Generalized Clustering via kernel embeddings

Stefanie Jegelka, Arthur Gretton, Bernhard Schölkopf,
Bharath K. Sriperumbudur, Ulrike von Luxburg

MAX-PLANCK-GESELLSCHAFT

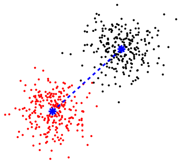
Max Planck Institute
for Biological Cybernetics

Tübingen, Germany

BIOLOGISCHE KYBERNETIK

Idea

Decompose sample into...

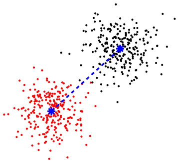


locally distinct clusters
separation by
first-order moments (means)

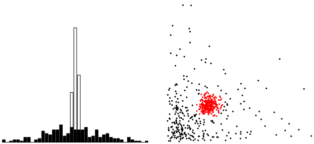
Mixture Model: $P = \pi_1 P_1 + \pi_2 P_2$

Idea

Decompose sample into...



generalize $\longrightarrow$



locally distinct clusters
separation by
first-order moments (means)

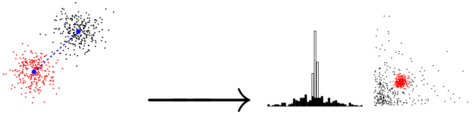
distinct distributions
separation by
higher-order moments

variance, kurtosis, ...

Mixture Model: $P = \pi_1 P_1 + \pi_2 P_2$

Decomposition

$$P = \pi_1 P_1 + \pi_2 P_2$$

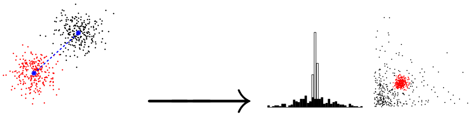


first-order moments (means) higher-order moments

$$\max_{P_1, P_2} \quad \underbrace{D(P_1, P_2)}_{\text{max. discrepancy}} + \underbrace{\lambda \Omega(P_1, P_2)}_{\text{favor "simplicity"}}$$

Decomposition

$$P = \pi_1 P_1 + \pi_2 P_2$$



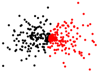
first-order moments (means) higher-order moments

$$\max_{P_1,P_2} \quad {\color{red} D(P_1, P_2)} \quad + \quad \lambda\Omega(P_1, P_2)$$

max. discrepancy favor "simplicity"

Which Discrepancy Measure?

means $\left|\mathbb{E}_{x\sim P_1}[x] - \mathbb{E}_{y\sim P_2}[y]\right| = |\mu_1 - \mu_2|$

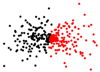


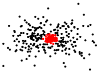
Which Discrepancy Measure?

means $\left|\mathbb{E}_{x\sim P_1}[x] - \mathbb{E}_{y\sim P_2}[y]\right| = |\mu_1 - \mu_2|$

variance $\left|\mathbb{E}_{x\sim P_1}[(x-\mu_1)^2] - \mathbb{E}_{y\sim P_2}[(y-\mu_2)^2]\right|$

dth moment $\left|\mathbb{E}_{x\sim P_1}[x^d] - \mathbb{E}_{y\sim P_2}[y^d]\right|$





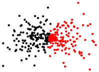
Which Discrepancy Measure?

means $\left|\mathbb{E}_{x\sim P_1}[x] - \mathbb{E}_{y\sim P_2}[y]\right| = |\mu_1 - \mu_2|$

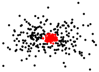
variance $\left|\mathbb{E}_{x\sim P_1}[(x-\mu_1)^2] - \mathbb{E}_{y\sim P_2}[(y-\mu_2)^2]\right|$

dth moment $\left|\mathbb{E}_{x\sim P_1}[x^d] - \mathbb{E}_{y\sim P_2}[y^d]\right|$

general $\left|\mathbb{E}_{x\sim P_1}[g(x)] - \mathbb{E}_{y\sim P_2}[g(y)]\right|$



which “discriminative” g?



Kernel Framework: Maximum Mean Discrepancy

$$P = \pi_1 P_1 + \pi_2 P_2 \qquad \max_{P_1, P_2} \underbrace{D(P_1, P_2)}_{\text{discrepancy}} + \lambda \underbrace{\Omega(P_1, P_2)}_{\text{"simplicity"}}$$

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

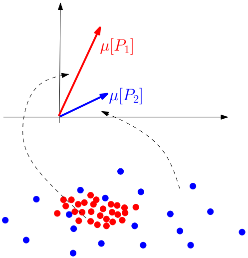
- implementation ?

Kernel Framework: Maximum Mean Discrepancy

$$P = \pi_1 P_1 + \pi_2 P_2$$

$$\max_{P_1,P_2} \underbrace{D(P_1,P_2)}_{\text{discrepancy}} + \lambda \underbrace{\Omega(P_1,P_2)}_{\text{"simplicity"}}$$

$$\mathrm{MMD}(P_1,P_2) = \sup_{g\in\mathcal{H},\|g\|_{\mathcal{H}}\le 1} \left|\mathbb{E}_{x\sim P_1} g(x) - \mathbb{E}_{y\sim P_2} g(y)\right|$$



Embedding

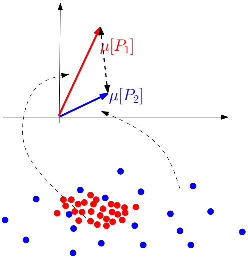
- represent each P_i by a mean function $\mu[P_i]$ in Hilbert space $\mathcal{H}$

Kernel Framework: Maximum Mean Discrepancy

$$P = \pi_1 P_1 + \pi_2 P_2$$

$$\max_{P_1,P_2} \underbrace{D(P_1,P_2)}_{\text{discrepancy}} + \lambda \underbrace{\Omega(P_1,P_2)}_{\text{“simplicity”}}$$

$$\begin{aligned}\mathrm{MMD}(P_1,P_2) &= \sup_{g\in\mathcal{H},\|g\|_{\mathcal{H}}\le 1} \left|\mathbb{E}_{x\sim P_1} g(x) - \mathbb{E}_{y\sim P_2} g(y)\right| \\ &= \|\mu[P_1]-\mu[P_2]\|_{\mathcal{H}}\end{aligned}$$



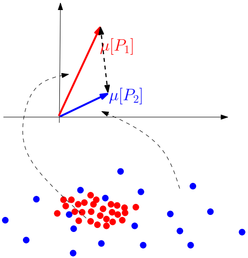
- represent each P_i by a mean function $\mu[P_i]$ in Hilbert space $\mathcal{H}$
- discrepancy steerable via kernel

Kernel Framework: Maximum Mean Discrepancy

$$P = \pi_1 P_1 + \pi_2 P_2$$

$$\max_{P_1,P_2} \underbrace{D(P_1,P_2)}_{\text{discrepancy}} + \lambda \underbrace{\Omega(P_1,P_2)}_{\text{"simplicity"}}$$

$$\pi_1\pi_2 \mathrm{MMD}(P_1,P_2)^2 = \pi_1\pi_2 \sup_{g\in\mathcal{H},\|g\|_\mathcal{H}\le 1} \left|\mathbb{E}_{x\sim P_1} g(x) - \mathbb{E}_{y\sim P_2} g(y)\right|^2$$
$$= \pi_1\pi_2 \|\mu[P_1] - \mu[P_2]\|_\mathcal{H}{}^2$$

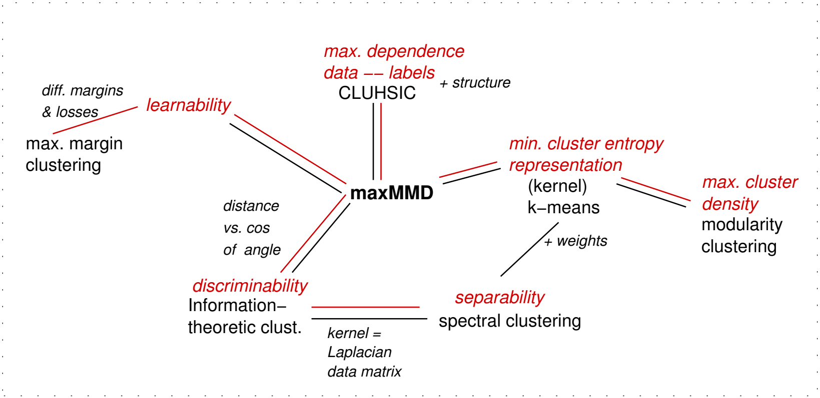


- represent each P_i by a mean function $\mu[P_i]$ in Hilbert space $\mathcal{H}$
- discrepancy steerable via kernel

Connections between clustering concepts – why?

- better understanding of methods – better understanding of results
- new algorithms for old objectives by transfer

Connections



Generalization of K-means

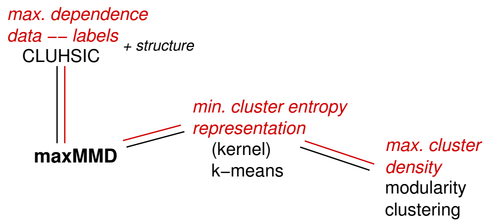
assignment of point x_i to cluster j $\searrow$ | x_i mapped into $\mathcal{H}$ $\downarrow$ | mean representative of "cluster" P_j $\swarrow$

$$D(P_1, P_2) = -\sum_{i=1}^{n} \sum_{j=1,2} \alpha_{ij} \|\varphi(x_i) - \mu[P_j]\|^2 + \text{const}$$

max distance of means $\rightarrow$ max discrepancy of moments

min variance $\rightarrow$ min "entropy"

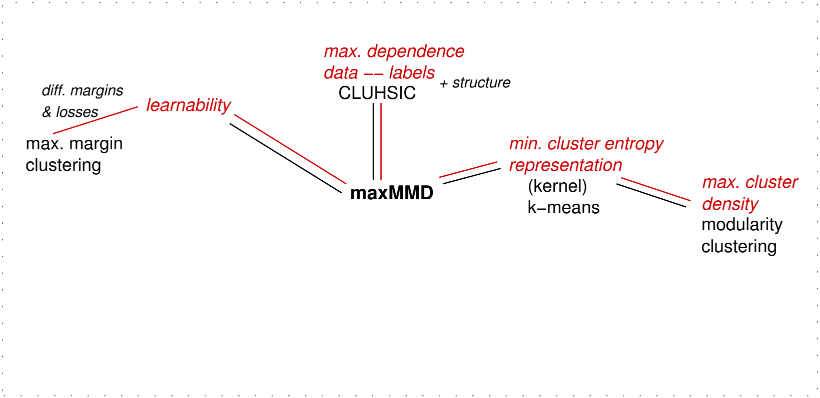
Connections



capture structure:

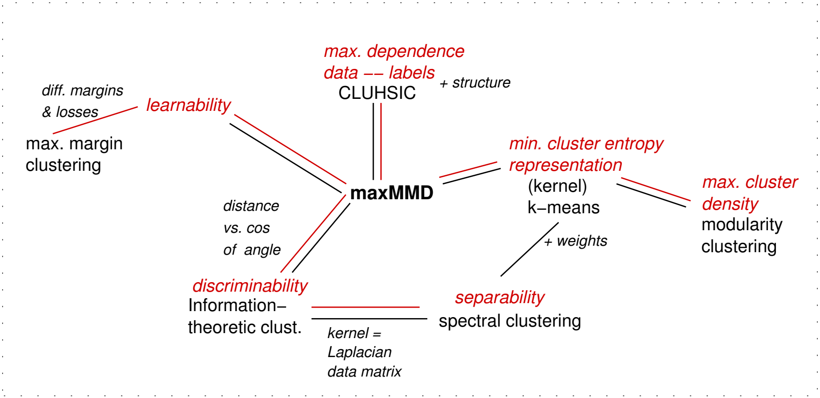
maximize statistical dependence between labels and data

Connections



capture structure:
minimize Bayes risk

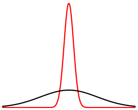
Connections

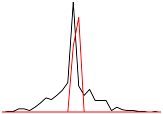


"Regularization"

$$P = \pi_1 P_1 + \pi_2 P_2$$

$$\max_{P_1,P_2} \underbrace{D(P_1,P_2)}_{\text{discrepancy}} + \lambda \underbrace{\Omega(P_1,P_2)}_{\textit{"simplicity"}}$$



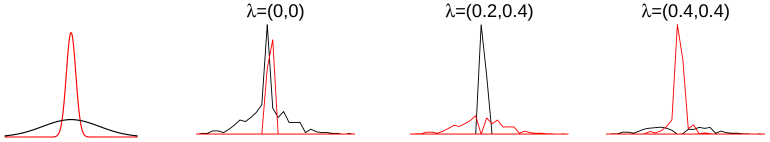


"Regularization"

$$P = \pi_1 P_1 + \pi_2 P_2 \qquad \max_{P_1,P_2} \underbrace{D(P_1,P_2)}_{\text{discrepancy}} + \lambda \underbrace{\Omega(P_1,P_2)}_{\text{"simplicity"}}$$

kernel measure of simplicity (spread)

$$\lambda\Omega(P_1,P_2) = -\lambda_1 \|\mu[P_1]\|^2_{\mathcal{H}} - \lambda_2 \|\mu[P_2]\|^2_{\mathcal{H}}$$



Implementation

- non-convex optimization problem:
  - without regularization: kernel k-means
  - with regularization: solver
- empirical performance: comparable to spectral clustering (Normalized cut), kernel k-means

Summary

- generalized clustering: by discrepancy of distributions
- flexible framework
- incorporates several clustering concepts