Technical Report A60, Helsinki University of Technology,
Laboratory of Computer and Information Science, Espoo, Finland, August 2000.

Semisupervised clustering based on conditional distributions in an auxiliary space

Janne Sinkkonen and Samuel Kaski
Neural Networks Research Centre
Helsinki University of Technology
P. O. Box 5400, 02015 HUT, Finland
samuel.kaski@hut.fi

Abstract

We study the problem of learning groups or categories that are local in the continuous *primary* space, but homogeneous by the distributions of an associated auxiliary random variable over a discrete *auxiliary* space. Assuming variation in the auxiliary space is meaningful, categories will emphasize similarly meaningful aspects of the primary space. From a data set consisting of pairs of primary and auxiliary items, the categories are learned by minimizing a Kullback–Leibler divergence-based distortion between (implicitly estimated) distributions of the auxiliary data, conditioned on the primary data. Still, the categories are solely defined in terms of the primary space. An on-line algorithm resembling the traditional Hebb-type competitive learning is introduced for learning the categories. Minimizing the distortion criterion turns out to be equivalent to maximizing the mutual information between the categories and the auxiliary data. In addition, connections to density estimation and to the related Information Bottleneck principle are outlined. In a case study text documents are clustered by the similarity of keywords associated to the documents.

1 Introduction

Clustering algorithms and their goals vary, but it is common to aim at clusters that are relatively homogeneous while data in different clusters is dissimilar. The results depend heavily on the criterion of similarity. The difficult problem of selecting a suitable criterion is commonly addressed by feature extraction and variable selection methods. Recently, metrics have additionally been derived by fitting a generative model to the data and using information-geometric methods for extracting a metric from the model (Hofmann, 2000; Jaakkola and Haussler, 1999; Tipping, 1999).

We study the related case in which additional useful information exists about the data items during the model fitting process. The information is available as *auxiliary samples* c_k; they form pairs $(\mathbf{x}_k, c_k)$ with the primary samples $\mathbf{x}_k$. In this paper $\mathbf{x}_k \in \mathbb{R}^n$ and the c_k is multinomial. The extra information may, for example, be class labels or keywords like in our case study.

It is assumed that differences in the auxiliary data indicate what is important in the primary data space. More precisely, the difference between samples $\mathbf{x}_k$ and $\mathbf{x}_l$ is significant if the corresponding values c_k and c_l are different. The usefulness of this assumption depends, of course, on the choice of the auxiliary data.

Since the relationship between the auxiliary data and the primary data is stochastic, we get a better description of the difference between values $\mathbf{x}$ and $\mathbf{x}'$ by measuring differences between the *distributions* of c, given $\mathbf{x}$ and $\mathbf{x}'$. The conditional densities $p(c|\mathbf{x})$ and $p(c|\mathbf{x}')$ are not known, however. Only the set of sample pairs $\{(\mathbf{x}_k, c_k)\}_k$ is available. Because our aim is to minimize within-cluster dissimilarities, this similarity criterion leads to clusters that are homogeneous in terms of the (estimated) distributions $p(c|\mathbf{x})$.

In order to retain the potentially useful structure of the primary space, we only use the auxiliary data to indicate importance, and define the clusters in terms of localized basis functions within the primary space. Such a clustering can then be used later for new samples from the primary data space even when the corresponding auxiliary samples are not available.

Very loosely speaking, our aim is to preserve the topology of the primary space, but measure distances by similarity in the auxiliary space.

Clearly almost any kind of paired data is applicable, but only good auxiliary data improves clustering. If the auxiliary data is closely related to the goal of the clustering task, as for example a performance index would be, then the auxiliary data guides the clustering to emphasize the important dimensions of the primary data space, and to disregard the rest. This automatic relevance detection is the main practical motivation for the present work.

We have earlier constructed, by fitting a set of basis functions to a paired data set, an explicit local metric that is an approximation to the distances in the auxiliary space (Kaski and Sinkkonen, 2000). In the present work we will show that when the goal is clustering (or, more exactly, vector quantization), there is no need to construct a metric explicitly but it is possible to directly optimize a cost function that measures within-cluster dissimilarity in terms of distortion in the auxiliary space. The result is defined in the input space, and the clusters are, in terms of the distributions of the auxiliary data, as homogeneous as possible.

We will additionally show that minimizing the within-cluster distortion is equivalent to maximizing the mutual information between the basis functions used for defining the clusters (interpreted as a multinomial random variable) and the auxiliary data. Maximization of mutual information has been previously used for constructing neural representations (Becker, 1996; Becker and Hinton, 1992).

In another related work, the information bottleneck (Pereira et al., 1993; Tishby et al., 1999), data is also clustered by maximizing mutual information with a variable describing relevance. The main difference from our work is that the bottleneck method treats discrete data items whereas we create the categories by partitioning a continuous input space.

2 Clustering based on the Kullback-Leibler divergence

We seek to cluster items $\mathbf{x}$ of the data space by utilizing the information within a set of pairs $(\mathbf{x}_k, c_k)$ of data. The set consists of paired samples of two random variables. The vector-valued random variable X takes values $\mathbf{x} \in \mathbb{X} \subset \mathbb{R}^n$, and the c_k (or sometimes just c) are values of the multinomial random variable C. We wish to keep the clusters local with respect to $\mathbf{x}$, but measure similarities between the samples $\mathbf{x}$ by the differences of the corresponding conditional distributions $p(c|\mathbf{x})$. These distributions are unknown and will be implicitly estimated from the data.

Vector quantization (VQ) or, equivalently, K-means clustering, is one approach to categorization. In VQ the goal is to minimize the average distortion E between the data and the prototypes or codebook vectors $\mathbf{m}_j$, defined by

$$E = \sum_j \int y_j(\mathbf{x}) D(\mathbf{x}, \mathbf{m}_j)\; p(\mathbf{x})\; d\mathbf{x}\; . \tag{1}$$

Here $D(\mathbf{x}, \mathbf{m}_j)$ denotes the measure of distortion between $\mathbf{x}$ and $\mathbf{m}_j$, and $y_j(\mathbf{x})$ is the cluster membership function that fulfills $0 \leq y_j(\mathbf{x}) \leq 1$ and $\sum_j y_j(\mathbf{x}) = 1$. In the classic "hard" vector quantization the membership function is binary valued: $y_j(\mathbf{x}) = 1$ if $D(\mathbf{x}, \mathbf{m}_j) \leq D(\mathbf{x}, \mathbf{m}_i)$ for all i, and $y_j(\mathbf{x}) = 0$ otherwise. Such functions define a partitioning of the space into discrete cells, Voronoi regions, and the goal of learning is to find the partitioning that minimizes the average distortion. If the membership functions $y_j(\mathbf{x})$ may attain any values between zero and one, the approach may be called soft vector quantization (Nowlan, 1990, has studied a maximum likelihood solution).

We measure distortions D as differences between the distributions $p(c|\mathbf{x})$ and model distributions. The measure of the differences will be the Kullback–Leibler divergence, defined for two discrete-valued distributions with event probabilities $\{p_i\}$ and $\{\psi_i\}$ as $D_{KL}(p, \psi) \equiv \sum_i p_i \log(p_i/\psi_i)$. In our case, the first distribution is the multinomial distribution in the auxiliary space that corresponds to the data $\mathbf{x}$, namely $p_i \equiv p(c_i|\mathbf{x})$. The second distribution is the prototype; let us denote the jth prototype by $\boldsymbol{\psi}_j$.

When the Kullback–Leibler distortion measure is plugged into the error function of VQ, equation (1), the average distortion becomes

$$E_{KL} = \sum_j \int y_j(\mathbf{x}) D_{KL}(p(c|\mathbf{x}), \boldsymbol{\psi}_j)\; p(\mathbf{x})\; d\mathbf{x}\; . \tag{2}$$

Instead of the distortions between the vectorial samples and vectorial prototypes as in (1), we now compute point-wise distortions between the distributions $p(c|\mathbf{x})$ and the prototypes $\boldsymbol{\psi}_j$. The prototypes are distributions in the auxiliary space.

The average distortion will be minimized by parameterizing the functions $y_j(\mathbf{x})$, and optimizing the distortion with respect to the cluster membership parameters and the prototypes. When the parameters of $y_j(\mathbf{x})$ are denoted by $\boldsymbol{\theta}_j$, the average distortion can be written as

$$E_{KL} = -\sum_{i,j} \int y_j(\mathbf{x}; \boldsymbol{\theta}_j) \log \psi_{ji}\; p(c_i, \mathbf{x})\; d\mathbf{x} + \text{const.}\; , \tag{3}$$

where the constant is independent of the parameters.

The membership functions $y_j(\mathbf{x}; \boldsymbol{\theta}_j)$ can be interpreted as conditional densities $p(v_j|\mathbf{x}) \equiv y_j(\mathbf{x})$ of a multinomially distributed random variable V that indicates the cluster identity. The value of the random variable V will be denoted by $v \in \{v_j\}$ and the value of the random variable C corresponding to the multinomially distributed auxiliary distribution will be denoted by $c \in \{c_i\}$. Given $\mathbf{x}$, the choice of the cluster v does not depend on the c. In other words, C and V are conditionally independent: $p(c, v|\mathbf{x}) = p(c|\mathbf{x})p(v|\mathbf{x})$. It follows that $p(c, v) = \int p(c|\mathbf{x})p(v|\mathbf{x})p(\mathbf{x})d\mathbf{x}$.

It can be shown (see Appendix 1) that if the membership distributions are of the normalized exponential form

$$y_j(x; \boldsymbol{\theta}_j) = \frac{\exp f(\mathbf{x}; \boldsymbol{\theta}_j)}{\sum_l \exp f(\mathbf{x}; \boldsymbol{\theta}_l)}\; , \tag{4}$$

then the gradient of E_{KL} with respect to the parameters $\boldsymbol{\theta}_l$ becomes

$$\frac{\partial E_{KL}}{\partial \boldsymbol{\theta}_l} = \sum_{i,j} \int \frac{\partial f(\mathbf{x}; \boldsymbol{\theta}_l)}{\partial \boldsymbol{\theta}_l} \log \frac{\psi_{ji}}{\psi_{li}}\; p(c_i, v_j, v_l, \mathbf{x})\; d\mathbf{x}\; . \tag{5}$$

The prototypes $\boldsymbol{\psi}_j$ are probabilities of multinomial distributions, and therefore they must fulfill the conditions $0 \leq \psi_{ji} \leq 1$ and $\sum_i \psi_{ji} = 1$. We will incorporate these conditions into our model by reparameterizing the prototypes as follows:

$$\log \psi_{ji} \equiv \gamma_{ji} - \log \sum_m e^{\gamma_{jm}} \ . \tag{6}$$

Any choice of the values of the new parameters γ_{ji} fulfills the conditions posed on the $\boldsymbol{\psi}_j$, and hence unconstrained optimization of the γ_{ji} corresponds to constrained optimization of the $\boldsymbol{\psi}_j$.

The gradient of the average distortion (3) with respect to the new parameters of the prototypes is

$$\frac{\partial E_{KL}}{\partial \gamma_{lm}} = \sum_i \int (\psi_{lm} - \delta_{mi})\, p(c_i, v_l, \mathbf{x})\ d\mathbf{x} \ , \tag{7}$$

where the Kronecker symbol $\delta_{mi} = 1$ when $m = i$ and $\delta_{mi} = 0$ otherwise.

The average distortion can be minimized with stochastic approximation, by sampling from $y_j(\mathbf{x})y_l(\mathbf{x})p(c_i, \mathbf{x}) = p(v_j, v_l, c_i, \mathbf{x})$. This leads to an on-line algorithm in which the following steps are repeated for $t = 0, 1, \ldots$ with $\alpha(t)$ gradually decreasing towards zero:

1. At the step t of stochastic approximation, draw a data sample $(\mathbf{x}(t), c(t))$. Assume that the value of $c(t)$ is c_i; this defines the value of i in the following steps.
2. Draw two basis functions, j and l, from the multinomial distribution having probabilities $\{y_k(\mathbf{x}(t))\}_k$.
3. Adapt the parameters $\boldsymbol{\theta}_l$ and γ_{lm}, $m = 1, \ldots, N_c$, by

$$\boldsymbol{\theta}_l(t+1) = \boldsymbol{\theta}_l(t) - \alpha(t) \left[\frac{\partial f(\mathbf{x}; \boldsymbol{\theta}_l)}{\partial \boldsymbol{\theta}_l} \right]_{\boldsymbol{\theta}_l = \boldsymbol{\theta}_l(t)} \log \frac{\psi_{ji}}{\psi_{li}} \tag{8}$$

$$\gamma_{lm}(t+1) = \gamma_{lm}(t) - \alpha(t)(\psi_{lm} - \delta_{mi}) \ , \tag{9}$$

where N_c is the number of possible values of the random variable C. Due to the symmetry between j and l, it is possible to adapt the parameters twice for one t by swapping j and l in (8) and (9) for the second adaptation. Note that $\boldsymbol{\theta}_l(t+1) = \boldsymbol{\theta}_l(t)$ if $j = l$.

In stochastic approximation the α should fulfill the conditions $\sum_t \alpha(t) = \infty$ and $\sum_t (\alpha(t))^2 < \infty$. In practice we have used piecewise-linear decreasing schedules.

Below we will consider two special cases. In the demonstrations in Figure 1 the basis functions are normalized Gaussians in the Euclidean space $\mathbb{X} = \mathbb{R}^n$. In the second case in Section 3, text documents mapped onto a hypersphere $\mathbb{X} = \mathbb{S}^n$ are clustered by using normalized von Mises–Fisher distributions (Mardia, 1975) as the basis functions.

For Gaussians parameterized by their locations $\boldsymbol{\theta}_l$ and having a diagonal covariance matrix in which the variance σ^2 is equal in each dimension, $f(\mathbf{x}; \boldsymbol{\theta}_l) = -\|\mathbf{x} - \boldsymbol{\theta}_l\|^2 / 2\sigma^2$, and

$$\frac{\partial f(\mathbf{x}; \boldsymbol{\theta}_l)}{\partial \boldsymbol{\theta}_l} = \frac{1}{\sigma^2}(\mathbf{x} - \boldsymbol{\theta}_l) \ . \tag{10}$$

The von Mises–Fisher (vMF) distribution is an analogue of the Gaussian distribution on a hypersphere (Mardia, 1975) in that it is the maximum entropy distribution when the first two moments are fixed. The density of an n-dimensional vMF distribution is

$$\text{vMF}(\mathbf{x}; \boldsymbol{\theta}) = \frac{1}{Z(\kappa)} \exp \kappa \frac{\mathbf{x}^T \boldsymbol{\theta}}{\|\boldsymbol{\theta}\|} \ . \tag{11}$$

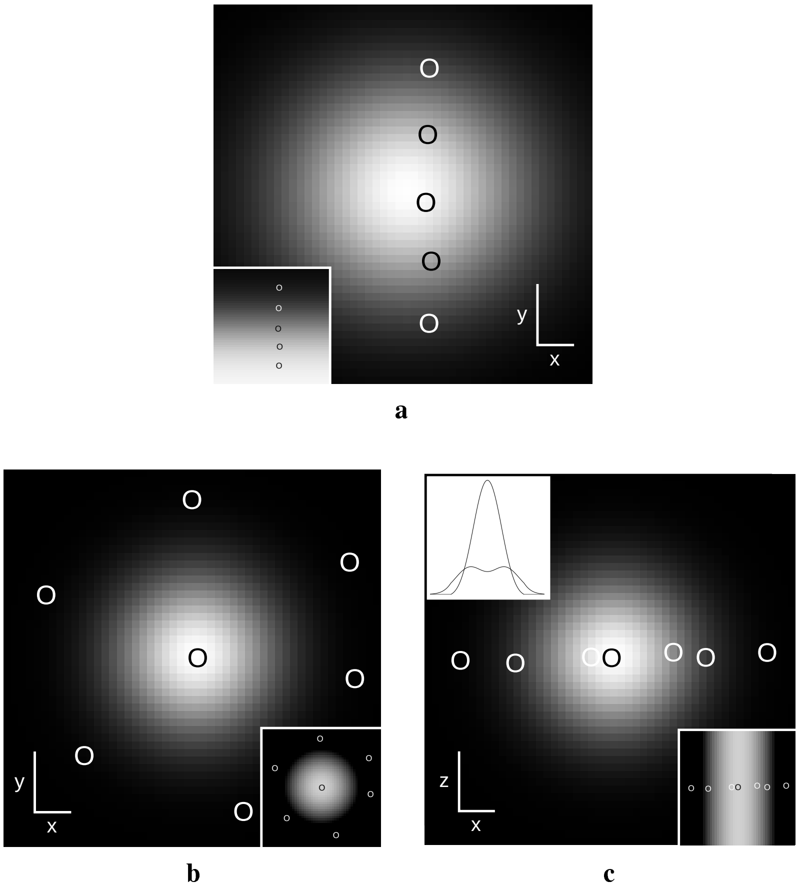

Figure 1: The location parameters $\boldsymbol{\theta}_j$ (small circles) of the Gaussian basis functions of two models, optimized for two-dimensional and three-dimensional data sets. The shades of gray at the background depict densities from which the data was sampled. **a**: 2D data, with roughly Gaussian $p(\mathbf{x})$. The inset shows the conditional density $p(c_0|\mathbf{x})$ that is monotonically decreasing as a function of the y-dimension. The model has learned to represent only the dimension on which $p(c|\mathbf{x})$ changes. **b** and **c**: Two projections of 3D data, with a symmetric Gaussian $p(\mathbf{x})$. The insets show the conditional density $p(c_0|\mathbf{x})$, which decreases monotonically as a function of a two-dimensional radius, and stays constant with respect to the orthogonal third dimension z. The one-dimensional cross section describing $p(c_0|\mathbf{x})$ and $p(c_1|\mathbf{x})$ as a function of the two-dimensional radius is shown in the inset of **c**. The model has learned to represent only variation in the direction of the radius and along the dimensions x and y, and discards the dimension z as irrelevant.

Here the parameter vector $\boldsymbol{\theta}$ represents the mean direction vector. The normalizing coefficient $Z(\kappa) \equiv (2\pi)^{\frac{1}{2}n} I_{\frac{1}{2}n-1}(\kappa)/\kappa^{\frac{1}{2}n-1}$ is not relevant here but it will be used in the mixture model of Section 3. The function $I_r(\kappa)$ is the modified Bessel function of the first kind and order r.

In the clustering algorithm described above we will use vMF basis functions $f(\mathbf{x}; \boldsymbol{\theta}_l) = \kappa \mathbf{x}^T \boldsymbol{\theta}_l / \|\boldsymbol{\theta}_l\|$ with a constant dispersion κ. Then,

$$\frac{\partial f(\mathbf{x}; \boldsymbol{\theta}_l)}{\partial \boldsymbol{\theta}_l} = \kappa(\mathbf{x} - \mathbf{x}^T \boldsymbol{\theta}_l \boldsymbol{\theta}_l / \|\boldsymbol{\theta}_l\|^2) / \|\boldsymbol{\theta}_l\| \ .$$

The norm $\|\boldsymbol{\theta}_l\|$ does not affect f, and we may normalize $\boldsymbol{\theta}_l$, whereby the gradient becomes

$$\frac{\partial f(\mathbf{x}; \boldsymbol{\theta}_l)}{\partial \boldsymbol{\theta}_l} = \kappa(\mathbf{x} - \mathbf{x}^T \boldsymbol{\theta}_l \boldsymbol{\theta}_l) \ . \tag{12}$$

2.1 Connections to mutual information and density estimation

It is shown in Appendix 2 that at the minimum of the distortion E_{KL} the prototype $\boldsymbol{\psi}_l$ takes the form

$$\psi_{li} = p(c_i | v_l) \ . \tag{13}$$

Hence, the distortion (3) can be expressed as

$$E_{KL} = -\sum_{i,j} p(c_i, v_j) \log \frac{p(c_i, v_j)}{p(c_i) p(v_j)} + \text{const.} = -I(C; V) + \text{const.} \ , \tag{14}$$

where $I(C; V)$ denotes the mutual information between the random variables C and V. Thus, minimizing the average distortion E_{KL} is equivalent to maximizing the mutual information between the auxiliary variable C and the cluster memberships V.

Minimization of the distortion has a connection to density estimation as well; the details are described in Appendix 3. It can be shown that minimization of E_{KL} minimizes an upper limit of the mean Kullback-Leibler divergence between the real distribution $p(c|\mathbf{x})$ and a certain estimate $\hat{p}(c|\mathbf{x})$. The estimate is

$$\hat{p}(c_i | \mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp \sum_j y_j(\mathbf{x}; \boldsymbol{\theta}_j) \log \psi_{ji} \tag{15}$$

where $Z(\mathbf{x})$ is a normalizing coefficient selected such that $\sum_i \hat{p}(c_i|\mathbf{x}) = 1$.

For Gaussian basis functions, the upper limit becomes tight when σ approaches zero and the cluster membership functions become binary. Then the cluster membership functions approach indicator functions of Voronoi regions. Note that the solution cannot be computed in practice for $\sigma = 0$; smooth vector quantization with $\sigma > 0$ results in a compromise in which the solution is tractable but only an upper limit of the error will be minimized.

Furthermore, the mean divergence can be expressed in terms of the divergence of joint distributions as follows:

$$E_X\{D_{KL}(p(c|\mathbf{x}), \hat{p}(c|\mathbf{x}))\} = D_{KL}(p(c, \mathbf{x}), \hat{p}(c|\mathbf{x}) p(\mathbf{x})) \ . \tag{16}$$

Here $E_X\{\cdot\}$ denotes the expectation over values of X and the latter D_{KL} is the divergence of the joint distribution. The expression (16) is a cost function of an estimate of the conditional probability $p(c|\mathbf{x})$. Intuitively speaking, by minimizing (16) resources are not wasted in estimating the marginal $p(\mathbf{x})$ but all resources are concentrated on estimating $p(c|\mathbf{x})$.

It can further be shown that maximum likelihood estimation of the model $\hat{p}(c|\mathbf{x})$ using a finite data set is asymptotically equivalent to minimizing (16).

In summary, the algorithm presented in this paper computes a certain kind of estimate of the conditional distribution $p(c|\mathbf{x})$.

2.2 Related works

2.2.1 Competitive learning

It is well known that the early works on competitive learning or adaptive feature detectors (e.g., Didday, 1976; Grossberg, 1976; Nass and Cooper, 1975; Pérez, et al., 1975) have a close connection to vector quantization (Gersho, 1979; Gray, 1984; Makhoul et al., 1985) and K-means clustering (Forgy, 1965; MacQueen, 1967). The neurons in a competitive-learning network are parameterized by vectors describing the synaptic weights, denoted by $\mathbf{m}_j$ for neuron j. In the simplest models the activity of a neuron due to external inputs is a nonlinear function f of the inputs $\mathbf{x}$ multiplied by the synaptic weights, $f(\mathbf{x}^T\mathbf{m}_j)$. The activities of the neurons compete; the activity of each neuron reduces the activity of the others via negative feedback.

If the competition is of the winner-take-all type (Kaski and Kohonen, 1994) only the neuron with the largest $f(\mathbf{x}^T\mathbf{m}_j)$ remains active. Each neuron therefore functions as a feature detector that detects whether the input comes from a particular domain of the input space. During Hebbian-type learning the neurons gradually specialize in representing different types of domains. For recent more detailed accounts see, for example (Kohonen, 1984; Kohonen, 1993; Kohonen and Hari, 1999).

In the previous discussion the winner was defined in terms of inner products but it is possible to generalize the model to an arbitrary metric. If the usual Euclidean metric is used, the learning corresponds to minimization of a mean-squared vector quantization distortion or, equivalently, minimization of the distance to the closest cluster center in the K-means clustering. The domain of the input space that a neuron detects can hence be interpreted as a Voronoi region in vector quantization.

The relationship of competitive learning to our work is that the "cluster membership functions" $y_j(\mathbf{x};\boldsymbol{\theta}_j)$ in Section 2 may be interpreted as the outputs of a set of neurons, and in the limit of crisp membership functions (for Gaussians, $\sigma \to 0$) only one neuron, the one having the largest external input, is active and can be interpreted as the winner. After learning our algorithm therefore corresponds to a traditional competitive network. The learning procedure makes the difference by making the network detect features that are as homogeneous as possible with regard to the auxiliary data.

For normalized Gaussian membership functions $y_j(\mathbf{x};\boldsymbol{\theta}_j)$ and $\sigma \neq 0$ our model is a kind of a variant of Gaussian mixture models or soft vector quantization.

At the limit of crisp feature detectors (for Gaussians, $\sigma \to 0$), the output of the network reduces to a 1-of-C-coded discrete value. Similarly, the outputs of the soft version can be interpreted as probability density functions of a multinomial random variable. Such an interpretation has already been made by Becker in some of her work, discussed below.

2.2.2 Multinomial variables

Becker et al. (Becker and Hinton, 1992; Becker, 1996) have introduced a learning goal called *Imax* for neural networks. Their networks consist of two separate modules having different inputs, and the learning algorithms aim at maximizing the mutual information between the outputs of the modules. For example, if the inputs are two-dimensional arrays of random dots with stereoscopic displacements simulating the views of two eyes, the networks are able to infer depth from the data.

The variant called discrete Imax (Becker, 1996) is closely related to the clustering algorithm of the present paper. In Imax, the outputs of the neurons in each module are interpreted as the probabilities of a multinomial random variable, and the goal of learning is to maximize the mutual information between the variables of the two modules.

Our model differs from Becker's in two ways. First, Becker uses (normalized) $\exp(\mathbf{x}^T \boldsymbol{\theta}_j)$ as basis functions, whereas our parameterization makes the basis functions invariant to the norms of $\boldsymbol{\theta}_j$ (cf. eqn 11). Without such invariance, the units with largest norms may dominate the representation, a phenomenon noted by Becker as well.

The other difference is that Becker optimizes the model using gradient descent based on the whole batch of input vectors, whereas we have a simple Hebb-type on-line algorithm (eqns 8 and 9) adapting on the basis of one data sample at a time.

The gradient of the discrete Imax with respect to the parameters $\boldsymbol{\theta}_l$ is, after simplification and in our notation,

$$\frac{\partial I}{\partial \boldsymbol{\theta}_j} = \sum_{i,j} \int \mathbf{x} \log \frac{p(c_i|v_l)}{p(c_i|v_j)}\, p(v_j, v_l, c_i, \mathbf{x})\, d\mathbf{x}\ . \tag{17}$$

It would be possible to apply stochastic approximation here by sampling from $p(v_l, v_j, c_i, \mathbf{x})$, which leads to a different adaptation rule than ours.

Becker (1996) has also used Gaussian basis functions, but with some approximations and ending up having a different formula for the gradient.

2.2.3 Continuous variables

The mutual information between continuous-valued outputs of two neurons can be maximized as well (Becker and Hinton, 1992; Becker, 1996). Some assumptions about the continuously-valued signals and the noise has to be made, however. In (Becker and Hinton, 1992) the outputs were assumed to consist of Gaussian signals corrupted by independent, additive Gaussian noise.

In this paper the multinomial Imax has been reinterpreted as (soft) vector quantization in the Kullback-Leibler "metric" in which the distance is measured in an auxiliary space. In neural terms, the model builds a representation of the input space: each neuron is a detector specialized to represent a certain domain. In contrast, the continuous version tries to represent the input space by generating a parametric transformation to a continuous, lower-dimensional output space. If the parameterization and the assumptions about the noise are correct, the continuous representations are potentially more accurate. The advantage of the quantized representation is that no such assumptions need to be made; the model[1] is almost purely data-driven (semi-parametric), and of course very useful for clustering-type applications.

It is particularly difficult to maximize the mutual information if each module has several continuous-valued outputs. In some recent works (Fisher and Principe, 1998; Torkkola and Campbell, 2000) the Shannon entropy has been replaced by the quadratic Renyi entropy, yielding simpler formulae for the mutual information.

2.2.4 Information bottleneck

Tishby et al. (1999) get motivation to their *information bottleneck* method from the rate distortion theory of Shannon and Kolmogorov (see Cover and Thomas, 1991, for a review). In the rate distortion theory, the aim is to find an optimal codebook for a set of discrete symbols when a 'cost' in the form of a distortion function describing the effects of a transmission line is given.

[1]We have made some assumptions by parameterizing the $f(\mathbf{x}; \boldsymbol{\theta})$ in (4). However, the model becomes semi-parametric as a scale parameter, similar to the σ for Gaussians and $1/\kappa$ for the vMF kernels, approaches zero.

In our notation, the authors consider the problem of building an optimal representation V for a *discrete* random variable X. In the rate distortion theory, a real-valued distortion function $d(x, v)$ is assumed known, and $I(X; V)$ is minimized with respect to the representation (or conventionally, the codebook) $p(v|x)$ subject to the constraint $E_{X,V}\{d(x, v)\} < k$. At the minimum the conditional distributions defining the codebook are

$$p(v_l|x) = \frac{p(v_l)\exp\left[-\beta d(x, v_l)\right]}{\sum_j p(v_j)\exp\left[-\beta d(x, v_j)\right]} , \tag{18}$$

where β depends on k. The authors realized that if the average distortion $E_{X,V}\{d(x, v)\}$ is replaced by the mutual information $-I(C; V)$, then the rate distortion theory gives a solution that captures as much information of the "relevance variable" C as possible. Here the multinomial random variable C has the same role as our auxiliary data.

The functional to be minimized becomes $I(X; V) - \beta I(C; V)$, and its variational optimization with respect to the conditional densities $p(v|x)$ leads to the solution (18) with

$$d(x, v_j) = D_{KL}(p(c|x), p(c|v_j)) . \tag{19}$$

Together these two equations give a characterization of the optimal representation V when X is discrete and we accept the criterion $I(X; V) - \beta I(C; V)$ for the goodness of the representation. (Note that the characterization is self-referential through $p(c|v)$ and therefore does not in itself present an algorithm for finding the $p(c|v)$; Tishby et al. introduced an algorithm for finding the prototypes in the case of a multinomial X.)

The connection to our work is that, as shown in Section 2.1, minimizing the distortion E_{KL} is equivalent to maximizing the mutual information $I(C; V)$. If our X were discrete, using the distortion criterion would therefore be a limit case of Tishby's $I(X; V) - \beta I(C; V)$, with $\beta \to \infty$.

In our setup, X is continuous. It is possible, however, that under suitable conditions the variational approach of Tishby et al. could be generalized to the case of a continuous X. Then (18) and (19) would describe the optimal continuous basis functions. Of course, the result would not be of direct practical help unless $p(c|\mathbf{x})$ were known in advance.

The shape of the basis functions of such a theoretical continuous information bottleneck would be indirectly (and elegantly) limited by the term $I(X; V)$ of the cost function, which keeps the information the representation V offers of X low.

In our setup $p(c|\mathbf{x})$, needed for (18), must be estimated from the data. We have implicitly done this, and parameterized the conditional density in the form of the basis functions $y_j(\mathbf{x}; \boldsymbol{\theta}_j)$. The parameterization limits the shape of the basis functions even when the β of the bottleneck is infinite and the mutual information $I(C; V)$, which we maximized, dominates.

Nevertheless, eqn 18 offers fundamental insight to the problem and a starting point for further developments, like alternative second-order parameterizations.

3 Case study: Clustering of text documents

We tested our approach by clustering a large, high-dimensional data set, a collection of abstracts of scientific papers from the INSPEC database. Such clusterings are used in the information retrieval field to speed up document searches, and for document categorization.

Lists of keywords were used as the distributions of the auxiliary space that guides the clustering. The keywords have been manually chosen to reflect the contents of the abstracts, and they should therefore contain some information about what is most essential in the abstracts.

We compared our model with a totally unsupervised mixture of von Mises–Fisher distributions, fitted to the data with the EM algorithm. The model is analogous to the usual mixture of Gaussians; the Gaussian mixture components are simply replaced by von Mises–Fisher components. The model is

$$p(\mathbf{x}) = \sum_j p(\mathbf{x}|v_j)p_j \tag{20}$$

where $p(\mathbf{x}|v_j) = \text{vMF}(\mathbf{x}; \boldsymbol{\theta}_j)$, and vMF is defined in (11). The p_j are the mixing parameters.

3.1 Preprocessing of data

The data was collected from nine general, partially overlapping classes: nuclear physics; optics; materials science; circuit theory and circuits; information and communication theory; power systems and applications; systems theory and cybernetics; software techniques and systems; and engineering computing.

We carried out two sets of experiments, each with different preprocessing methods. In both of them we constructed the input vector using the words in the abstracts and titles, converted into their base form.

In the first experiment no prior information about the relevance of the input variables, namely the words, was utilized. We picked randomly 500 words from the set of words that occurred over 50 times in the whole collection. The components of each of the 500-dimensional input vectors were set to the frequencies of occurrence of these words in the corresponding document, and finally the length of the input vectors was normalized. The encoding corresponds to the "vector space model" (Salton and McGill, 1983) of information retrieval, with term frequency weighting.

In the second experiment we encoded the abstracts with a method that utilizes more prior information about which kind of encoding is good in information retrieval. We removed words belonging to a stop-list of 1335 uninformative words, and used all the remaining 4748 words that occurred over 50 times in the documents. The frequency of occurrence of the words was weighted with the well-known IDF-weight, which is the inverse of the number of the documents where the term occurs.

In both experiments we removed documents in which less than 5 words remained after the preprocessing, resulting in 41,307 documents in the first experiment and 54,555 in the second.

3.2 Results

In both models, our own and the vMF mixture (20), we first estimated the width of the basis functions, parameter κ in (11), using a validation set. We then compared our approach with the mixture of vMF distributions by computing how much information the discovered (soft) categories convey of the nine *document classes*, selected independently by informaticians and meant to as well as possible describe the fields into which the abstracts belong. Note that the classes were not utilized in teaching of either model. We tested how much the extra auxiliary information, the keywords that are known to be relevant, helps in automatic construction of meaningful categories.

We computed the mutual information between the class labels and (the posterior probabilities of) the learned categories. In the KL clustering model the posterior probabilities $p(v_j|\mathbf{x}) = \text{vMF}(\mathbf{x}; \boldsymbol{\theta}_j)$ are given by (11), and in the mixture model by $p(v_j|\mathbf{x}) = p(\mathbf{x}|v_j)p_j/p(\mathbf{x})$. A test set used neither in learning nor in validation (selection of κ) was used for estimating the mutual information.

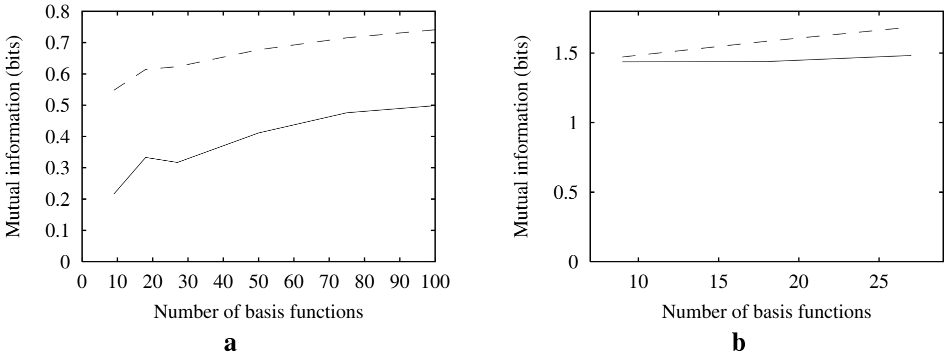

Figure 2: Mutual information between the posterior probabilities of the document categories generated by the models and an independent classification of the documents (*not* used in learning), as a function of the number of categories. **a** First experiment. Prior knowledge was not utilized in feature extraction. **b** Second experiment. Prior knowledge was utilized. Solid line: mixture of vMFs, dashed line: our model.

The results of the first experiment are shown in Fig. 2**a** and of the second experiment in Fig. 2**b**, respectively. In the first experiment, in which no prior knowledge was utilized in feature extraction, our method clearly outperformed the mixture model. In the second experiment with the improved feature extraction method the margin clearly reduced. The probable reason is that the nine classes, with abstracts collected from very different fields, corresponded to nine relatively well separated clusters. Even the basic mixture model was then able to find the clusters easily.

4 Conclusions

In this paper we have derived and analyzed a simple on-line algorithm that utilizes auxiliary information in forming soft data categories. The categories minimize the within-cluster distortion in an auxiliary space, while still being defined and local in terms of the primary data space.

Our experimental results suggest that the algorithm is particularly useful for categorization when good feature extraction stages are not available but there exists auxiliary information that reflects relevance to the goal of the processing task.

Appendix 1: Gradient of the error criterion with respect to the parameters of the basis functions

If we denote

$$K_{il}(\mathbf{x}) \equiv \sum_j \frac{\partial y_j(\mathbf{x};\boldsymbol{\theta}_j)}{\partial \boldsymbol{\theta}_l} \log \psi_{ji} \ , \tag{21}$$

then

$$\frac{\partial E_{KL}}{\partial \boldsymbol{\theta}_l} = -\sum_i \int K_{il}(\mathbf{x}) \, p(c_i, \mathbf{x}) \, d\mathbf{x} \ . \tag{22}$$

Since the basis functions $y_j(x;\boldsymbol{\theta}_j)$ are of the normalized exponential form (4) the gradient of y_j is[2]

$$\frac{\partial y_j(\mathbf{x};\boldsymbol{\theta}_j)}{\partial \boldsymbol{\theta}_l} = \delta_{lj} y_j(\mathbf{x};\boldsymbol{\theta}_j) \frac{\partial f_j(\mathbf{x};\boldsymbol{\theta}_j)}{\partial \boldsymbol{\theta}_l} - y_j(\mathbf{x};\boldsymbol{\theta}_j) y_l(\mathbf{x};\boldsymbol{\theta}_l) \frac{\partial f_l(\mathbf{x};\boldsymbol{\theta}_l)}{\partial \boldsymbol{\theta}_l} \ . \tag{23}$$

Substituting the result into (21) gives

$$\begin{aligned} K_{il}(\mathbf{x}) &= y_l(\mathbf{x};\boldsymbol{\theta}_l) \left[\log \psi_{li} - \sum_j y_j(\mathbf{x};\boldsymbol{\theta}_j) \log \psi_{ji} \right] \frac{\partial f_l(\mathbf{x};\boldsymbol{\theta}_l)}{\partial \boldsymbol{\theta}_l} \\ &= y_l(\mathbf{x};\boldsymbol{\theta}_l) \left[\sum_j y_j(\mathbf{x};\boldsymbol{\theta}_j) \log \frac{\psi_{li}}{\psi_{ji}} \right] \frac{\partial f_l(\mathbf{x};\boldsymbol{\theta}_l)}{\partial \boldsymbol{\theta}_l} \ . \end{aligned} \tag{24}$$

Since the V and C are conditionally independent with respect to X, we may write

$$y_j(\mathbf{x};\boldsymbol{\theta}_j) y_l(\mathbf{x};\boldsymbol{\theta}_l) p(c_i,\mathbf{x}) = p(c_i, v_j, v_l, \mathbf{x}) \ . \tag{25}$$

Substituting this and eqn (24) into (22) we arrive at (5).

Appendix 2: Optimal prototypes

Based on (3) and the conditional independence of C and V with respect to X,

$$\begin{aligned} \text{const.} - E_{KL} &= \int \sum_{i,j} y_j(\mathbf{x};\boldsymbol{\theta}_l) \log \psi_{ji} \ p(c_i,\mathbf{x}) \ d\mathbf{x} && (26) \\ &= \sum_{i,j} \log \psi_{ji} \int p(c_i, v_j, \mathbf{x}) \ d\mathbf{x} && (27) \\ &= \sum_{i,j} p(c_i, v_j) \log \psi_{ji} = \sum_j p(v_j) \sum_i p(c_i|v_j) \log \psi_{ji} \ . && (28) \end{aligned}$$

Note that both $\{p(c_i|v_j)\}_i$ and $\{\psi_{ji}\}_i$ are probability densities. It is well known in the information theory literature (information inequality; e.g. Cover and Thomas, 1991) that for such densities the expression

$$\sum_i p(c_i|v_j) \log \psi_{ji}$$

is maximized when $\psi_{ji} = p(c_i|v_j)$. At the minimum of E_{KL} we therefore necessarily have $\psi_{ji} = p(c_i|v_j)$.

[2] Note that the normalized versions of the densities of the so called exponential family are included—if the $y_j(\mathbf{x};\boldsymbol{\theta}_j)$ are interpreted as densities of the random variables $p(v_j|\mathbf{x})$.

Appendix 3: Connection to conditional density estimation

Let us denote $y_j(\mathbf{x};\boldsymbol{\theta}_j) = y_j(\mathbf{x})$ for brevity. The distortion (3) can be expressed as

$$E_{KL} = -\sum_{i,j} \int y_j(\mathbf{x}) \log \psi_{ji}\ p(c_i, \mathbf{x})\ d\mathbf{x} + \text{const.} \tag{29}$$

$$= \int \sum_i \left[\log p(c_i|\mathbf{x}) - \sum_j y_j(\mathbf{x}) \log \psi_{ji} \right] p(c_i, \mathbf{x})\ d\mathbf{x} + \text{const.} \tag{30}$$

$$= \int \sum_i p(c_i|\mathbf{x}) \log \frac{p(c_i|\mathbf{x})}{\exp \sum_j y_j(\mathbf{x}) \log \psi_{ji}}\ p(\mathbf{x})\ d\mathbf{x} + \text{const.} \tag{31}$$

$$= \int \sum_i p(c_i|\mathbf{x}) \log \frac{p(c_i|\mathbf{x})}{q_i(\mathbf{x})}\ p(\mathbf{x})\ d\mathbf{x} + \text{const.}\ , \tag{32}$$

$$\tag{33}$$

where

$$q_i(\mathbf{x}) \equiv \exp \sum_j y_j(\mathbf{x}) \log p(c_i|v_j)\ . \tag{34}$$

The $\{q_i(\mathbf{x})\}_i$ is not a proper density. However, $\sum_i q_i(\mathbf{x}) \leq 1$ based on Jensen's inequality. Hence, for all $\mathbf{x}$,

$$\sum_i p(c_i|\mathbf{x}) \log \frac{p(c_i|\mathbf{x})}{q_i(\mathbf{x})}\ p(\mathbf{x}) \geq \sum_i p(c_i|\mathbf{x}) \log \frac{p(c_i|\mathbf{x})}{\hat{p}(c_i|\mathbf{x})}\ p(\mathbf{x})\ , \tag{35}$$

where

$$\hat{p}(c_i|\mathbf{x}) \equiv \frac{q_i(\mathbf{x})}{\sum_i q_i(\mathbf{x})}\ . \tag{36}$$

Therefore, minimizing the clustering criterion E_{KL} minimizes an upper limit of

$$\int \sum_i p(c_i|\mathbf{x}) \log \frac{p(c_i|\mathbf{x})}{\hat{p}(c_i|\mathbf{x})}\ p(\mathbf{x})\ d\mathbf{x} = E_X\{D_{KL}(p(c|\mathbf{x}), \hat{p}(c|\mathbf{x}))\}\ , \tag{37}$$

the expected KL-divergence between the conditional density $p(c|\mathbf{x})$ and its estimate $\hat{p}(c|\mathbf{x})$. (Here $E_X\{\cdot\}$ denotes the expectation over $\mathbf{x}$.) One can also write

$$\begin{aligned} E_X\{D_{KL}(p(c|\mathbf{x}), \hat{p}(c|\mathbf{x}))\} &= \int \sum_i p(c_i, \mathbf{x}) \log \frac{p(c_i, \mathbf{x})}{\hat{p}(c|\mathbf{x}) p(\mathbf{x})} d\mathbf{x} \\ &= D_{KL}(p(c, \mathbf{x}), \hat{p}(c|\mathbf{x}) p(\mathbf{x}))\ . \end{aligned} \tag{38}$$

Maximizing the likelihood of the model $\hat{p}(c_k|\mathbf{x}_k)$ for data $\{c_k, \mathbf{x}_k\}_k$ sampled from $p(c, \mathbf{x})$ is asymptotically equivalent to minimizing the Kullback-Leibler divergence (38). This is because for the N i.i.d. samples $\{c_k, \mathbf{x}_k\}_k$ the scaled log-likelihood

$$\frac{1}{N} \sum_k \log \hat{p}(c_k|\mathbf{x}_k)$$

converges to

$$\begin{aligned} \int \sum_i p(c_i, \mathbf{x}) \log \hat{p}(c_i|\mathbf{x}) d\mathbf{x} &= -\int \sum_i p(c_i|\mathbf{x}) \log \frac{p(c_i|\mathbf{x})}{\hat{p}(c_i|\mathbf{x})} p(\mathbf{x}) d\mathbf{x} + \text{const.} \\ &= -E_X\{D_{KL}(p(c|\mathbf{x}), \hat{p}(c|\mathbf{x}))\} + \text{const.} \end{aligned} \tag{39}$$

with probability 1 when $N \to \infty$, and because maximizing (39) minimizes (38).

In the special case when the $y_j(\mathbf{x})$ are binary-valued, the $q_i(\mathbf{x}) = \hat{p}(c_i|\mathbf{x})$ is a proper density and the equality in (35) holds for almost every $\mathbf{x}$. Therefore, the minimization of (an approximation of) the average distortion E_{KL} on a finite data set is equivalent to maximizing the likelihood of $\hat{p}(c|\mathbf{x})$ on the same sample.

Acknowledgments

This work was supported by the Academy of Finland.

References

Becker, S. (1996). Mutual information maximization: models of cortical self-organization. *Network: Computation in Neural Systems, 7*, 7–31.

Becker, S. & Hinton, G. E. (1992). Self-organizing neural network that discovers surfaces in random-dot stereograms. *Nature, 355*, 161–163.

Cover, T. M. & Thomas, J. A. (1991). *Elements of Information Theory*. New York: Wiley.

Didday, R. L. (1976). A model of visuomotor mechanisms in the frog optic tectum. *Mathematical Biosciences, 30*, 169–180.

Fisher III, J. W., & Principe, J. (1998). A methodology for information theoretic feature extraction. In *Proc. IJCNN'98, International Joint Conference on Neural Networks* (pp. 1712–1716). Piscataway, NJ: IEEE Service Center.

Forgy, E. W. (1965). Cluster analysis of multivariate data: Efficiency vs interpretability of classifications. *Biometrics, 21*, 768–769.

Gersho, A. (1979). Asymptotically optimal block quantization. *IEEE Transactions on Information Theory, 25*, 373–380.

Gray, R. M. (1984). Vector quantization. *IEEE ASSP Magazine*, April, 4–29.

Grossberg, S. (1976). On the development of feature detectors in the visual cortex with applications to learning and reaction-diffusion systems. *Biological Cybernetics, 21*, 145–159.

Hofmann, T. (2000). Learning the similarity of documents: An information-geometric approach to document retrieval and categorization. In Solla, S. A., Leen, T. K., & Müller, K.-R. (Eds.), *Advances in Neural Information Processing Systems, 12* (pp. 914–920). Cambridge, MA: MIT Press.

Jaakkola, T. S. & Haussler, D. (1999). Exploiting generative models in discriminative classifiers. In Kearns, M. S., Solla, S. A., and Cohn, D. A. (Eds.), *Advances in Neural Information Processing Systems, 11* (pp. 487–493). San Mateo, CA: Morgan Kaufmann.

Kaski, S. & Kohonen, T. (1994). Winner-take-all networks for physiological models of competitive learning. *Neural Networks, 7*, 973–984.

Kaski, S. & Sinkkonen, J. (2000). Metrics that learn relevance. In *Proc. IJCNN-2000, International Joint Conference on Neural Networks*. To appear.

Kohonen, T. (1984). *Self-Organization and Associative Memory* (3rd ed. 1989). Berlin: Springer-Verlag.

Kohonen, T. (1993). Physiological interpretation of the self-organizing map algorithm. *Neural Networks, 6*, 895–905.

Kohonen, T. & Hari, R. (1999). Where the abstract feature maps of the brain might come from. *TINS, 22*, 135–139.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In Le Cam, L. M., & Neyman, J. (Eds.), *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. Vol. I: Statistics* (pp. 281–297). Berkeley, CA: University of California Press.

Makhoul, J., Roucos, S., & Gish, H. (1985). Vector quantization in speech coding. *Proceedings of the IEEE, 73*, 1551–1588.

Mardia, K. V. (1975). Statistics of directional data. *Journal of the Royal Statistical Society B 37*, 349–393.

Nass, M. M. & Cooper, L. N. (1975). A theory for the development of feature detecting cells in visual cortex. *Biological Cybernetics, 19*, 1–18.

Nowlan, S. J. (1990). Maximum likelihood competitive learning. In Touretzky, D. S. (Ed.), *Advances in Neural Information Processing Systems, 2* (pp. 574–582). San Mateo, CA: Morgan Kaufmann.

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

Pérez, R., Glass, L., & Shlaer, R. J. (1975). Development of specificity in cat visual cortex. *Journal of Mathematical Biology, 1*, 275–288.

Salton, G. & McGill, M. J. (1983). *Introduction to modern information retrieval.* McGraw-Hill, New York.

Tipping, M. E. (1999). Deriving cluster analytic distance functions from Gaussian mixture models. In *Proc. ICANN99, Ninth International Conference on Artificial Neural Networks* (pp. 815–820). London: IEE.

Tishby, N., Pereira, F. C., & Bialek, W. (1999). The information bottleneck method. In *37th Annual Allerton Conference on Communication, Control, and Computing*, Urbana, Illinois.

Torkkola, K. & Campbell, W. M. (2000). Mutual information in learning feature transformations. In *Proc. ICML'2000, The Seventeenth International Conference on Machine Learning*. To appear.