Generalized clustering via kernel embeddings

Stefanie Jegelka[1], Arthur Gretton[2,1], Bernhard Schölkopf[1], Bharath K. Sriperumbudur[3], and Ulrike von Luxburg[1]

[1] Max Planck Institute for Biological Cybernetics, Tübingen, Germany
[2] Carnegie Mellon University, Pittsburgh, PA 15213, USA
[3] Dept. of ECE, UC San Diego, La Jolla, CA 92093, USA

Abstract. We generalize traditional goals of clustering towards distinguishing components in a non-parametric mixture model. The clusters are not necessarily based on point locations, but on higher order criteria. This framework can be implemented by embedding probability distributions in a Hilbert space. The corresponding clustering objective is very general and relates to a range of common clustering concepts.

1 Introduction

In this paper we consider a statistical, non-parametric framework for clustering. Assuming the data points to be drawn from some underlying distribution P, we treat a cluster as a sample from a component distribution P_k. A clustering can then be described as a decomposition of the underlying distribution of the form $P = \sum_{k=1}^{K} \pi_k P_k$ with mixture weights π_k.

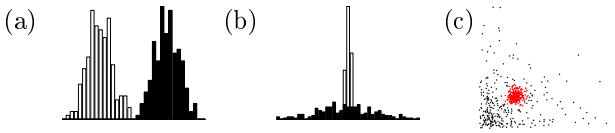

Fig. 1. (a) Traditional clustering, (b) generalized clustering; (c) example 2D result labeled by the MMD approach.

A common statistic to quantify the separation between clusters is the distance between the cluster means (e.g. Figure 1.a). However, separation based on "location" is not always what we want to achieve. The example in Figure 1.b is a mixture of two Gaussians with identical means, but with different variances. In this situation a decomposition is desirable where the difference between the variances of P_1 and P_2 is large. The difference between cluster means or between cluster variances are just two examples of distance functions between distributions. A straightforward generalization of the traditional clustering problem is to replace the distance between the means by a more general distance function. To avoid unnecessarily complicated solutions, we additionally require that the components P_k be "simple". This leads to the following generalized clustering problem:

> *Generalized clustering:* Decompose the density into *"simple"* components P_i, while *maximizing a given distance* function between the P_i.

A particularly suitable distribution representation and associated distance measure is given by Gretton et al. (2006). In this framework, a probability distribution P is embedded as $\mu[P]$ into a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ corresponding to some kernel function k. The Hilbert space norm $\|\mu[P]\|_{\mathcal{H}}$ can be interpreted as a "simplicity score": the smaller the norm, the "simpler" the corresponding distribution (e.g., the smoother the density). The maximum mean discrepancy $\mathrm{MMD}(P, Q)$ between two distributions P and Q is defined as the Hilbert space distance $\|\mu[P] - \mu[Q]\|_{\mathcal{H}}$ between the two embedded distributions. MMD with a dth-order polynomial kernel only discriminates between distributions based on the first d moments; with a linear kernel, it is simply the distance of means. At the most general level, i.e., with a characteristic kernel (e.g., a Gaussian kernel), *all* moments are accounted for (Sriperumbudur et al., 2008). Thus, the combination of simplicity score and distance between distributions afforded by RKHS embeddings yields a straightforward expression for the two objectives of generalized clustering. Our formulation will turn out to be very generic and to relate to many well-known clustering criteria. In this sense, the main contribution of this work is to reveal and understand the properties of the MMD approach and its relations to existing clustering algorithms. We will discuss them in Section 4 and simplicity in Section 5, after formally introducing MMD in Section 2 and the optimization problem in Section 3.

Alternative approaches to generalized clustering exist in the literature, but they are less general than the MMD approach. We summarize them in Section 4.

2 Maximum mean discrepancy (MMD)

We begin with a concise presentation of kernel distribution embeddings and the MMD, following Gretton et al. (2006). Given a kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ on some space $\mathcal{X}$, it is well known that points $x \in \mathcal{X}$ can be embedded into the corresponding reproducing kernel Hilbert space $\mathcal{H}$ via the embedding $\varphi : \mathcal{X} \to \mathcal{H},\ x \mapsto k(x, \cdot)$. If P denotes a probability distribution on $\mathcal{X}$, one can show that the expectation $\mu[P] := \mathbb{E}_{x\sim P}[\varphi(x)]$ realizes an embedding[4] $\mu : \mathcal{P} \to \mathcal{H}$ of the space of all probability distributions $\mathcal{P}$ in $\mathcal{H}$. The Maximum Mean Discrepancy between two distributions P_1 and P_2 can be defined in two equivalent ways:

$$\mathrm{MMD}(P_1, P_2) = \|\mu[P_1] - \mu[P_2]\|_{\mathcal{H}} \tag{1}$$

$$= \sup_{g\in\mathcal{H}, \|g\|_{\mathcal{H}}\leq 1} \left(\mathbb{E}_{x\sim P_1} g(x) - \mathbb{E}_{y\sim P_2} g(y)\right). \tag{2}$$

The first form shows that MMD is a metric. The second form shows that two probability distributions P_1 and P_2 are particularly "far" from each other if there exists a smooth function g that has largest magnitude where the probability mass of P_1 differs most from that of P_2. Given samples $\{X_i^{(1)}\}_{i=1}^n$ and $\{X_i^{(2)}\}_{i=1}^n$, the embedding and the MMD can be empirically estimated as

[4] Assume $\mathbb{E}_P[k(x, x)] < \infty$ and k is *characteristic*, then the embedding is injective (see Sriperumbudur et al. (2008) for the proof and the definition of 'characteristic').

$$\mu[\hat{P}_1] = \frac{1}{n}\sum_{i=1}^{n}\varphi(X_i) = \frac{1}{n}\sum_{i=1}^{n}k(X_i,\cdot) \tag{3}$$
$$\widehat{\mathrm{MMD}}(P_1,P_2) = \mathrm{MMD}(\hat{P}_1,\hat{P}_2) := \|\mu[\hat{P}_1] - \mu[\hat{P}_2]\|_{\mathcal{H}}.$$

3 The generalized clustering optimization problem

We now describe how kernel distribution embeddings can be used to implement a generalized clustering algorithm. For simplicity, we focus on the case of two clusters only. Our goal is to decompose the underlying distribution P such that $\mathrm{MMD}(P_1,P_2)$ is large and $\|\mu[P_1]\|_{\mathcal{H}}$ and $\|\mu[P_2]\|_{\mathcal{H}}$ are small. With only a finite sample $\{X_i\}_{i=1}^n$, we must estimate these quantities empirically. To this end, we parameterize the empirical distributions $\hat{P}_k$ via assignments $\alpha_i^{(k)}$ of Dirac measures δ_{X_i} on the sample points X_i:

$$\hat{\pi}_k\hat{P}_k = \frac{1}{n}\sum_{i=1}^{n}\alpha_i^{(k)}\delta_{X_i} \quad \text{with} \quad \hat{\pi}_k = \frac{1}{n}\sum_{i=1}^{n}\alpha_i^{(k)}$$

for $\alpha_i^{(1)} + \alpha_i^{(2)} = 1$. For a soft clustering we allow $\alpha^{(1)},\alpha^{(2)} \in [0,1]^n$; for a hard clustering we constrain $\alpha^{(1)},\alpha^{(2)} \in \{0,1\}^n$. The resulting decomposition takes the form $\hat{P} = \hat{\pi}_1\hat{P}_1 + \hat{\pi}_2\hat{P}_2$. These estimates lead to the following optimization problem (note that $\alpha^{(2)} = \mathbf{1} - \alpha^{(1)}$ is determined by $\alpha^{(1)}$):

$$\max_{\alpha^{(1)}}\Psi(\alpha^{(1)}) := \max_{\alpha^{(1)}} \quad \text{MMD-Term}(\hat{P}_1,\hat{P}_2) + \lambda\cdot\text{Regularization-Term}(\hat{P}_1,\hat{P}_2).$$

Let $\mathbf{K} = (k(X_i,X_j))_{i,j=1,\ldots,n}$ denote the kernel matrix of the sample. The MMD-Term and its parametric form are then

$$\begin{aligned}\hat{\pi}_1\hat{\pi}_2\mathrm{MMD}(\hat{P}_1,\hat{P}_2) &= \hat{\pi}_1\hat{\pi}_2\|\mu[\hat{P}_1] - \mu[\hat{P}_2]\|_{\mathcal{H}}^2 \\ &= \frac{\hat{\pi}_2}{n^2\hat{\pi}_1}(\alpha^{(1)})^\top\mathbf{K}\alpha^{(1)} + \frac{\hat{\pi}_1}{n^2\hat{\pi}_2}(\alpha^{(2)})^\top\mathbf{K}\alpha^{(2)} - \frac{2}{n^2}(\alpha^{(1)})^\top\mathbf{K}\alpha^{(2)}.\end{aligned} \tag{4}$$

The product of the cluster sizes $\hat{\pi}_1,\hat{\pi}_2$ acts as a balancing term to avoid particularly small clusters. We will call the maximization of (4) *maxMMD* and give various interpretations in Section 4. As a regularization term we use

$$\lambda_1\|\mu[\hat{P}_1]\|_{\mathcal{H}}^2 + \lambda_2\|\mu[\hat{P}_2]\|_{\mathcal{H}}^2 = \frac{\lambda_1}{n^2\hat{\pi}_1^2}(\alpha^{(1)})^\top\mathbf{K}\alpha^{(1)} + \frac{\lambda_2}{n^2\hat{\pi}_2^2}(\alpha^{(2)})^\top\mathbf{K}\alpha^{(2)} \tag{5}$$

To avoid empty clusters, we introduce a constraint for the minimum size $\varepsilon > 0$ of a cluster. This leads to the final optimization problem

$$\max_{\alpha^{(1)}\in[0,1]^n}\Psi(\alpha^{(1)}) \qquad \text{s. t. } \varepsilon \le \sum_i \alpha_i^{(1)} \le (1-\varepsilon).$$

As we shall see in Section 4, maxMMD alone can be optimized efficiently via a variant of the kernel k-means algorithm ensuring the minimum size constraint. For the full objective, we used the Ipopt solver (Wächter and Biegler, 2006). Even though we evaluated our criterion and variants in many experiments, we will exclude them to save space and concentrate on the theory, our main contribution.

4 MaxMMD, discriminability, and related approaches

We will now describe how the *discriminability* criterion maxMMD (Eqn. (4)) encompasses the concepts behind a number of classical clustering objectives. Figure 2 gives an overview.

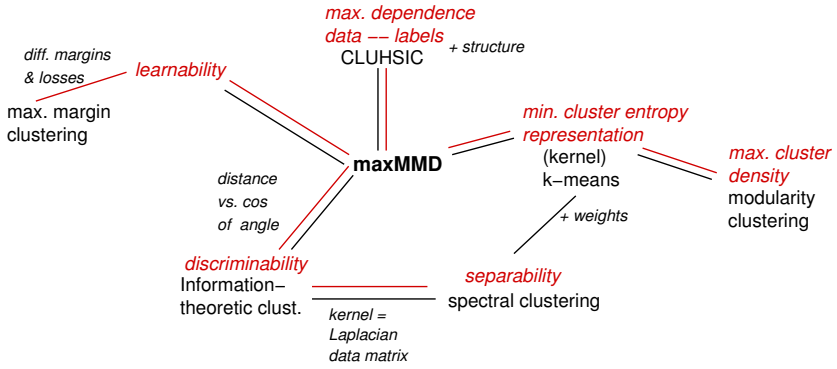

Fig. 2. Overview of connections. Red links are conceptual, black links mathematical. First, *discriminability* is conceptually close to *learnability*. Learning seeks to detect patterns in the data, i.e., dependencies between data and labels (cluster assignments). Only if we capture those dependencies can we reliably predict labels. In other words, we want clusters to maximize the *dependence between data and labels*. If the label distribution is closely linked to the data, it *represents* the data well. Representation conceptually connects to compression and coding. Clusters that are less complex to describe have *lower entropy*. Small entropy means *dense clusters*, which leads back to a generalization of the k-means criterion. Extended by weights, this criterion encompasses spectral clustering, e.g., graph cut criteria. Those cuts favor sparse connections between clusters, simply another measure of discriminability. Spectral clustering also relates to discriminability via the angle of embeddings $\mu[P_1], \mu[P_2]$.

Concept 1 (Discriminability) *MaxMMD seeks dissimilar components P_k.*

Since MMD is a metric for distributions, maxMMD strives for distant component distributions by construction. Hence, it indirectly *promotes discriminability* of the clusters by their statistics.

Moreover, MaxMMD behaves similarly to the Jensen-Shannon (JS) divergence in the clustering context. For a mixture $P = \pi_1 P_1 + \pi_2 P_2$, the latter is $D_{\mathrm{JS}}(P_1, P_2) = \pi_1 D_{\mathrm{KL}}(P_1, P) + \pi_2 D_{\mathrm{KL}}(P_2, P)$ (Fuglede and Topsøe, 2004). If we replace the KL divergence D_{KL} by the squared MMD, we arrive at the parallel form

$$\pi_1 \mathrm{MMD}(P_1, P)^2 + \pi_2 \mathrm{MMD}(P_2, P)^2 = \pi_1 \pi_2 \mathrm{MMD}(P_1, P_2)^2. \tag{6}$$

Discriminability via projections or some moments is used by, e.g., Chaudhuri and Rao (2008), Arora and Kannan (2001), but for specific distributions. Information-theoretic clustering (Gokcay and Principe, 2002, Jenssen et al., 2004) measures discriminability by the cosine of the angle between the $\mu[P_k]$. Its motivation, however, restricts k to have a specific form, whereas MMD is more general.

Concept 2 (Learnability) *MaxMMD finds clusters that are well learnable.*

It turns out that our objective also connects unsupervised and supervised learning. In supervised learning, the Bayes risk R^* measures the difficulty of a learning problem. If R^* is large, then there is no good rule (among the simple ones we choose from) to tell the classes apart, i.e., they are almost indiscriminable. We will see that the negative MMD corresponds to a particular Bayes risk.
Assume for a moment that the cluster assignments are known, i.e., the data points X_i have labels $Y_i \in \{-1, 1\}$, all sampled from an unknown joint distribution $P(X, Y)$. We search for a classifier $g : \mathcal{X} \to \{-1, 1\}$ from a space $\mathcal{G}$ of candidate functions. Here, $\mathcal{G}$ is the set of all functions in $\mathcal{H}$ with $\|g\|_{\mathcal{H}} \leq 1$. Let $P(X) = \pi_1 P_1 + \pi_2 P_2$ with $\pi_1 = P(Y = 1)$ and $\pi_2 = P(Y = -1)$. Choosing loss

$$\ell(y, g(x)) = \begin{cases} \frac{-g(x)}{\pi_1} & \text{if } y = 1 \\ \frac{g(x)}{\pi_2} & \text{if } y = -1 \end{cases}$$

and using the definition (1) of MMD, the Bayes risk becomes

$$\begin{aligned} R^* &= \inf_{g \in \mathcal{G}} - \left(\int g dP_1 - \int g dP_2 \right) = - \sup_{g \in \mathcal{G}} \left(\int g dP_1 - \int g dP_2 \right) \\ &= -\mathrm{MMD}(P_1, P_2). \end{aligned} \tag{7}$$

A large MMD hence corresponds to a low Bayes risk. The classifier g^* that minimizes the risk is (Gretton et al., 2008)

$$g^* = \arg \inf_{g \in \mathcal{H}, \|g\|_{\mathcal{H}} \leq 1} - \left(\int g dP_1 - \int g dP_2 \right) = \frac{\mu[P_1] - \mu[P_2]}{\|\mu[P_1] - \mu[P_2]\|_{\mathcal{H}}}, \quad \text{i.e.,}$$
$$g^*(x) = \langle g^*, \varphi(x) \rangle_{\mathcal{H}} \propto \int k(x, x') dP_1(x') - \int k(x, x'') dP_2(x'').$$

Estimating $\mu[P_k]$ as in (3) with the assignments $\alpha_i^{(k)}$ yields a Parzen window classifier with the window function k, assuming k is chosen appropriately.
In clustering, we choose labels Y_i and hence implicitly P_1 and P_2. Maximizing their MMD defines classes that are well learnable with $\mathcal{G}$. This concept is reminiscent of Maximum Margin Clustering (MMC) that seeks to minimize a 0-1-loss (Xu et al., 2005). As opposed to MMC, however, we do not maximize a minimum separation (margin) between the cluster points in the RKHS, but strive for discrepancy of the means in $\mathcal{H}$.

Concept 3 (Dependence maximization) *MaxMMD seeks the dependence between data and cluster assignments.*

What is the aim of learning? We assume an association between the data and the labels that can be expressed by a pattern or learnable rule. Well-learnable cluster assignments are in conformity with such a dependence.
In matrix notation and with binary assignments, Criterion (4) becomes

$$\hat{\pi}_1 \hat{\pi}_2 \|\mu[\hat{P}_1] - \mu[\hat{P}_2]\|_{\mathcal{H}}^2 = \mathrm{tr}(\mathbf{KL}) - const = -\mathrm{HSIC}(P, \alpha^{(1)}) - const \tag{8}$$

where $\mathbf{K}$ is the kernel matrix of the data and $\mathbf{L}$ that of the labels with entries $l_{ij} = n^{-1}\hat{\pi}_1^{-1/2}$ if $\alpha_i^{(1)} = \alpha_j^{(1)} = 1$, $l_{ij} = n^{-1}\hat{\pi}_2^{-1/2}$ if $\alpha_i^{(1)} = \alpha_j^{(1)} = 0$ and $l_{ij} = 0$ otherwise. HSIC (Gretton et al., 2005) is a measure of statistical dependence between random variables. Hence, *maxMMD is equivalent to maximizing the dependence between cluster assignments and the data distribution.*
This criterion has been used in existing clustering approaches. Song et al. (2007) exploit the HSIC formulation to introduce structural constraints. The criterion by Aghagolzadeh et al. (2007) is similar to (8) but derived as an estimate of the Mutual Information between labels and data.

Concept 4 (Representation) *MaxMMD aims to find functions $\mu[\hat{P}_k]$ that represent the data well.*

We can rewrite the maxMMD criterion as

$$2\hat{\pi}_1\hat{\pi}_2\|\mu[\hat{P}_1] - \mu[\hat{P}_2]\|_{\mathcal{H}}^2 = const - \sum_{k=1}^{2}\sum_{i=1}^{n}\alpha_i^{(k)}\|\varphi_i - \mu[\hat{P}_k]\|_{\mathcal{H}}^2. \tag{9}$$

Consider a probabilistic encoding with a limited number of deterministic codebook vectors $y(l) = \mu[\hat{P}_l]$. Choose a label $l(X_i)$ for a point X_i with probability $\hat{P}(l|X_i) = \alpha_i^{(l)}$ and encode it as $y(X_i) = y(l(X_i))$ (Rose, 1994). Then Criterion (9) is an estimate of the average distortion $\mathbb{E}[D] = \int d(X, y(X))dP(X,k)$ with divergence $d(X_i, y(X_i)) = \|\varphi_i - y(X_i)\|_{\mathcal{H}}^2$. An assignment with minimal distortion, as favored by (9), is a good encoding, i.e., *it represents the data well.* Following further algebra, the minimization of the average distortion corresponds to the maximization of the sum of dot products

$$\sum_{k=1}^{2}\sum_{i=1}^{n}\alpha_i^{(k)}\langle\varphi(X_i), \mu[\hat{P}_k]\rangle_{\mathcal{H}} = \sum_{k=1}^{2}\sum_{i=1}^{n}\alpha_i^{(k)} f_{\mu[\hat{P}_k]}(X_i). \tag{10}$$

The functions $f_{\mu[\hat{P}_k]}$ have the form of a Parzen window estimator: $f_{\mu[\hat{P}_k]}(X) = \sum_{j=1}^{n}\alpha_j^{(k)}k(X, X_j)$. Criterion (10) is large if the $f_{\mu[\hat{P}_k]}$ represent the density structure of the data, i.e., if each X_i is likely under the estimator it belongs to.

Concept 5 (Entropy) *MaxMMD finds clusters with low generalized entropy.*

With the definition $n\mu[\hat{P}_k] = \hat{\pi}_k^{-1}\sum_i \alpha_i^{(k)}\varphi(X_i)$, Criterion (10) becomes

$$\max \sum_{k=1}^{2}\hat{\pi}_k\|\mu[\hat{P}_k]\|_{\mathcal{H}}^2. \tag{11}$$

The related term $H_2(P_k) = -\log\|\mu[P_k]\|_{\mathcal{H}}^2$, as shown by Erdogmus and Principe (2006), is the Parzen window estimate of a generalized entropy, the Renyi entropy (Renyi, 1960). Consequently, large norms $\|\mu[\hat{P}_k]\|_{\mathcal{H}}$ result in small Renyi entropies of the $\hat{P}_k$. Thus, similar to the analogy in Equation (6), Criterion (11) parallels the JS divergence: maximizing D_{JS} minimizes the weighted sum of Shannon entropies, $D_{\mathrm{JS}}(P_1, P_2) = H_S(P) - \pi_1 H_S(P_1) - \pi_2 H_S(P_2)$ (Lin, 1991).

Criterion (9) is in fact the kernel k-means objective. While linear k-means seeks clusters with minimal variance, its kernel version generalizes to clusters with minimal entropy, as shown by Concept 5, and in line with the viewpoint of Erdogmus and Principe (2006) that entropy is a generalization of variance to non-Gaussian settings.
Entropy can be viewed as an information theoretic complexity measure; recall the association between coding length and Shannon entropy. This means that maxMMD favors less complex clusters in an information theoretic sense.
The function $\mu[P_k]$ has a larger norm if (the image of) its support is narrower. In light of this view, small entropies relate to criteria that favor small, dense clusters. For graphs, those are densely connected subgraphs. Construct a graph with vertices X_i. Let the kernel matrix $\mathbf{K}$ be its adjacency matrix (with $k(X, X) = c$ for all X) and cluster C_k have n_k nodes. Let $m(C_k)$ denote the sum of the weights of the within-cluster edges. Then maxMMD promotes the average connectedness of a vertex within its cluster:

$$\hat{\pi}_k \|\mu[\hat{P}_k]\|^2_{\mathcal{H}} = \hat{\pi}_k^{-1} (\alpha^{(k)})^\top K \alpha^{(k)} = nc + n\frac{m(C_k)}{n_k}.$$

Edge density is vital in the modularity criterion (Newman and Girvan, 2004) that compares the achieved to the expected within-cluster connectivity.

5 Regularization

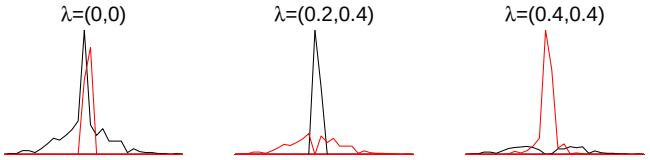

Fig. 3. Effect of increasing regularization with Gaussian kernel (estimates by binning).

As we have seen, the maxMMD criterion is well supported by classical clustering approaches. That said, clustering by maxMMD alone can result in $\hat{P}_k$ with non-smooth, steep boundaries, as in Figure 3. Smoothness of a function f in form of "simplicity" is commonly measured by the norm $\|f\|_{\mathcal{H}}$, for example in Support Vector Machines. To avoid steep or narrow $\hat{P}_k$ — unless the data allows for that — we add a smoothness term for the estimates $f_{\mu[\hat{P}_k]}$,

$$\rho(\alpha^{(1)}, \alpha^{(2)}) := \lambda_1 \|\mu[\hat{P}_1]\|^2_{\mathcal{H}} + \lambda_2 \|\mu[\hat{P}_2]\|^2_{\mathcal{H}}.$$

The weights $\hat{\pi}_k$ do not come into play here. If $\mu[P_k]$ is small, then its support in $\mathcal{H}$ is broader, and P_k has higher entropy (uncertainty). Thus, constraining the norms can avoid "overfitting". Furthermore, we can introduce prior knowledge about different entropies of the clusters by choosing $\lambda_1 \neq \lambda_2$.
To enforce overlap and smoothness, more direct restriction of the assignments $\alpha^{(k)}$ is also conceivable, similar to the updates in soft k-means. This type of regularization is motivated by the analogy (9) to kernel k-means. Both ways of regularization restrict the set of candidate distributions for P_k.

6 Conclusion

The literature on clustering is overwhelming, and it is difficult even to get an overview of what is out there. We believe that it is very important to "tidy up" and discover relationships between different clustering problems and algorithms. In this paper we study a generalized clustering problem which considers clustering from a higher level point of view, based on embeddings of distributions to high dimensional vector spaces. This approach reveals connections to the concepts behind many well-known clustering criteria.

Acknowledgments. We thank Bob Williamson, Mark Reid and Dominik Janzing for discussions.

References

M. Aghagolzadeh, H. Soltanian-Zadeh, B. Araabi, and A. Aghagolzadeh. A hierarchical clustering based on maximizing mutual information. In *ICIP*, 2007.

S. Arora and R. Kannan. Learning mixtures of arbitrary Gaussians. In *STOC*, 2001.

K. Chaudhuri and S. Rao. Beyond Gaussians: Spectral methods for learning mixtures of heavy-tailed distributions. In *COLT*, 2008.

D. Erdogmus and J. C. Principe. From linear adaptive filtering to nonlinear information processing. *IEEE Signal Processing Magazine*, 23(6):14–33, November 2006.

B. Fuglede and F. Topsøe. Jensen-Shannon divergence and Hilbert space embedding. In *Proc. of the Int. Symp. on Information Theory (ISIT)*, 2004.

E. Gokcay and J. C. Principe. Information theoretic clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(2):158–170, 2002.

A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. In *ALT*, 2005.

A. Gretton, K. M. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola. A kernel method for the two-sample problem. In *NIPS*, 2006.

A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola. A kernel method for the two-sample problem. *Tech.Report 157, MPI for Biol. Cyb.*, 2008.

R. Jenssen, D. Erdogmus, J. Principe, and T. Eltoft. The laplacian pdf distance: a cost function for clustering in a kernel feature space. In *NIPS*, 2004.

J. Lin. Divergence measures based on the Shannon entropy. *IEEE Transaction on Information Theory*, 37(1), 1991.

M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69, 2004.

A. Renyi. On measures of entropy and information. In *Proc. Fourth Berkeley Symp. Math., Statistics and Probability*, pages 547–561, 1960.

K. Rose. A mapping approach to rate-distortion computation and analysis. *IEEE Transactions on Information Theory*, 40(6):1939–1952, 1994.

L. Song, A. Smola, A. Gretton, and K. Borgwardt. A dependence maximization view of clustering. In *24th International Conference on Machine Learning (ICML)*, 2007.

B. K. Sriperumbudur, A. Gretton, K. Fukumizu, G. Lanckriet, and B. Schölkopf. Injective Hilbert space embeddings of probability measures. In *COLT*, 2008.

A. Wächter and L. T. Biegler. On the implementation of a primal-dual interior point filter line search algorithm for large-scale nonlinear programming. *Mathematical Programming*, 106(1):25–57, 2006.

L. Xu, J. Neufeld, B. Larson, and D. Schuurmans. Maximum margin clustering. In *NIPS*, 2005.