# Discriminative Clustering

Samuel Kaski
*Helsinki University of Technology, Neural Networks Research Centre*
*P.O. Box 9800*
*FIN-02015 HUT, Finland*
*samuel.kaski@hut.fi*

**Summary.** Discriminative clustering (DC) uses auxiliary data to define what is relevant in the primary data. It partitions the continuous primary data space to local clusters that have maximally homogeneous (categorical) auxiliary data. The task has several interpretations: searching for maximally predictive clusters, clusters that maximize mutual information with the auxiliary data, clusters for which contingency tables detect optimally dependency with the auxiliary data, or K-means clusters in the so-called Fisher or learning metric. DC can be applied to adjust the resolution of an existing classification, or to guide clustering with auxiliary data.

## 1. Introduction

The task is to cluster continuous primary data $\mathbf{x} \in \mathbb{R}^n$ in a way that the clusters become relevant for or informative of the discrete auxiliary data $c$, i.e., capable of predicting $p(c|\mathbf{x})$. Compared to modeling of the joint distribution (Hastie et al., 1995), the clusters become more informative about $c$. Classical distributional clustering (Pereira et al., 1993) searches for informative clusters as well, but is only applicable to discrete data. The difference from classification is that the goal is to explore the primary data, not merely to predict the classes.

Motivation for the model comes from a central problem in clustering: After the (feature extraction and) metric is fixed, it is not possible to distinguish between relevant and irrelevant variation in data. Assuming a suitable dependent variable such as a class labeling is available, however, it might be advantageous to use it to guide clustering. We have earlier suggested to incorporate such supervision to an exploratory task by computing a supervised metric, coined the learning metric (Kaski et al., 2001; Kaski and Sinkkonen, to appear; Sinkkonen and Kaski, 2002).

Charting companies based on financial indicators is one example application for discriminative clustering (Kaski et al., 2001). The binary variable indicating whether the company went bankrupt or not is natural auxiliary data. Others include clustering of gene activity patterns (Sinkkonen and Kaski, 2002), where functional classes are the auxiliary data, and segmentation of customers based on what they buy.

In this paper the DC model is reviewed briefly, and some new results are presented on regularization and extension to continuous auxiliary data.

## 2. The discriminative clustering model

The goal is to partition the primary data space into clusters that (i) are local and (ii) have homogeneous auxiliary data. Locality is enforced by defining the clusters as (Euclidean) Voronoi regions in the primary data space: $\mathbf{x}$ belongs to cluster $j$, $\mathbf{x} \in V_j$, if $\|\mathbf{x} - \mathbf{m}_j\| \leq \|\mathbf{x} - \mathbf{m}_k\|$ for all $k$. The Voronoi regions are uniquely determined by the parameters $\{\mathbf{m}_j\}$.

Homogeneity is enforced by assigning a *distributional prototype* $\boldsymbol{\psi}_j$, a prototype density, to each Voronoi region $j$, and searching for partitionings capable of predicting auxiliary data with the prototypes. The resulting model is a piecewise-constant generative model of $p(c|\mathbf{x})$, with the log likelihood

$$(1) \quad L = \sum_j \sum_{\mathbf{x} \in V_j} \log \psi_{j,c(\mathbf{x})} \ ,$$

where $c(\mathbf{x})$ is the class of sample $\mathbf{x}$. The probability of class $i$ within Voronoi region $V_j$ is predicted to be $\psi_{j,i}$. The motivation for this model is that asymptotically

$$\frac{1}{N}L \to -\sum_j \int_{V_j} D_{KL}(p(c|\mathbf{x}), \boldsymbol{\psi}_j)p(\mathbf{x})d\mathbf{x} + \text{const.} \ , \tag{2}$$

where $D_{KL}$ is the Kullback-Leibler divergence between the prototype and the observed distribution of auxiliary data, and $N$ is the number of samples. This is the cost function of K-means clustering or vector quantization with the $D_{KL}$ as distortion. In this sense, maximizing the likelihood of the model maximizes the distributional homogeneity of the clusters.

The likelihood is to be maximized with respect to the both sets of prototypes, $\mathbf{m}_j$ and $\boldsymbol{\psi}_j$. Since the goal is to cluster the primary data, however, the prototype distributions are actually not needed. If the (log) posterior probability is maximized instead of the likelihood, it can be shown (Sinkkonen et al., 2002) that for suitable uninformative priors,

$$\begin{aligned} \log p(\{\mathbf{m}_j\}|D^{(c)}, D^{(x)}) &= \log \int_{\{\boldsymbol{\psi}_j\}} p(\{\mathbf{m}_j\}, \{\boldsymbol{\psi}_j\}|D^{(c)}, D^{(x)})d\{\boldsymbol{\psi}_j\} \\ &= \sum_{ij} \log\Gamma(n_{ji} + n^0) - \sum_j \log\Gamma(n_{j\cdot} + N_c n^0) + \mathit{const.} \end{aligned} \tag{3}$$

Here $D^{(c)}$ is the observed auxiliary and $D^{(x)}$ the primary data set, $n_{ji}$ is the count of data from (auxiliary) category $i$ in cluster $j$, $n_{j\cdot} = \sum_i n_{ji}$, $N_c$ is the number of auxiliary data classes, and $n^0$ comes from the Dirichlet prior.

It can be shown that maximizing (2) maximizes the mutual information between the auxiliary data and the clusters considered as a random variable (Sinkkonen and Kaski, 2002). A connection to the so-called learning metrics is derived in (Kaski and Sinkkonen, to appear): in the learning metric the class distribution changes isotropically, and DC is asymptotically K-means clustering in this metric, with the constraint that the Voronoi regions are Euclidean.

The objective function (3) can be optimized by gradient methods after smoothing the cluster assignments by normalized Gaussians. We have used the conjugate gradient algorithm.

## 3. Regularization

Since DC in effect is K-means in a new metric, it is expected to suffer from the "dead unit problem" as well. The number of data in the clusters can be equalized by multiplying the middle term in (3) by a constant $\beta > 1$. Asymptotically the resulting extra term is the entropy.

Another potential problem in DC is that since it models solely the conditional probability $p(c|\mathbf{x})$, it cannot explicitly take into account uncertainty in $\mathbf{x}$. DC can be complemented with a generative mixture-type model for $\mathbf{x}$, to model the joint density by $p(c, \mathbf{x}|\{\mathbf{m}_j\}\{\boldsymbol{\psi}_j\}) = p(c|\mathbf{x}, \{\mathbf{m}_j\}, \{\boldsymbol{\psi}_j\})p(\mathbf{x}|\{\mathbf{m}_j\})$. If $p(\mathbf{x}|\{\mathbf{m}_j\})$ comes from the so-called classification mixture, the objective function has an interpretation as a compromise between K-means and DC,

$$\log p(\{\mathbf{m}_j\}|D^{(c)}, D^{(x)}) = MAP_{DC} - \lambda E_{VQ} \ ,$$

where $MAP_{DC}$ is given by (3) and $E_{VQ}$ is the cost function of K-means clustering. Alternatively, $E_{VQ}$ can be replaced by the log likelihood of a standard mixture of Gaussians. Changing the value of $\lambda$ shifts the focus of the clustering between DC and K-means. This possibility for a tunable compromise distinguishes the model from standard models of the joint density.

## 4. Associative clustering: Extension to continuous auxiliary data

Discriminative clustering optimizes contingency tables (Sinkkonen et al., 2002), that is, data cross-tabulated in a matrix of clusters (rows) vs. categories of auxiliary data (columns).

Formally, DC maximizes the Bayes factor (Good, 1976) of the hypothesis of dependent vs. independent rows in the contingency table. The column margins are fixed. In other words, DC finds such clusters that the rows and columns of the contingency table become as dependent as possible.

The interpretation suggests a generalization of DC to continuous auxiliary data. A set of clusters is postulated to each of the two continuous spaces, and the clusters are optimized to maximize the Bayes factor. The total number of data samples is fixed.

It can be shown (details omitted for brevity) that the (log) Bayes factor becomes

$$\log BF = \sum_{ji} \log \Gamma(n_{ji} + n^0) - \sum_{j} \log \Gamma(n_{j\cdot} + Ln^0) - \sum_{i} \log \Gamma(n_{\cdot i} + Kn^0) + const.$$

where $K$ and $L$ are the numbers of clusters in the two spaces, and the third term is the only difference from the objective function of DC; the function is now symmetric with respect to the two spaces. The model can be optimized in the same way as DC.

## 5. Experiments

We verified that DC does what it promises by computing the empirical mutual information of clusters and auxiliary data for two standard data sets. The basic discriminative clustering (DC), the equalized variant (entDC), combination of DC and VQ (DC-VQ), and combination of DC and mixture of Gaussians (DC-MoG) were compared to alternative clustering methods: standard K-means (VQ), mixture of Gaussians (MoG), and a joint mixture model (MDA2; Hastie et al., 1995). The mutual information was larger for DC (Table 1). The difference between the variants of DC was usually not significant but the difference from the others was. (Significant difference from the **best**, t-test over separate test sets, $P < 0.01$; almost significant, $P < 0.05$.)

***Table 1. Mutual information between clusters and auxiliary data***

| Data | DC | entDC | DC-VQ | DC-MoG | VQ | MoG | MDA2 |
|---|---|---|---|---|---|---|---|
| Letter | 1.89 | **1.98** | 1.91 | 1.97 | 0.95 | 0.97 | 1.68 |
| TIMIT | 1.49 | 1.52 | 1.52 | **1.53** | 1.37 | 1.36 | 1.51 |

The following task would be an ideal application for associative clustering: Cluster customers and products based on their properties, such that the customer clusters are informative about what the customers buy, and product clusters are popular for certain types of customers.

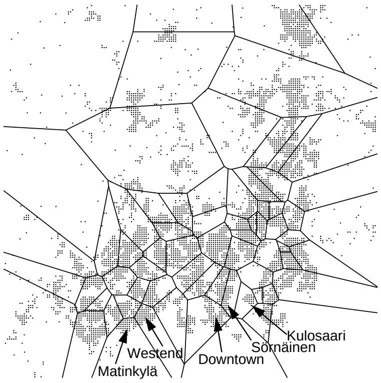

***Figure 1. Helsinki area partitioned into demographically homogeneous regions***

Since such applications are typically confidential, we demonstrate the idea by associative clustering of coordinates of small geographic squares (primary data) and a set of 146 demographic variables at the location (auxiliary data). Figure 1 shows the 70 primary data clusters in Helsinki area, where dots denote squares with more than 10 inhabitants. Many of the clusters and cluster borders have a clear interpretation; examples include the well-off Westend and the neighboring suburb Matinkylä, and the former industrial area Sörnäinen close to downtown and the embassy area Kulosaari.

## 6. Discussion

Discriminative clustering is a new step towards learning from data what is relevant or interesting. It helps bridge the gap between unsupervised learning (descriptive modeling or modeling of $p(\mathbf{x})$) and supervised learning (predictive modeling or modeling of $p(c|\mathbf{x})$). More work is still needed on connections with recent works on "clustering with side data," and on generalizing the idea to other types of data and other kinds of models besides clustering.

## Acknowledgments

I thank Janne Sinkkonen and Arto Klami for help with the simulations, Ricardo Vigário for translation, the Academy of Finland for grant 52123, and Xtract Oy and Statistics Finland for collaboration with the demographics. DC has been developed together with Janne Sinkkonen.

## REFERENCES

Good, I. J. (1976). On the application of symmetric Dirichlet distributions and their mixtures to contingency tables. Annals of Statistics, 4, 1159–1189.

Hastie, T., Tibshirani, R., and Buja, A. (1995). Flexible discriminant and mixture models. In Neural Networks and Statistics (eds. J. Kay and D. Titterington). Oxford University Press.

Kaski, S. and Sinkkonen, J. (to appear). Principle of learning metrics for data analysis. The Journal of VLSI Signal Processing-Systems.

Kaski, S., Sinkkonen, J., and Peltonen, J. (2001). Bankruptcy analysis with self-organizing maps in learning metrics. IEEE Transactions on Neural Networks, 12, 936–947.

Pereira, F., Tishby, N., and Lee, L. (1993). Distributional clustering of English words. In Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics, 183–190. ACL, Columbus, OH.

Sinkkonen, J. and Kaski, S. (2002). Clustering based on conditional distributions in an auxiliary space. Neural Computation, 14, 217–239.

Sinkkonen, J., Kaski, S., and Nikkilä, J. (2002). Discriminative clustering: Optimal contingency tables by learning metrics. In Proceedings of the 13th European Conference on Machine Learning (eds. T. Elomaa, H. Mannila, and H. Toivonen), 418–430. Berlin, Springer.

## RÉSUMÉ

*Le groupement discriminatif (discriminative clustering, DC) utilise des données auxiliaires pour définir ce qui est important dans les données primaires. Il sépare l'espace continu de données primaires en partitions locales de sorte qu'elles presentent une homogeneité maximale dans les données (catégorique) auxiliaires associées. Le processus a plusieurs interprétations: la recherche des groupements les plus prévisibles, des groupements qui maximisent l'information conjointe avec les données auxiliaires, des groupements pour lequels les tables de contingence détectent, d'une façon optimale, une dépendance avec les donées auxiliaires, ou encore des groupes definis par la méthode de k-means en utilisant des métriques (de Fisher) qui apprènent. DC peut etre utilisé pour ajuster la résolution d'une classification existante, ou bien pour guider le groupement de données en s'appuyant sur des données auxiliaires appropriées.*