifold, but it is assumed that the probability mass within them can be made arbitrarily small by increasing the number of regions. Assume further that the densities $p(c|\mathbf{x})$ are differentiable. Then the class distributions $p(c|\mathbf{x})$ can be made arbitrarily close to linear within each region V_j by increasing the number of Voronoi regions.

Let E_{V_j} denote the expectation over the Voronoi region V_j with respect to the probability density $p(\mathbf{x})$. At the optimum of the cost E_{KL}, we have $\boldsymbol{\psi}_j = E_{V_j}[p(c|\mathbf{x})]$, i.e. the parameters $\boldsymbol{\psi}_j$ are equal to the means of the conditional distribution within the Voronoi regions (see [2]; this holds even for the soft clusters).

Since $p(c|\mathbf{x})$ is linear within each Voronoi region, there exists a linear operator L_j for each V_j, for which $p(c|\mathbf{x}) = L_j\mathbf{x}$. The distributional prototypes then become

$$\boldsymbol{\psi}_j = E_{V_j}[p(c|\mathbf{x})] = E_{V_j}[L_j\mathbf{x}] = L_j E_{V_j}[\mathbf{x}] \equiv L_j\tilde{\mathbf{m}}_j = p(c|\tilde{\mathbf{m}}_j) ,$$

and the cost function becomes

$$E_{KL} = \sum_j \int_{V_j} D_{KL}(p(c|\mathbf{x}), p(c|\tilde{\mathbf{m}}_{j(\mathbf{x})}))\ p(\mathbf{x})\ d\mathbf{x} .$$

That is, given a locally linear $p(c|\mathbf{x})$, there exists a point $\tilde{\mathbf{m}}_j = E_{V_j}[\mathbf{x}]$ for each Voronoi region such that the Kullback-Leibler divergence appearing in the cost function can be measured with respect to the distribution $p(c|\tilde{\mathbf{m}}_{j(\mathbf{x})})$ instead of the average over the whole Voronoi region.

Since the Kullback-Leibler divergence is locally equal to a quadratic form of the Fisher information matrix, we may expand the divergence around $\tilde{\mathbf{m}}_j$ to get

$$E_{KL} = \sum_j \int_{V_j} (\mathbf{x} - \tilde{\mathbf{m}}_{j(\mathbf{x})})^T \mathbf{J}(\tilde{\mathbf{m}}_{j(\mathbf{x})})(\mathbf{x} - \tilde{\mathbf{m}}_{j(\mathbf{x})})\ p(\mathbf{x})\ d\mathbf{x} , \tag{5}$$

where $\mathbf{J}(\tilde{\mathbf{m}}_{j(\mathbf{x})})$ is the Fisher information matrix evaluated at $\tilde{\mathbf{m}}_{j(\mathbf{x})}$.

Note that the Voronoi regions V_j are still defined by the parameters $\mathbf{m}_j$ and in the original, usually Euclidean metric.

In summary, discriminative clustering or maximization of mutual information asymptotically finds a partitioning from the family of local Euclidean Voronoi partitionings, for which the within-cluster distortion in the Fisher metric is minimized. In other words, discriminative clustering asymptotically performs vector quantization in the Fisher metric by Euclidean Voronoi regions: Euclidean metrics define the family of Voronoi partitionings $\{V_j\}_j$ over which the optimization is done, and the Fisher metric is used to measure distortion inside the regions.

3 Estimation from Finite Data

3.1 Maximum Likelihood

Note that for finite data minimizing the cost function (4) is equivalent to maximizing

$$L = \sum_j \sum_{\mathbf{x} \in V_j} \log \psi_{j,c(\mathbf{x})} , \tag{6}$$

where $c(\mathbf{x})$ is the index of the class of the sample $\mathbf{x}$. This is the log likelihood of a piece-wise constant conditional density estimator. The estimator predicts the distribution of C to be $\boldsymbol{\psi}_j$ within the Voronoi region j. The likelihood is maximized with respect to both the $\boldsymbol{\psi}_j$ and the partitioning, under the defined constraints.

3.2 Maximum A Posteriori

The natural extension of maximum likelihood estimation is to introduce a prior and to find the maximum a posterior (MAP) estimate. The Bayesian framework is particularly natural for discriminative clustering since we are actually interested only on the resulting clusters, not the distribution of the auxiliary data within them. The class distributions can therefore be conveniently integrated out from the posterior (although seemingly paradoxical, the auxiliary data of course guides the clustering).

Denote the observed auxiliary data set by $D^{(c)}$, and the primary data set by $D^{(x)}$. We then wish to find the set of clusters $\{\mathbf{m}\}$ which maximizes the posterior

$$p(\{\mathbf{m}\}|D^{(c)}, D^{(x)}) = \int_{\{\boldsymbol{\psi}\}} p(\{\mathbf{m}\}, \{\boldsymbol{\psi}\}|D^{(c)}, D^{(x)}) d\{\boldsymbol{\psi}\} \ ,$$

or equivalently $\log p(\{\mathbf{m}\}|D^{(c)}, D^{(x)})$. Here the integration is over all $\boldsymbol{\psi}_j$.

Denote the number of classes by N_c, the number of clusters by k, and the total number of samples by N. Denote the part of the data assigned to cluster j by $D_j^{(c)}$, and the number of data samples of class i in cluster j by n_{ji}. Further denote $N_j = \sum_i n_{ji}$.

Assume the improper and separable prior $p(\{\mathbf{m}\}, \{\boldsymbol{\psi}\}) \propto p(\{\boldsymbol{\psi}\}) = \prod_j p(\boldsymbol{\psi}_j)$. Then,

$$\begin{aligned} p(\{\mathbf{m}\}|D^{(c)}, D^{(x)}) &\propto \int_{\{\boldsymbol{\psi}\}} p(D^{(c)}|\{\mathbf{m}\}, \{\boldsymbol{\psi}\}, D^{(x)}) p(\{\boldsymbol{\psi}\}) d\{\boldsymbol{\psi}\} \\ &= \prod_j \int_{\boldsymbol{\psi}_j} p(D_j^{(c)}|\{\mathbf{m}\}, \boldsymbol{\psi}_j, D^{(x)}) p(\boldsymbol{\psi}_j)\, d\boldsymbol{\psi}_j \\ &= \prod_j \int_{\boldsymbol{\psi}_j} \prod_i \psi_{ji}^{n_{ji}} p(\boldsymbol{\psi}_j)\, d\boldsymbol{\psi}_j \equiv \prod_j Q_j \ . \end{aligned}$$

We will use a conjugate (Dirichlet) prior, $p(\boldsymbol{\psi}_j) \propto \prod_i \psi_{ji}^{n_i^0 - 1}$, where $n^0 = \{n_i^0\}_i$ are the prior parameters common to all j, and $N^0 = \sum_i n_i^0$. Then the "partition-specific" density $p(D_j^{(c)}|\{\mathbf{m}\}, \boldsymbol{\psi}_j) p(\boldsymbol{\psi}_j)$ is Dirichlet with respect to ψ and the factors Q_j of the total posterior become

$$Q_j = \int_{\boldsymbol{\psi}_j} p(D_j^{(c)}|\{\mathbf{m}\}, \boldsymbol{\psi}_j, D^{(x)}) p(\boldsymbol{\psi}_j)\, d\boldsymbol{\psi}_j \propto \int_{\boldsymbol{\psi}_j} \prod_i \psi_{ij}^{n_i^0 + n_{ji} - 1}\, d\boldsymbol{\psi}_j = \frac{\prod_i \Gamma(n_i^0 + n_{ji})}{\Gamma\left(N^0 + N_j\right)} \ .$$

The log of the posterior probability then is

$$\log p(\{\mathbf{m}\}|D^{(c)}, D^{(x)}) = \sum_{ij} \log \Gamma(n_i^0 + n_{ji}) - \sum_j \log \Gamma(N^0 + N_j) \ . \qquad (7)$$

In MAP estimation this function needs to be maximized.

3.3 Asymptotic Connection to Maximization of Mutual Information

It is shown that for a fixed number of clusters, the cost function (7) of the new method approaches mutual information as the number of data samples increases.

Denote $s_{ji} \equiv n_i^0 + n_{ji} - 1$, $S_j \equiv \sum_i s_{ji} = N^0 + N_j - N_c$, and $S = \sum_j S_j$. Then,

$$\log p(\{\mathbf{m}\}|D^{(c)}, D^{(x)}) = \sum_{ij} \log \Gamma(s_{ji} + 1) - \sum_j \log \Gamma[(S_j + N_c - 1) + 1] \; . \quad (8)$$

It is straightforward to show using the Stirling approximation and Taylor approximations (Appendix A), that

$$\frac{\log p(\{\mathbf{m}\}|D^{(c)}, D^{(x)})}{S} = \sum_{ij} s_{ji}/S \log \frac{s_{ji}/S}{S_j/S} + \mathcal{O}\left(\frac{N_c k(\log S + 1)}{S}\right) \; , \quad (9)$$

where s_{ji}/S approaches p_{ji}, the probability of class i in cluster j, and S_j/S approaches p_j as the number of data samples increases. Hence, (9) approaches the mutual information, added by a constant.

4 Discriminative Clustering Optimizes Contingency Tables

Contingency tables (see [4]) are classical methods for measuring statistical dependency between discrete-valued (categorical) random variables. The categories are fixed before the analysis, and for two variables the co-occurrences of the categories in a sample are tabulated into a two-dimensional table. A classic example due to Fisher is to measure whether the order of adding milk and tea affects the taste. The first variable indicates the order of adding the ingredients, and the second whether the taste is better or worse. In medicine the other variable could indicate health status and the other one demographic groups.

The resulting contingency table is tested for dependency between the row and column variables. The literature for various kinds of tests and uses of contingency tables is extensive, see for example [4–7]. The effect of small sample sizes and/or small cell frequencies has been the subject of much controversy. Bayesian methods are principled means for coping with small data sets; below we will derive a connection between the Bayesian approach presented in [7], and our discriminative clustering method.

Given discrete-valued auxiliary data, the result of any clustering method can be analyzed as a contingency table. The possible values of the auxiliary variable correspond to columns and the clusters to rows. Clustering compresses a potentially large number of multivariate continuous-valued observations into a manageable number of categories, and the contingency table can, at least in principle, be tested for dependency. Note that the difference from the traditional use of contingency tables is that the row categories are not fixed but clustering tries to find a suitable categorization. The question here is, *is discriminative* *clustering a good way of constructing such contingency tables?* The answer is that it is optimal in the sense introduced below.

Good [7] derived a "Bayesian test" for dependency in contingency tables by computing the *Bayes factor against H*,

$$\frac{P(\{n_{ij}\}|\bar{H})}{P(\{n_{ij}\}|H)}, \tag{10}$$

where H is the hypothesis of statistical independence of the row and column categories. The probabilities are derived assuming mixtures of Dirichlet distributions as priors.

In the special case of one fixed margin (the auxiliary data) in the contingency table, and the prior defined in Section 3.2[1], the Bayes factor is

$$\begin{aligned}
&\frac{P(\{n_{ij}\}|\{n(c_i)\},\bar{H})}{P(\{n_{ij}\}|\{n(c_i)\},H)} \\
&\qquad = \frac{\prod_{i,j}\Gamma(n_{ji}+n^0)}{\prod_j (N_j+N^0)} \times \frac{\Gamma(N^0)^k}{\Gamma(n^0)^{N_c k}} \frac{\Gamma(kn^0)^{N_c}\Gamma(N+kN^0)}{\prod_i \Gamma(n(c_i)+kn^0)\Gamma(kN^0)} \\
&\qquad\qquad = p(\{\mathbf{m}\}|D^{(c)}, D^{(x)}) \times const.\ ,
\end{aligned} \tag{11}$$

where the constant does not depend on N_j or n_{ij}. Here $n(c_i)$ denotes the number of samples in the (auxiliary) class c_i. MAP estimation for discriminative clustering is thus equivalent to constructing a dependency table that results in a maximal Bayes factor, under the constraints of the model.

5 Algorithms

Optimization of both variants of discriminative clustering, the finite data version (7) and the infinite-data version (4), is hard since the gradient is zero except on the Voronoi borders. Hence gradient-based optimization algorithms are not applicable. We have earlier [2] proposed a "smoothed" infinite-data variant which can be optimized by an on-line algorithm, reviewed below. A similar smoothed variant will be introduced for MAP estimation as well.

5.1 Algorithm for Large Data Sets

Smooth parameterized *membership functions* $y_j(\mathbf{x};\{\mathbf{m}\})$) were introduced to the cost function (4). Their values vary between 0 and 1, and $\sum_j y_j(\mathbf{x}) = 1$. The smoothed cost function is

$$E'_{KL} = \sum_j \int y_j(\mathbf{x};\{\mathbf{m}\}) D_{KL}(p(c|\mathbf{x}), \boldsymbol{\psi}_j)\ p(\mathbf{x})\ d\mathbf{x}\ . \tag{12}$$

[1] In contrast to [7], we used priors with equal total amount of "prior data" for both hypotheses.

The membership functions can be for instance normalized Gaussians, $y_j(\mathbf{x}) = Z^{-1}(\mathbf{x})e^{-\|\mathbf{x}-\mathbf{m}_j\|^2/\sigma^2}$, where Z normalizes the sum to unity for each $\mathbf{x}$.

The cost function can be minimized by the following stochastic approximation algorithm. Denote the i.i.d. data pair at the on-line step t by $(\mathbf{x}(t), c(t))$ and index the (discrete) value of $c(t)$ by i, that is, $c(t) = c_i$. Draw two clusters, j and l, independently with probabilities given by the membership functions $\{y_k(\mathbf{x}(t))\}_k$. Reparameterize the distributional prototypes by the "soft-max", $\log \psi_{ji} = \gamma_{ji} - \log \sum_m \exp(\gamma_{jm})$, to keep them summed up to unity. Adapt the prototypes by

$$\mathbf{m}_j(t+1) = \mathbf{m}_j(t) - \alpha(t)\,[\mathbf{x}(t) - \mathbf{m}_j(t)] \log \frac{\psi_{li}(t)}{\psi_{ji}(t)} \tag{13}$$

$$\gamma_{jm}(t+1) = \gamma_{jm}(t) - \alpha(t)\,[\psi_{jm}(t) - \delta_{mi}] \ , \tag{14}$$

where δ_{mi} is the Kronecker delta. Due to the symmetry between j and l, it is possible (and apparently beneficial) to adapt the parameters twice for each t by swapping j and l in (13) and (14) for the second adaptation. Note that no updating takes place if $j = l$, i.e. then $\mathbf{m}_j(t+1) = \mathbf{m}_j(t)$. During learning the parameter $\alpha(t)$ decreases gradually toward zero according to a schedule that fulfills the conditions of the stochastic approximation theory.

5.2 MAP Algorithm for Finite Data Sets

In an analogous fashion to the infinite-data variant we postulate smooth membership functions $y_j(\mathbf{x}; \{\mathbf{m}\})$ that govern the assignment of the data $\mathbf{x}$ to the clusters. Then the smoothed "number" of samples of class i within cluster j becomes $n_{ij} = \sum_{c(\mathbf{x})=i} y_j(\mathbf{x})$, and the MAP cost function (7) becomes

$$\log p(\{\mathbf{m}\}|D^{(c)}, D^{(x)}) = \sum_{ij} \log \Gamma\left(n_i^0 + \sum_{c(\mathbf{x})=i} y_j(\mathbf{x})\right) - \sum_j \log \Gamma\left(N_j^0 + \sum_{\mathbf{x}} y_j(\mathbf{x})\right) . \tag{15}$$

For normalized Gaussian membership functions the gradient of the cost function with respect to the jth model vector is (Appendix B)

$$\sigma^2 \frac{\partial}{\partial \mathbf{m}_j} \log p(\{\mathbf{m}\}|D^{(c)}, D^{(x)}) = \sum_{\mathbf{x},l} (\mathbf{x} - \mathbf{m}_j) y_l(\mathbf{x}) y_j(\mathbf{x}) (L_{c(\mathbf{x}),j} - L_{c(\mathbf{x}),l}) \ , \tag{16}$$

where

$$L_{ij} \equiv \Psi(n_{ji} + n_i^0) - \Psi(N_j + N_j^0) \ .$$

Here Ψ is the digamma function, derivative of the logarithm of Γ.

The MAP estimate can then be solved with general-purpose nonlinear optimization methods. We have used the conjugate gradient algorithm.

Note that Ψ approaches the logarithm when its argument grows, and hence for large data sets the gradient approaches the average of (13) over the data and the lth membership function, with $\psi_{ji} \equiv n_{ij}/N_j$.

6 Empirical Results

The algorithm is first demonstrated with a toy example in Figure 1. The data (10,000 samples) comes from a two-dimensional spherically symmetric Gaussian distribution. The two-class auxiliary data changes only in the vertical dimension, indicating that only the vertical dimension is relevant. The algorithm learns to model only the relevant dimension.

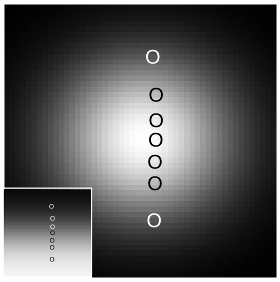

Fig. 1. A demonstration of the MAP algorithm. The probability density function of the data is shown in shades of gray and the cluster centers with circles. The conditional density of one of the two auxiliary classes is shown in the inset. (Here $\sigma = 0.4$)

As far as we know there do not exist alternative methods for precisely the same task, partitioning the primary data space to clusters that are homogeneous in terms of the auxiliary data. We have earlier compared the older mutual information maximizing variant (section 5.1) with two clustering methods: the plain mixture of Gaussians and MDA2 [8, 9], a mixture model for the joint distribution of primary and auxiliary data. For gene expression data our algorithm outperformed the alternatives [2]. Here we will add the new variant (section 5.2) to the comparison.

A random half of the Landsat satellite data set from the UCI Machine Learning Repository (36 dimensions, six classes, and 6435 samples) was partitioned into 2–10 clusters, using the six-fold class indicator as the auxiliary data.

For each number of clusters, solutions were computed for 30 values of the smoothing parameter σ, ranging from two to 100 on the logarithmic scale. All the prior parameters n_i^0 were set to unity. The models were evaluated by computing the log-posterior probability (7) of the left-out data.

The log-posterior probabilities of the validation set are presented in Figure 2. For all numbers of clusters the new algorithm performed better, having a larger edge at smaller numbers of clusters. Surprisingly, in constrast to earlier experiments with other data sets, for this data set the alternative clustering methods seem to outperform the older variant of discriminative clustering.

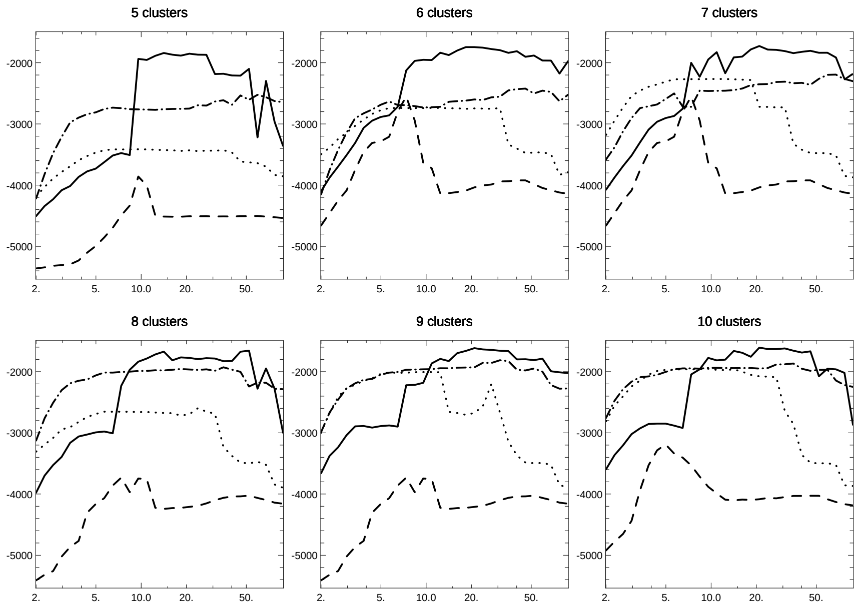

Fig. 2. The performance of the conjugate-gradient MAP algorithm *(solid line)* compared to the older discriminative clustering algorithm *(dashed line)*, plain mixture of Gaussians *(dotted line)* and MDA2, a mixture model for the joint distribution of primary and auxiliary data *(dash-dotted line)*. Sets of clusters were computed with each method with several values of the smoothing parameter σ, and the posterior log-probability (7) of the validation data is shown for a hard assignment of each sample to exactly one cluster. Results measured with empirical mutual information (not shown) are qualitatively similar. The smallest visible value corresponds to assigning all samples to the same cluster

For 4–7 clusters, the models were compared by ten-fold cross-validation. The best value for σ was chosen with validation data, in preliminary tests. The new model was significantly better for all cluster numbers (paired t test, $p < 0.001$).

7 Conclusions

In summary, we have applied the learning metrics principle to clustering, and coined the approach discriminative clustering. It was shown that discriminative clustering is asymptotically, in the limit of a large number of clusters, equivalent to clustering in Fisher metrics, with the additional constraint that the clusters are (Euclidean) Voronoi regions in the primary data space. In the earlier work [1] Fisher metrics were derived from explicit conditional density estimators for clustering with Self-Organizing Maps; discriminative clustering has the advantage that the (arbitrary) density estimator is not required.

We have derived a finite-data discriminative clustering method that maximizes the posterior probability of the cluster centroids. There exist related methods for infinite data, proposed by us and others, derived by maximizing the mutual information [2, 10, 11]. For discrete primary data there exist also finite-data generative models [12, 13]; the main difference in our methods is the ability to derive a metric to continuous primary data spaces.

Finally, we have shown that the cost function is equivalent to the Bayes factor of a contingency table with the marginal distribution of the auxiliary data fixed. The Bayes factor is the odds of the data likelihood given the hypothesis that the rows and columns are independent, vs. the alternative hypothesis of dependency. Hence, discriminative clustering can be interpreted to find a set of clusters that maximize the statistical dependency with the auxiliary data.

Acknowledgments

This work was supported by the Academy of Finland, in part by the grants 50061 and 52123.

References

1. Kaski, S., Sinkkonen, J., Peltonen, J.: Bankruptcy analysis with self-organizing maps in learning metrics. IEEE Trans. Neural Networks **12** (2001) 936–947
2. Sinkkonen, J., Kaski, S.: Clustering based on conditional distributions in an auxiliary space. Neural Computation **14** (2002) 217–239
3. Kullback, S.: Information Theory and Statistics. Wiley, New York (1959)
4. Agresti, A.: A survey of exact inference for contingency tables. Statistical Science **7** (1992) 131–153
5. Fisher, R. A.: On the interpretation of χ^2 from the contingency tables, and the calculation of *p*. J. Royal Stat. Soc. **85** (1922) 87–94
6. Freeman, G. H., Halton, J. H.: Note on an exact treatment of contingency, goodness of fit and other problems of significance. Biometrika **38** (1951) 141–149
7. Good, I. J.: On the application of symmetric Dirichlet distributions and their mixtures to contingency tables. Annals of Statistics **4** (1976) 1159–1189
8. Hastie, T., Tibshirani, R., Buja, A.: Flexible discriminant and mixture models. In Kay, J. Titterington, D. (eds): Neural Networks and Statistics. Oxford University Press (1995)
9. Miller, D. J., Uyar, H. S.: A mixture of experts classifier with learning based on both labelled and unlabelled data. In Mozer, M., Jordan, M., Petsche, T. (eds): Advances in Neural Information Processing Systems 9. MIT Press, Cambridge, MA (1997) 571–577
10. Becker, S.: Mutual information maximization: models of cortical self-organization. Network: Computation in Neural Systems **7** (1996) 7–31
11. Tishby, N., Pereira, F. C., Bialek, W.: The information bottleneck method. In: 37th Annual Allerton Conference on Communication, Control, and Computing. Urbana, Illinois (1999)
12. Hofmann, T., Puzicha, J., Jordan, M. I.: Learning from dyadic data. In: Kearns, M. S., Solla, S. A., Cohn, D. A. (eds): Advances in Neural Information Processing Systems 11. Morgan Kaufmann Publishers, San Mateo, CA (1998) 466–472

13. Hofmann, T.: Unsupervised learning by probabilistic latent semantic analysis. Mach. Learning (2001) 177–196

A Connection of MAP Estimation to Maximization of Mutual Information

The Stirling approximation $\log\Gamma(s+1) = s\log s - s + \mathcal{O}(\log s)$ applied to (8) yields

$$\begin{aligned}\log p(\{\mathbf{m}\}|D^{(c)},D^{(x)}) = \sum_{ij} s_{ji}\log s_{ji} - \sum_j S_j \log(S_j + N_c - 1) \\ - (N_c - 1)\sum_j \log(S_j + N_c - 1) + \mathcal{O}(N_c k(\log S + 1)) \ .\end{aligned}$$

The zeroth-order Taylor expansion $\log(S+n) = \log S + \mathcal{O}\left(\frac{n}{S}\right)$ gives after rearrangements, for $S_j > 1$,

$$\log p(\{\mathbf{m}\}|D^{(c)},D^{(x)}) = \sum_{ij} s_{ji}\log s_{ji} - \sum_j S_j \log S_j + \mathcal{O}(N_c k(\log S + 1)) \ .$$

Division by S then gives (9).

B Gradient of the MAP Cost Function

Denote for brevity $t_{ji} = n_{ji} + n_i^0$ and $T_j = \sum_i t_{ji}$. The gradient of (15) with respect to $\mathbf{m}_j$ is

$$\begin{aligned}\frac{\partial}{\partial \mathbf{m}_j}\log p(\{\mathbf{m}\}|D^{(c)},D^{(x)}) &= \sum_{il}\sum_{c(\mathbf{x})=i}\frac{\partial}{\partial \mathbf{m}_j}y_l(\mathbf{x})\Psi(t_{li}) - \sum_{\mathbf{x},l}\frac{\partial}{\partial \mathbf{m}_j}y_l(\mathbf{x})\Psi(T_l) \\ &= \sum_{\mathbf{x},l}\frac{\partial}{\partial \mathbf{m}_j}y_l(\mathbf{x})[\Psi(t_{l,c(\mathbf{x})}) - \Psi(T_l)] \ .\end{aligned}$$

It is straightforward to show that for normalized Gaussian membership functions

$$\frac{\partial}{\partial \mathbf{m}_j}y_l(\mathbf{x}) = \frac{1}{\sigma^2}(\mathbf{x}-\mathbf{m}_j)(\delta_{lj} - y_l(\mathbf{x}))y_j(\mathbf{x}) \ .$$

Substituting this to the gradient gives

$$\sigma^2\frac{\partial}{\partial \mathbf{m}_j}\log p(\{\mathbf{m}\}|D^{(c)},D^{(x)}) = \sum_{\mathbf{x},l}(\mathbf{x}-\mathbf{m}_j)(\delta_{lj}-y_l(\mathbf{x}))y_j(\mathbf{x})[\Psi(t_{l,c(\mathbf{x})})-\Psi(T_l)] \ . \tag{17}$$

The final form (16) for the gradient results from applying the identity

$$\sum_l(\delta_{lj} - y_l)y_j L_l = \sum_l y_l y_j (L_j - L_l) \ ,$$

to (17).