Nonparametric Clustering Based on Energy Statistics

Guilherme França
Johns Hopkins University
guifranca@gmail.com

Joshua T. Vogelstein
Johns Hopkins University
jovo@jhu.edu

Abstract

blabla

1 Introduction

Mention why energy is important, main results, where it was applied, etc. Motivate how this can be used for clustering. Mention most important papers on this . . . Explain main results of this paper and give a brief outline.

2 Energy Statistics and RKHS

In this section we briefly review the main concepts from energy statistics and its relation to reproducing kernel Hilbert spaces (RKHS), which form the basis of our work. For more details we refer the reader to [1] (and references therein) and also [2].

Consider random variables in $\mathbb{R}^D$ such that $X, X' \overset{iid}{\sim} P$ and $Y, Y' \overset{iid}{\sim} Q$, where P and Q are cumulative distribution functions with finite first moments. The quantity [1]

$$\mathcal{E}(P, Q) = 2\mathbb{E}\|X - Y\| - \|X - X'\| - \|Y - Y'\|, \tag{1}$$

called *energy distance*, is rotational invariant and nonnegative, $\mathcal{E}(P, Q) \geq 0$, where equality to zero holds if and only if $P = Q$. Above, $\|\cdot\|$ denotes the Euclidean norm in $\mathbb{R}^D$. Energy distance provides a characterization of equality of distributions, and $\mathcal{E}^{1/2}$ is a metric on the space of distributions.

The energy distance can be generalized as, for instance,

$$\mathcal{E}_\alpha(P, Q) = 2\mathbb{E}\|X - Y\|^\alpha - \|X - X'\|^\alpha - \|Y - Y'\|^\alpha, \tag{2}$$

where $0 < \alpha \leq 2$. This quantity is also nonnegative, $\mathcal{E}_\alpha(P, Q) \geq 0$. Furthermore, for $0 < \alpha < 2$ we have that $\mathcal{E}_\alpha(P, Q) = 0$ if and only if $P = Q$, while for $\alpha = 2$ we have $\mathcal{E}_2(P, Q) = 2\|\mathbb{E}(X) - \mathbb{E}(Y)\|^2$, showing that equality to zero only requires equality of the means, and thus it does not imply equality of distributions.

It is important to mention that (2) can be even further generalized. Let $X, Y \in \mathcal{X}$ and replace the Euclidean norm by $\rho : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$, i.e. $\|X - Y\| \to \rho(X, Y)$, where ρ is a so-called semimetric of negative type [2], which satisfy

$$\sum_{i,j=1}^{n} \alpha_i \alpha_j \rho(X_i, X_j) \leq 0, \tag{3}$$

where $X_i \in \mathcal{X}$, and $\alpha_i \in \mathbb{R}$ such that $\sum_{i=1}^n \alpha_i = 0$. In this case, there is a Hilbert space $\mathcal{H}$ and a map $\varphi : \mathcal{X} \to \mathcal{H}$ such that $\rho(X, Y) = \|\varphi(X) - \varphi(Y)\|^2_{\mathcal{H}}$. Even though the semimetric ρ may not satisfy the triangle inequality, $\rho^{1/2}$ does since it can be shown to be a legit metric.

There is an equivalence between energy distance, commonly used in statistics, and distances between embeddings of distributions in RKHS, commonly used in machine learning. This equivalence was

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

established in [2]. Let us first recall the definition of RKHS. Let $\mathcal{H}$ be a Hilbert space of real-valued functions over $\mathcal{X}$. A function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a reproducing kernel of $\mathcal{H}$ if it satisfies the following two conditions:

1. $h_x \equiv K(\cdot, x) \in \mathcal{H}$ for any $x \in \mathcal{X}$.
2. $\langle h_x, f \rangle_{\mathcal{H}} = f(x)$ for any $x \in \mathcal{X}$ and $f \in \mathcal{H}$.

In other words, for any $x \in \mathcal{X}$ there is a unique function $h_x \in \mathcal{H}$ that reproduces $f(x)$ through the inner product of $\mathcal{H}$. If such a *kernel* function K exists, then $\mathcal{H}$ is called a RKHS. From this we have $\langle h_x, h_y \rangle = h_y(x) = K(x, y)$. This implies that $K(x, y)$ is symmetric and positive definite, $\sum_{i,j=1}^{n} c_i c_j K(x_i, x_j) \geq 0$ for $c_i, c_j \in \mathbb{R}$.

The Moore-Aronszajn theorem establishes the converse [3]. For every symmetric and positive definite function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, there is an associated RKHS, $\mathcal{H}_K$, with reproducing kernel K. The map $\varphi : x \mapsto h_x \in \mathcal{H}_K$ is called the canonical feature map. Given a kernel K, this theorem enables us to define an embedding of a probability measure P into the RKHS: $P \mapsto h_P \in \mathcal{H}_K$ such that $\int f(x) dP(x) = \langle f, h_P \rangle$ for all $f \in \mathcal{H}_K$, or alternatively $h_P = \int K(\cdot, x) dP(x)$. We can now introduce the notion of distance between two probability measures using the inner product of $\mathcal{H}_K$. This is called the maximum mean discrepancy (MMD) and is given by

$$\gamma_K(P, Q) = \| h_P - h_Q \|_{\mathcal{H}_K}, \tag{4}$$

which can also be written as [4]

$$\gamma_K^2(P, Q) = \mathbb{E} K(X, X') + \mathbb{E} K(Y, Y') - 2\mathbb{E} K(X, Y) \tag{5}$$

where $X, X' \overset{iid}{\sim} P$ and $Y, Y' \overset{iid}{\sim} Q$. From the equality between (4) and (5) we also have

$$\langle h_P, h_Q \rangle_{\mathcal{H}_K} = \mathbb{E}\, K(X, Y). \tag{6}$$

Therefore, in practice, we can estimate the inner product between the embedded distributions by averaging the kernel function over sampled data.

The following important result shows that semimetrics of negative type and symmetric positive definite kernels are closely related [5]. Let $\rho : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a semimetric, and $x_0 \in \mathcal{X}$ an arbitrary but fixed point. Define

$$K(x, y) = \tfrac{1}{2} \left\{ \rho(x, x_0) + \rho(y, x_0) - \rho(x, y) \right\}. \tag{7}$$

Then, K is positive definite if and only if ρ is of negative type (3). Here we have a family of kernels, one for each choice of x_0. Conversely, if ρ is a semimetric of negative type and K is a kernel in this family, then

$$\rho(x, y) = K(x, x) + K(y, y) - 2K(x, y) = \| h_x - h_y \|_{\mathcal{H}_K}^2, \tag{8}$$

and the canonical feature map $\varphi : x \mapsto h_x$ is injective [2]. We say that the kernel K generates the semimetric ρ. If two different kernels generate the same ρ, they are equivalent kernels.

Now we can state the equivalence between energy distance $\mathcal{E}$ and inner products on RKHS, which is one of the main results of [2]. If ρ is a semimetric of negative type and K a kernel that generates ρ, then

$$\begin{aligned} \mathcal{E}(P, Q) &\equiv 2\mathbb{E}\, \rho(X, Y) - \mathbb{E}\, \rho(X, X') - \mathbb{E}\, \rho(Y, Y') && (9) \\ &= 2\left[\mathbb{E}\, K(X, X') + \mathbb{E}\, K(Y, Y') - 2\mathbb{E}\, K(X, Y) \right] && (10) \\ &= 2\gamma_K^2(P, Q) && (11) \end{aligned}$$

This result follows simply by substituting (8) into (9), and using (5). Since $\gamma_k^2(P, Q) = \| h_P - h_Q \|_{\mathcal{H}_K}^2$, we can compute the energy distance using the inner product of $\mathcal{H}_K$. Moreover, this can be computed from the data according to (6).

3 Energy Distance based Clustering

Now we formulate an optimization problem for clustering based on energy statistics and RKHS, introduced in the previous section. Assume we have data $\mathbb{X} = \{x_1, x_2, \ldots, x_n\}$, where $x_i \in \mathcal{X}$ is in some space of negative type (3), and a partition $\mathbb{X} = \cup_{j=1}^{k} \mathcal{C}_j$, where $\mathcal{C}_i \cap \mathcal{C}_j = \emptyset$. Each expectation in (9) can be computed through the function

$$g(\mathcal{C}_i, \mathcal{C}_j) \equiv \frac{1}{n_i n_j} \sum_{x \in \mathcal{C}_i} \sum_{y \in \mathcal{C}_j} \rho(x, y) \tag{12}$$

where $n_i = |\mathcal{C}_i|$ is the number of elements in partition $\mathcal{C}_i$. In energy statistics [1] we have the within energy dispersion

$$W = \sum_{j=1}^{k} \frac{n_j}{2} g(\mathcal{C}_j, \mathcal{C}_j), \tag{13}$$

and also the between-sample energy statistic

$$S = \sum_{1 \le j < l \le k} \frac{n_j n_l}{2n} \left[2g(\mathcal{C}_j, \mathcal{C}_l) - g(\mathcal{C}_j, \mathcal{C}_j) - g(\mathcal{C}_l, \mathcal{C}_l) \right], \tag{14}$$

where $n = \sum_{j=1}^{k} n_j$. Given a set of distributions $\{P_j\}_{j=1}^{k}$ where $x \in \mathcal{C}_j \sim P_j$, the quantity S_α provides a *nonparametric* test statistic for equality of distributions. Under the null hypothesis $H_0 : P_1 = P_2 = \cdots = P_k$, S_α is small, and under the alternative hypothesis $H_1 : P_i \neq P_j$ for at least two $i \neq j$, $S_\alpha \to \infty$ as the sample size is large, $n \to \infty$. This test is nonparametric in the sense that it does not make any assumptions about the distributions P_j.

Based on this test statistic for equality of distributions, a possible criteria for clustering data is to maximize S. However, it can be shown that the total dispersion of the data obeys [1]

$$T(\mathbb{X}) = W + S = \frac{n}{2} g(\mathbb{X}, \mathbb{X}). \tag{15}$$

Note that T only depends on the pooled data, so it does not depend on how we partition $\mathbb{X}$. Therefore, maximizing S is equivalent to minimizing W, which has a simpler form. Thus, our clustering problem corresponds to find the best partition of $\mathbb{X}$ to solve the optimization problem

$$\min_{\{\mathcal{C}_j\}} W(\{\mathcal{C}_1, \ldots, \mathcal{C}_k\}) \tag{16}$$

where each datapoint belongs to one and only one partition (hard assignments).

Now fix an arbitrary point $x_0 \in \mathcal{X}$, and suppose that the kernel K generates ρ, such that (7) and (8) hold. Then we can write (13) as

$$W(\{\mathcal{C}_j\}) = \frac{1}{2} \sum_{j=1}^{k} \frac{1}{n_j} \sum_{x,y \in \mathcal{C}_j} \rho(x, y) = \sum_{j=1}^{k} \sum_{x \in \mathcal{C}_j} \left(K(x, x) - \frac{1}{n_j} \sum_{y \in \mathcal{C}_j} K(x, y) \right) \tag{17}$$

When minimizing W, the first term is just a constant and does not contribute, therefore the problem reduces to

$$\max_{\{\mathcal{C}_j\}} \sum_{j=1}^{k} \frac{1}{n_j} \sum_{x,y \in C_j} K(x, y). \tag{18}$$

Let us introduce the binary matrix $Z \in \{0, 1\}^{n \times k}$ such that

$$Z_{ij} = \begin{cases} 1 & \text{if } x_i \in \mathcal{C}_j \\ 0 & \text{otherwise.} \end{cases} \tag{19}$$

Notice that $D = Z^T Z = \operatorname{diag}(n_1, n_2, \ldots, n_k)$ contains the number of elements in each partition. Introducing the kernel matrix $G \in \mathbb{R}^{n \times n}$ such that

$$G_{ij} = K(x_i, x_j), \tag{20}$$

then (18) is equal to $\max \operatorname{Tr} \left\{ D^{-1} Z^\top G Z \right\}$. Therefore, our clustering optimization problem based on energy statistics can be formulated as

$$\begin{aligned} &\max_{Z} \operatorname{Tr} \left\{ \left(Z D^{-1/2} \right)^\top G \left(Z D^{-1/2} \right) \right\} \\ &\text{s.t. } Z_{ij} \in \{0, 1\}, \textstyle\sum_{j=1}^{k} Z_{ij} = 1, \sum_{i=1}^{n} Z_{ij} = n_j, \text{ and } D = Z^\top Z. \end{aligned} \tag{21}$$

This is a quadratic problem with integer constraints, which is NP-hard. Let us now write this in terms of $Y \equiv ZD^{-1/2}$, which componentwise is

$$Y_{ij} = \begin{cases} \frac{1}{\sqrt{n_j}} & \text{if } x_i \in \mathcal{C}_j \\ 0 & \text{otherwise.} \end{cases} \tag{22}$$

We thus have

$$\max_{Y} \operatorname{Tr}\left\{ Y^\top G Y \right\} \qquad \text{s.t. } Y \ge 0, Y^\top Y = I, YY^\top \boldsymbol{e} = \boldsymbol{e}, \tag{23}$$

where $\boldsymbol{e} = (1, 1, \dots, 1)^\top \in \mathbb{R}^n$ is the all-ones vector, and G is the pairwise kernel matrix (20) obtained from (7). Therefore, to cluster data $\{x_i\}_{i=1}^n \in \mathcal{X}$ into k partitions, assuming that k is given, we first compute G — which is defined by an arbitrary semimetric of negative type on $\mathcal{X}$ — and then solve the optimization problem (23) for $Y \in \mathbb{R}^{n \times k}$. The ith row of Y will contain a single nonzero element in some jth column, indicating that $x_i \in \mathcal{C}_j$.

In general, problem (23) is NP-hard since it is a quadratically constrained quadratic problem (QCQP). There are few methods available to tackle this kind of problem directly, which is computational prohibitive even for relatively small datasets. However, one can find an approximate solution by relaxing some of the constraints. For instance, an approximation can be given by $\max_Y \operatorname{Tr}\left\{Y^\top G Y\right\}$ subject to $Y^\top Y = I$, and requiring that the rows of Y are normalized. It is possible to find a global solution to this problem by choosing Y as the top k eigenvectors of G, which results in $\max \operatorname{Tr}\left\{Y^\top G Y\right\} = \sum_{i=1}^k \lambda_i(G)$, which is the sum of the top k eigenvalues of G.

It is important to note that (23) has the same formulation as kernel k-means, spectral clustering, and the maximum cut problem on graphs [6]. The result (23) brings energy statistics based clustering into this broad picture, and (23) should have interesting applications in graph partitioning problems and unsupervised learning in general. Furthermore, and most importantly, our analysis is valid for any space $\mathcal{X}$ equiped with a semimetric of negative type ρ. This method is nonparametric since it does not assume any form of the distribution of the data, contrary to k-means and gaussian mixture models (GMM), for example. Also, there is no concept of cluster center involved in this approach.

3.1 A Simple Algorithm

We can formulate an iterative algorithm to find an approximate solution to (23) on the same lines as kernel k-means. Let t be the iteration time. First precompute the kernel matrix G, fix the number of clusters k, then perform the following steps:

1. Initialize clusters $\{\mathcal{C}_1^{(0)}, \dots, \mathcal{C}_k^{(0)}\}$, which determines the label matrix $Y^{(0)}$.
2. For each datapoint x_i compute its cluster assignment through

$$Y_{ij}^{(t+1)} = \begin{cases} \frac{1}{\sqrt{n_j}} & \text{if } j = \arg\max_\ell \frac{1}{n_\ell} \sum_{m=1}^n G_{im} Y_{m\ell}^{(t)} \\ 0 & \text{otherwise.} \end{cases} \tag{24}$$

3. If converged return $Y^{(t+1)}$, otherwise set $t = t + 1$ and repeat step 2.

If data is D-dimensional, computing G has complexity $O(n^2 D)$. Step 2 above has complexity $O(n)$ for each point, thus total complexity $O(n^2)$. Assuming we perform T iterations, the total complexity of the algorithm is $O(n^2(D + T))$. We can initialize the algorithm in step 1 with any method we want, a good alternative is the initialization from k-means++.

3.2 Two-Class Problem in One Dimension

If data is one-dimensional and we choose $\rho(x, y) = |x - y|$, we can actually compute (12) in $O(n \log n)$ instead of $O(n^2 D)$ and find a direct solution to (16). This is done by noticing that

$$|x - y| = \mathbb{1}(x \ge y)(x - y) - \mathbb{1}(x < y)(x - y). \tag{25}$$

Denote

$$n_j^-(x) = \sum_{y \in \mathcal{C}_j} \mathbb{1}(y \le x), \quad n_j^+(x) = \sum_{y \in C_j} \mathbb{1}(y > x), \quad \mathcal{D}_j(x) = \frac{n_j^-(x) - n_j^+(x)}{n_j} \tag{26}$$

where $n_j^+(x)$ is the number of elements in $\mathcal{C}_j$ which are larger than $x \in \mathcal{C}_i$, and so on. Now (12) can be written as

$$g(\mathcal{C}_i, \mathcal{C}_j) = \frac{1}{n_i} \sum_{x \in \mathcal{C}_i} \mathcal{D}_j(x)\, x - \frac{1}{n_j} \sum_{y \in \mathcal{C}_j} \mathcal{D}_i(y)\, y. \tag{27}$$

If we merge $\mathcal{C}_i$ and $\mathcal{C}_j$ and sort the data, this expression can be computed in $O(n)$, where $n = n_i + n_j$. For a two-class problem, we have to loop through each point and compute (13) to find the best split of the data. Since sorting sorting usually takes $O(n \log n)$, the complexity of this procedure is $O(n(\log n + n)) = O(n^2)$.

4 Numerical Experiments

5 Conclusion

Acknowledgements

We thank ...

References

[1] G. J. Székely and M. L. Rizzo. Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143:1249–1272, 2013.

[2] D. Sejdinovic, B. Sriperumbudur, A. Gretton, and K. Fukumizu. Equivalence of Distance-Based and RKHS-Based Statistic in Hypothesis Testing. *The Annals of Statistics*, 41(5):2263–2291, 2013.

[3] N. Aronszajn. Theory of Reproducing Kernels. *Transactions of the American Mathematical Society*, 68:337–404, 1950.

[4] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A Kernel Two-Sample Test. *J. Mach. Learn. Res.*, 13:723–773, 2012.

[5] C. Berg, J. P. R. Christensen, and P. Ressel. *Harmonic Analysis on Semigroups: Theory of Positive Definite and Related Functions*. Graduate Text in Mathematics 100. Springer, New York, 1984.

[6] I. S. Dhillon, Y. Guan, and B. Kulis. Kernel K-means: Spectral Clustering and Normalized Cuts. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pages 551–556, New York, NY, USA, 2004. ACM.