REGULARIZED DISCRIMINATIVE CLUSTERING

Samuel Kaski, Janne Sinkkonen, and Arto Klami
Neural Networks Research Centre
Helsinki University of Technology
P.O. Box 9800, FIN-02015 HUT, Finland
{samuel.kaski,janne.sinkkonen,arto.klami}@hut.fi

Abstract. A generative distributional clustering model for continuous data is reviewed and methods for optimizing and regularizing it are introduced and compared. Based on pairs of auxiliary and primary data, the primary data space is partitioned into Voronoi regions that are maximally homogeneous in terms of auxiliary data. Then only variation in the primary data associated with variation in the auxiliary data influences the clusters. Because the whole primary space is partitioned, new samples can be easily clustered in terms of primary data alone. In experiments, the approach is shown to produce more homogeneous clusters than alternative methods. Two regularization methods are demonstrated to further improve the results: An entropy-type penalty for unequal cluster sizes, and the inclusion of a K-means component to the model. The latter can alternatively be interpreted as special kind of joint distribution modeling where the emphasis between discrimination and unsupervised modeling of primary data can be tuned.

INTRODUCTION

Models exist for discovering components underlying co-occurrences of nominal variables [2, 3, 8], and for the joint distribution $p(c, \mathbf{x})$ of continuous $\mathbf{x} \in \mathbb{X} \subset \mathbb{R}^n$ and discrete data c [6, 7, 10]. We consider the related task of clustering the continuous primary data by conditional modelling such that the clusters become "relevant for" or "informative of" the discrete auxiliary data, i.e., capable of predicting $p(c|\mathbf{x})$. The discriminative approach is expected (and indeed found) to result in clusters more informative about c than those obtained by modeling the joint distribution. The continuity of $\mathbf{x}$ distinguishes the setting from that of (classic) distributional clustering [12, 16].

The task, coined *discriminative clustering* (DC), is different from classification in that the number of clusters need not be equal to the number of "classes" c. The goal is to discover clusters, and the clusters may represent combinations of classes or parts of classes, depending on the application. In DC, the derived cluster structure of the $\mathbb{X}$-space is the primary outcome, even to the degree that the distributional parameters predicting $p(c|\mathbf{x})$ within a cluster are integrated out.

The main application area for DC is in data exploration or mining. Alternatively, when c is interpreted as an existing probabilistic partitioning of $\mathbb{X}$, DC can be used to alter the coarseness of the partitioning.

A prototypical application is partitioning the customers of a company on the basis of background data ($\mathbf{x}$; including e.g. coordinates of residence, age, etc.), by grouping existing customers into clusters that are informative of the buying behavior across several product categories (c). New real or potential customers can then be clustered even before they have made their first purchases. Other potential applications include finding prototypical gene expressions to refine existing functional classification of genes [13], clustering of financial statements to discover different ways to descend into bankruptcy, and partitional clustering in general when a variable c is available to automatically guide the clustering.

In this paper a model for discriminative clustering and a Bayesian objective function for its optimization are reviewed. The model cannot be optimized directly by gradient-based algorithms, and we show that complementing conjugate gradient with a smoothing of partitions gives comparable results to the much more time-consuming simulated annealing. The model is additionally regularized in two alternative ways: by penalizing from unequal cluster sizes, or by adding a term interpretable as K-means to the cost function. The latter is equivalent to generative modeling of the full joint distribution $p(c, \mathbf{x})$ of primary and auxiliary data, but also interpretable as a tunable compromise between modeling of $p(\mathbf{x})$ and $p(c|\mathbf{x})$.

In experiments, all the proposed models outperform alternative mixture-based models in their task, and both regularization methods outperform unregularized DC.

DISCRIMINATIVE CLUSTERING

We will start by reviewing the basic discriminative clustering model [9, 13]. Its goal is to partition the primary data space into clusters that are (i) local and (ii) homogeneous and predictive in terms of auxiliary data. (The connection between homogeneity and predictivity of the clusters is detailed below.) Locality is enforced by defining the clusters as 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* $\boldsymbol{\psi}_j$ to each Voronoi region j, and by searching for partitionings capable of predicting auxiliary data with the prototypes. The prototypes represent multinomial distributions over the auxiliary data, and are parameterized by ψ_{ji}, the probability of class i within Voronoi region j. The resulting model is a piecewise-constant model of $p(c|\mathbf{x})$, with the log likelihood

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

Asymptotically for large data

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

where D_{KL} is the Kullback-Leibler divergence between the prototype and the observed distribution of auxiliary data. This is the cost function of K-means clustering or Vector Quantization (VQ) with the distortion measured by D_{KL}. In this sense, maximizing the likelihood of the model maximizes the distributional homogeneity of the clusters.

It can be shown that maximizing (2) is equivalent to maximizing the mutual information between the auxiliary variable and the partitioning, which is a connection to earlier models that use the empirical mutual information as a clustering criterion [1, 13].

For small data sets, empirical mutual information is a severely biased estimate of the within-cluster homogeneity. An alternative [14], potentially better behaving form of discriminative clustering is obtained by marginalizing the likelihood (1), as reviewed next. It turns out that the distributional prototypes $\{\boldsymbol{\psi}_j\}$ can be analytically integrated out from the posterior distribution to leave only the parameters $\{\mathbf{m}_j\}$ of the Voronoi regions, which is convenient given the goal of partitioning the primary space.

MAP Estimation of Clusters of DC

The improper prior $p(\{\mathbf{m}_j\}, \{\boldsymbol{\psi}_j\}) \propto p(\{\boldsymbol{\psi}_j\}) = \prod_j p(\boldsymbol{\psi}_j)$ is used, where the factors $p(\boldsymbol{\psi}_j) \propto \prod_i \psi_{ji}^{n_i^0 - 1}$ are Dirichlet priors with the parameters n_i^0 common to all j.

The auxiliary data is denoted by $D^{(c)}$, and the primary data by $D^{(x)}$. We then wish to find the set of clusters $\{\mathbf{m}_j\}$ which maximizes the posterior (the integration is over all the $\boldsymbol{\psi}_j$)

$$\mathit{MAP}_{DC} = p(\{\mathbf{m}_j\}|D^{(c)}, D^{(x)}) = \int_{\{\boldsymbol{\psi}_j\}} p(\{\mathbf{m}_j\}, \{\boldsymbol{\psi}_j\}|D^{(c)}, D^{(x)}) d\{\boldsymbol{\psi}_j\} \propto$$

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

$$\propto \prod_j \int_{\boldsymbol{\psi}_j} \prod_i \psi_{ji}^{n_i^0 + n_{ji} - 1}\, d\boldsymbol{\psi}_j = \prod_j \frac{\prod_i \Gamma(n_i^0 + n_{ji})}{\Gamma\left(N^0 + N_j\right)} \ . \tag{3}$$

Here n_{ji} is the number of samples of class i in cluster j, $N_j = \sum_i n_{ji}$, and $N^0 = \sum_i n_i^0$.

OPTIMIZATION

MAP_{DC} itself cannot be optimized by a gradient algorithm because the gradient would be affected only by samples at the (typically zero-probability) border of the clusters. We have therefore resorted to maximizing the logarithm of a smoothed variant of (3) with the conjugate gradient algorithm [14]. The smoothed "number" of samples is $n_{ji} = \sum_{c(\mathbf{x})=i} y_j(\mathbf{x})$, where $c(\mathbf{x})$ is the class of $\mathbf{x}$ and $y_j(\mathbf{x})$ is a smoothed cluster "membership function," defined by $y_j(\mathbf{x}) = Z(\mathbf{x})^{-1} \exp(-\|\mathbf{x} - \mathbf{m}_j\|^2/\sigma^2)$ with Z such that $\sum_j y_j(\mathbf{x}) = 1$, and σ governing the degree of smoothing. In the experiments smoothing is used only for optimization, not in evaluation of the clustering results.

Alternatively, the objective function (3) can be optimized directly by simulated annealing (SA). The above-described smoothed optimization method is compared with SA in the experimental section of this paper. In each iteration of SA, a candidate step is generated by making small random displacements to the prototype vectors. The step is accepted if it increases the value of the objective function. Even if it decreases the objective function it is accepted, with a probability that is a decreasing function of the change in the objective function.

We used Gaussian displacements having the covariance matrix $\sqrt{T}\sigma^2\mathbf{I}$, where $\mathbf{I}$ is the identity matrix. Here $T = 1/(log(t) + 10)$ is the decreasing temperature parameter, a function of the iteration step t. The parameter σ was chosen with a validation set in preliminary experiments. A displacement step that decreases the objective function by ΔE is accepted with the probability $\exp(-\Delta E/T_A)$, where T_A decreases linearly to zero.

REGULARIZATION

Two regularization methods for the marginalized DC (3) are introduced in this section. The first is a straightforward attempt to improve optimization, while the latter is interpretable as joint distribution modeling. Such an explicit modeling of the "covariates" (here $\mathbf{x}$) may improve discrimination, especially with small data sets [11].

Favoring Equal Cluster Sizes

In the first regularization method, equal distribution of data into the clusters is favored, which is useful especially for avoiding "dead clusters" after bad initialization. The "equalized" objective function is

$$\log MAP_{DC} = \sum_{ij} \log \Gamma(n_i^0 + n_{ji}) - (1 + \lambda_{EQ}) \sum_j \log \Gamma(N^0 + N_j) \,, \qquad (4)$$

where $\lambda_{EQ} > 0$. As the number of data samples increases, (4) divided by N approaches mutual information plus λ_{EQ} times the entropy of the clusters (plus a term that does not depend on the parameters; proof omitted for brevity). Hence, the larger λ_{EQ} is, the more solutions with roughly equal numbers of samples in the clusters are favored.

Modeling the Primary Data Marginal

Discriminative methods that model the conditional probability $p(c|\mathbf{x})$ may benefit from the regularizing effects of modeling the marginal $p(\mathbf{x})$ as well. To investigate whether this is the case with DC, we complemented it with a generative Gaussian mixture-type model for $p(\mathbf{x})$. The full joint distribution model then is

$$p(c, \mathbf{x}|\{\mathbf{m}_j\}\{\boldsymbol{\psi}_j\}) = p(c|\mathbf{x}, \{\mathbf{m}_j\}, \{\boldsymbol{\psi}_j\})p(\mathbf{x}|\{\mathbf{m}_j\}) \tag{5}$$

Uniquely, both factors of (5) are parameterized by the same centroids $\{\mathbf{m}_j\}$. As is made explicit in (7) and (8), the special kind of parameterization makes it possible to interpret (5) as a tunable compromise between modeling $p(\mathbf{x})$ and $p(c|\mathbf{x})$.

We present two alternative forms for the model of $p(\mathbf{x})$. The first defines $p(\mathbf{x}|\{\mathbf{m}_j\})$ piece-wise for the Voronoi regions as (unnormalized) Gaussians: For $\mathbf{x} \in V_j$,

$$p(\mathbf{x}|\{\mathbf{m}_j\}) = Z(\{\mathbf{m}_j\})^{-1} e^{-\lambda_{VQ}\|\mathbf{x}-\mathbf{m}_j\|^2} , \tag{6}$$

where $\lambda_{VQ} > 0$. The model is also interpretable as a "classification mixture" [4]. Despite the piecewise definition, the density is everywhere continuous with respect to $\mathbf{x}$, for the borders of Voronoi regions are always half-way between the cluster prototypes. If the normalization factor $Z(\{\mathbf{m}_j\})$ is interpreted as a prior, the model for $p(\mathbf{x})$ appears in the total likelihood as a term representing the traditional K-means cost.

The second investigated model for $p(\mathbf{x})$ is a standard mixture of isotropic Gaussians with covariances $\sigma^2_{MoG}\mathbf{I}$ and location parameters equal to the $\{\mathbf{m}_j\}$ of DC.

For brevity, details are below derived only for the simpler model (6).

MAP Estimation of Clusters of Joint Model. With the (improper) prior

$$p(\{\mathbf{m}_j\}, \{\boldsymbol{\psi}_j\}) \propto Z(\{\mathbf{m}_j\})p(\{\boldsymbol{\psi}_j\}) = Z(\{\mathbf{m}_j\}) \prod_j p(\boldsymbol{\psi}_j) ,$$

the posterior (3) gets the extra factor

$$\prod_j Z(\{\mathbf{m}_j\})p(D^{(x)}|\{\mathbf{m}_j\}) = \prod_j \prod_{\mathbf{x}\in V_j} \exp(-\lambda_{VQ}\|\mathbf{x}-\mathbf{m}_j\|^2) ,$$

and the log posterior of the joint model becomes

$$\log p(\{\mathbf{m}_j\}|D^{(c)}, D^{(x)}) \propto \sum_{ij} \log\Gamma(n_i^0 + n_{ji}) - \sum_j \log\Gamma(N^0 + N_j) - \sum_{j;\mathbf{x}\in V_j} \lambda_{VQ}\|\mathbf{x}-\mathbf{m}_j\|^2 . \tag{7}$$

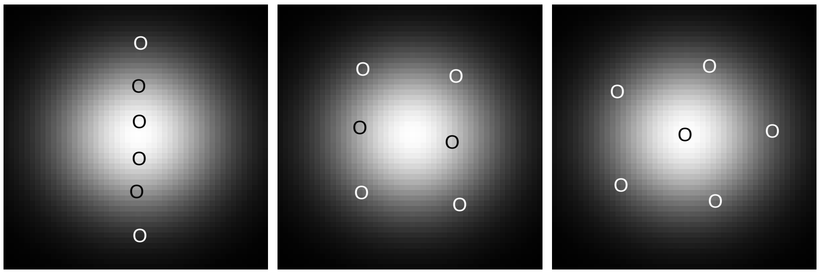

Figure 1: The VQ-regularized discriminative clustering (DC) model of (7) makes a compromise between the plain DC and ordinary K-means (VQ). From the viewpoint of plain DC (λ_{VQ}=0; left), only the vertical dimension is relevant as the distribution of the binary auxiliary data c was made to change monotonically and only in that direction. A compromise representation for the data is found at λ_{VQ}=0.02 (middle). The algorithm turns into ordinary VQ when $\lambda_{VQ} \rightarrow \infty$ (right). Circles denote the Voronoi region centroids $\{\mathbf{m}_j\}$ and gray shades the density $p(\mathbf{x})$.

Interpretations. The model has a Bayesian interpretation as a two-stage inference process: The posterior distribution of $\{\mathbf{m}_j\}$ is first inferred based on the primary data $D^{(x)}$. The posterior distribution is then used as a prior for the second, discriminative clustering stage. In the Bayesian context the prior restricts the complexity of the model.

The first K-means step can alternatively be interpreted as a regularization term of the cost function. The regularization interpretation is

$$MAP \equiv \log p(\{\mathbf{m}_j\}|D^{(c)}, D^{(x)}) = MAP_{DC} - \lambda_{VQ} E_{VQ} \; , \tag{8}$$

where E_{VQ} is the quantization error (that is, the cost function) of K-means clustering. A change in the value of λ_{VQ} makes the focus of the clustering shift between DC and VQ. In practice the value of λ_{VQ} will be chosen using a validation set.

EXPERIMENTS

Toy Demonstration. The Voronoi region centers $\{\mathbf{m}_j\}$ of the VQ-regularized model (7) are shown in Figure 1 for three different values of the parameter λ_{VQ}. The data (10,000 samples) were from an isotropic 2D Gaussian with a vertically varying $p(c|\mathbf{x})$. For small values of λ_{VQ}, the original cost function of discriminative clustering is minimized, and the clusters represent only the direction of the $\mathbb{X}$-space where the conditional distribution $p(c|\mathbf{x})$ changes. When λ_{VQ} increases the clusters gradually start to represent all variation in $\mathbf{x}$, converging to the K-means solution.

Real-Life Data. We compared the **plain discriminative clustering model and the regularized variants** on two data sets, with the final performance of the models measured by (3). The closest alternative mixture models have been included for reference. Since the effects of regularization were expected to be most apparent for small data sets, the data were split into a number of smaller subsets on which a set of independent tests were made. The DC models were optimized by conjugate gradients.

The Letter Recognition data from the UCI Machine Learning Repository (16 dimensions, 26 classes, and 20,000 samples) was split into five subsets. Two-fold modeling and testing for each subset gave a total of ten repetitions of ten-cluster solutions. The width of the mixture components and smoothing, and the regularization parameters were selected by 5-fold cross-validation within each learning set. The parameters $\{\mathbf{m}_j\}$ were initialized to a random set of training samples. (Results with the K-means initialization appearing in Table 1 are from experiments described later.)

The second data set consisted of 99,983 samples from the TIMIT collection, with 12 cepstral components as the primary data $\mathbf{x}$ and 41 phonemic classes as values of c. Since the set was larger it was divided into more (ten) subsets, resulting in twenty repetitions (with parameters within each repetition selected by 3-fold cross validation).

The best regularized methods were significantly better than plain discriminative clustering, which in turn produced better discriminative clusters than the reference methods (columns "Letter rand" and "TIMIT rand" in Table 1).

In Figure 2, the effect of tuning the compromise between K-means and DC in VQ-regularization is shown. As expected, increasing λ_{VQ} in general shifts the solution from optimizing the posterior probability (3) towards optimizing the K-means error. The new statistically almost significant finding is the slanting L-form: slight regularization improves the predictive power (the DC cost) of the clusters for the test set.

The **two optimization algorithms** are compared in Table 2. The number of clusters was halved to keep simulated annealing computationally man-

Method	Letter rand	Letter VQ	TIMIT rand	TIMIT VQ
DC	-4961.9	-4816.9	-12981	-12780
DC-VQ	-4933.4	-4779.5	-12905	-12767
DC-MoG	**-4857.9**	-4784.6	**-12866**	**-12722**
DC-EQ	-4864.1	**-4699.8**	-12942	-12757
MDA2	-5206.4	-5280.8	-13012	-12989
MoG	-6174.9	-6210.8	-13515	-13494
VQ	-6194.9	-6194.9	-13487	-13487

TABLE 1: Comparison of discriminative clustering (DC) and its regularized versions DC-VQ (7), DC-MoG ((7) with mixture of Gaussians model), and DC-EQ (4) on two data sets, Letter Recognition and TIMIT. **Best** posterior probability (3); significantly worse (t-test, $p < 0.01$) almost significantly worse ($p < 0.05$). Mixture of Gaussians (MoG), plain K-means (VQ), and joint density model MDA2 [7] have been included for reference. The results are presented for both random and K-means (VQ) initialization.

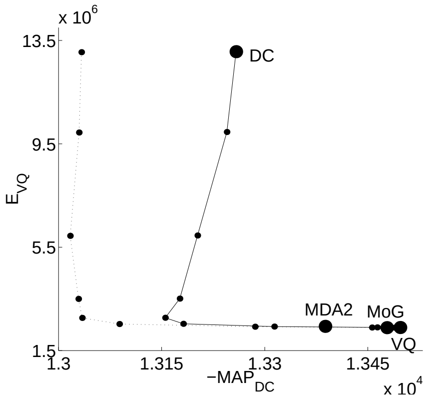

Figure 2: The effect of tuning the VQ-regularization on the two components of the cost: K-means cost (E_{VQ}) and predictive power (eqn 3; $\mathrm{MAP}_{\mathrm{DC}}$), on clusters found from the TIMIT data. Small dots on the curves: VQ-regularized DCs with varying parameter λ_{VQ}; large dots from left to right: plain DC, MDA2, mixture of Gaussians (MoG), plain K-means (VQ); solid line: test set; dashed line: learning set. Results are averages over cross-validation runs, and for computational reasons the parameter σ of the DC runs was not cross-validated but kept constant.

ageable. The smoothed conjugate gradient algorithm achieved about equally good costs as simulated annealing; SA seems to be slightly better for the TIMIT and CG for the letter data.

Finally we studied, by repeating the ten-cluster experiments, whether **replacing the random initialization with K-means** would improve the results and reduce variation between the data sets. The results of all DC variants improved significantly (columns "Letter VQ" and "TIMIT VQ" in Table 1). Regularized versions were still the best, but their relative goodness depended on the data.

DISCUSSION

An algorithm for distributional clustering of continuous data paired to a discrete variable was reviewed and extended. With a prototype distribution of the discrete variable associated to each Voronoi region of the primary

	Letter		TIMIT	
	CG	SA	CG	SA
DC	-5134.3	-5307.4	-13323	-13602
DC-VQ	-5128.3	-5115.4	-13209	-13194
DC-MoG	**-5075.6**	-5172.0	-13202	-13219
DC-EQ	-5157.5	-5105.9	-13185	**-13172**

TABLE 2: Comparison of the optimization algorithms. CG: Smoothing with conjugate gradients; SA: simulated annealing. Key: see Table 1.

data space, the regions are optimized to "predict" the discrete data well. In experiments the method produced more discriminative clusters than other methods.

The core DC model for $p(c|\mathbf{x})$ is very close to models proposed earlier for classification (RBF; [10]). In DC, however, the goal is to discover clusters, not to separate the data into the fixed pre-defined classes. The main outcome are clusters of $\mathbf{x}$, even to the extent that we were able to marginalize out the parameters producing predictions of c.

This paper contains three main new results. (i) The fast optimization of smoothed Voronoi regions by conjugate gradients produces clusters comparable to those obtained by the considerably more time-consuming simulated annealing. (ii) The two regularization methods, equalization of the cluster sizes and shifting the model towards a joint distribution model, improve the results compared to plain DC. No conclusion could be made of the relative goodness of the two methods. (iii) K-means initialization is superior to initialization by random data.

Regularization by joint distribution modeling is interpretable as the inclusion of a term of K-means quantization error in the cost function. The number of parameters in the regularized models is independent of the regularization parameter λ_{VQ}, and in this sense the model complexity is fixed. A regularized model therefore makes a compromise, tunable by λ_{VQ}, in representing variation of $\mathbf{x}$ associated to changes in $p(c|\mathbf{x})$ (the DC task), and in representing all variation isotropically (the K-means task). In the experiments with regularization, performance on learning data is not impaired while test set performance improves significantly. For some reason, therefore, allocating resources to model $p(\mathbf{x})$ improves generalization with respect to $p(c|\mathbf{x})$.

An adjustable combination of two mixture models was recently proposed for joint modeling of terms and links in text documents [5]. Here a similar combination improved a discriminative (conditional-density) model. The joint distribution modeling approach also makes it possible to treat primary data samples lacking the corresponding auxiliary part as partially missing data, along the lines proposed for classification tasks [15].

Finally, the improvement obtained by K-means initialization hints at an optimization by starting with standard clustering and tuning it gradually towards DC.

Acknowledgments. This work was supported by the Academy of Finland, grant 52123.

REFERENCES

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

[2] D. Blei, A. Y. Ng and M. I. Jordan, "Latent Dirichlet allocation," **Journal of Machine Learning Research**, vol. 3, pp. 993–1022, 2003.

[3] W. Buntine, “Variational extensions to EM and multinomial PCA,” in T. Elomaa, H. Mannila and H. Toivonen (eds.), **Proceedings of the ECML'02, 13th European Conference on Machine Learning**, Berlin: Springer, 2002, Lecture Notes in Artificial Intelligence 2430, pp. 23–34.

[4] G. Celeux and G. Govaert, “A Classification EM algorithm for clustering and two stochastic versions,” **Computational Statistics & Data Analysis**, vol. 14, pp. 315–332, 1992.

[5] D. Cohn and T. Hofmann, “The missing link—a probabilistic model of document content and hypertext connectivity,” in T. Leen, T. Dietterich and V. Tresp (eds.), **Advances in Neural Information Processing Systems 13**, Cambridge, MA: MIT Press, pp. 430–436, 2001.

[6] T. Hastie and R. Tibshirani, “Discriminant Analysis by Gaussian Mixtures,” **Journal of the Royal Statistical Society Series B: Methodological**, vol. 58, pp. 155–176, 1996.

[7] T. Hastie, R. Tibshirani and A. Buja, “Flexible Discriminant and Mixture Models,” in J. Kay and D. Titterington (eds.), **Neural Networks and Statistics**, Oxford University Press, 1995.

[8] T. Hofmann, “Unsupervised Learning by Probabilistic Latent Semantic Analysis,” **Machine Learning**, vol. 42, pp. 177–196, 2001.

[9] S. Kaski and J. Sinkkonen, “Principle of learning metrics for data analysis,” **The Journal of VLSI Signal Processing-Systems for Signal, Image, and Video Technology, Special Issue on Data Mining and Biomedical Applications of Neural Networks**, Accepted for publication.

[10] D. J. Miller and H. S. Uyar, “A Mixture of Experts Classifier with Learning Based on Both Labelled and Unlabelled Data,” in M. Mozer, M. Jordan and T. Petsche (eds.), **Advances in Neural Information Processing Systems 9**, Cambridge, MA: MIT Press, pp. 571–577, 1997.

[11] A. Y. Ng and M. I. Jordan, “On Discriminative vs. Generative classifiers: A comparison of logistic regression and naive Bayes,” in T. Dietterich, S. Becker and Z. Ghahramani (eds.), **Advances in Neural Information Processing Systems 14**, Cambridge, MA: MIT Press, 2002.

[12] F. Pereira, N. Tishby and L. Lee, “Distributional clustering of English words,” in **Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics**, Columbus, OH: ACL, pp. 183–190, 1993.

[13] J. Sinkkonen and S. Kaski, “Clustering based on conditional distributions in an auxiliary space,” **Neural Computation**, vol. 14, pp. 217–239, 2002.

[14] J. Sinkkonen, S. Kaski and J. Nikkilä, “Discriminative Clustering: Optimal Contingency Tables by Learning Metrics,” in T. Elomaa, H. Mannila and H. Toivonen (eds.), **Proceedings of the ECML'02, 13th European Conference on Machine Learning**, Berlin: Springer, 2002, Lecture Notes in Artificial Intelligence 2430, pp. 418–430.

[15] M. Szummer and T. Jaakkola, “Kernel expansions with unlabeled examples,” in T. Leen, T. Dietterich and V. Tresp (eds.), **Advances in Neural Information Processing Systems 13**, Cambridge, MA: MIT Press, pp. 626–632, 2001.

[16] N. Tishby, F. C. Pereira and W. Bialek, “The Information Bottleneck Method,” in **Proceedings of the 37th Annual Allerton Conference on Communication, Control, and Computing**, Urbana, Illinois, pp. 369-377, 1999.