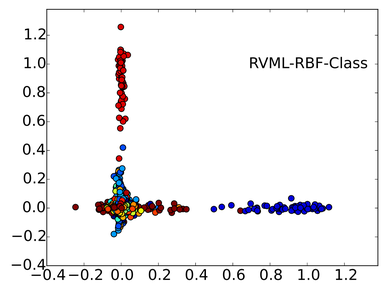


Regressive Virtual Metric Learning

Michaël Perrot and Amaury Habrard

Université Jean Monnet de Saint-Etienne, Université de Lyon,
Laboratoire Hubert Curien, CNRS, UMR5516, F-42000, Saint-Etienne, France.
{michael.perrot,amaury.habrard}@univ-st-etienne.fr

Objective: Propose a new framework for metric learning using only positive constraints between examples and a priori defined virtual points.

Metric Learning: Main Idea

Learning how to compare objects: learn a new space where some constraints are fulfilled, e.g. move closer circles of the same color (class) and keep far away circles of different colors (classes).

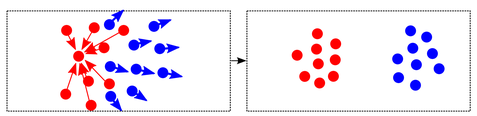
Mahalanobis-like Distance (with $\mathbf{M}$ a PSD matrix):

$$\begin{aligned} d_{\mathbf{M}}(\mathbf{x},\mathbf{x}') &= \sqrt{(\mathbf{x}-\mathbf{x}')^T\mathbf{M}(\mathbf{x}-\mathbf{x}')} \\ &= \sqrt{(\mathbf{L}^T\mathbf{x}-\mathbf{L}^T\mathbf{x}')^T(\mathbf{L}^T\mathbf{x}-\mathbf{L}^T\mathbf{x}')} \\ &= \sqrt{\mathbf{x}^T\mathbf{M}\mathbf{x}+\mathbf{x}'^T\mathbf{M}\mathbf{x}'-2\mathbf{x}^T\mathbf{M}\mathbf{x}'}. \end{aligned}$$

Metric Learning: Classical Approach

Classical approaches in metric learning use two kind of constraints:

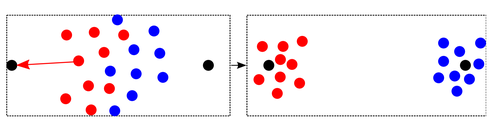
- Similarity constraints: the goal is to bring closer similar examples, e.g. examples of the same class.
- Dissimilarity constraints: the goal is to push far away dissimilar examples, e.g. examples of different classes.



It induces a quadratic number of constraints: $\mathcal{O}(n^2)$ (e.g. [JWZ09]).

Metric Learning: Virtual Points Approach

In our virtual points approach we only use similarity constraints. However instead of pairing examples with each other, we associate each example to a given virtual point.



It induces a linear number of constraints: $\mathcal{O}(n)$.

Formulation

Let $S=\{(\mathbf{x}_i,y_i)\}_{i=1}^n \subset (\mathcal{X}\times\mathcal{Y})^n$ be a set of examples. Let $f_{\mathbf{v}}:\mathcal{X}\times\mathcal{Y}\to\mathcal{V}$ where $\mathcal{V}\subseteq\mathbb{R}^{d'}$ be the function which associates each example to a virtual point. We consider the learning set $S_{\mathbf{v}}=\{(\mathbf{x}_i,\mathbf{v}_i)\}_{i=1}^n \subset \mathcal{X}\times\mathcal{V}$. Let $\mathbf{X}=(\mathbf{x}_1,\dots,\mathbf{x}_n)^T$ and $\mathbf{V}=(\mathbf{v}_1,\dots,\mathbf{v}_n)^T$, we learn $\mathbf{L}$ such that $\mathbf{M}=\mathbf{L}\mathbf{L}^T$ with the following optimization problem:

$$\min_{\mathbf{L}} \hat{R}(\mathbf{L}) + \lambda\|\mathbf{L}\|_{\mathcal{F}}^2 = \min_{\mathbf{L}} \frac{1}{n}\|\mathbf{X}\mathbf{L}-\mathbf{V}\|_{\mathcal{F}}^2 + \lambda\|\mathbf{L}\|_{\mathcal{F}}^2. \quad (1)$$

Using the closed form solution for $\mathbf{L}$, we get:

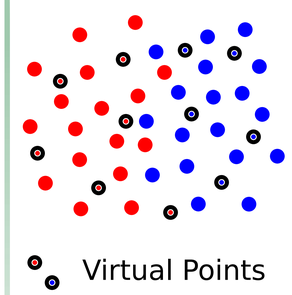
$$\mathbf{M}=\mathbf{L}\mathbf{L}^T=\mathbf{X}^T\left(\mathbf{X}\mathbf{X}^T+\lambda n\mathbf{I}\right)^{-1}\mathbf{V}\mathbf{V}^T\left(\mathbf{X}\mathbf{X}^T+\lambda n\mathbf{I}\right)^{-1}\mathbf{X}. \quad (2)$$

Let $K_{\mathbf{X}}=\phi(\mathbf{X})\phi(\mathbf{X})^T$, we can kernelize our approach as:

$$d^2_{\mathbf{M}_K}(\phi(\mathbf{x}),\phi(\mathbf{x}')) = \phi(\mathbf{x})^T\mathbf{M}_K\phi(\mathbf{x}) + \phi(\mathbf{x}')^T\mathbf{M}_K\phi(\mathbf{x}') - 2\phi(\mathbf{x})^T\mathbf{M}_K\phi(\mathbf{x}')$$

with $\mathbf{M}_K=\phi(\mathbf{X})^T\left(K_{\mathbf{X}}+\lambda n\mathbf{I}\right)^{-1}\mathbf{V}\mathbf{V}^T\left(K_{\mathbf{X}}+\lambda n\mathbf{I}\right)^{-1}\phi(\mathbf{X})$.

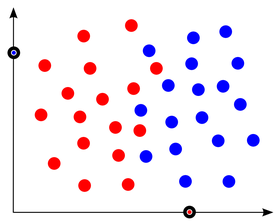
Virtual Points Selection



Optimal Transport Based Approach This is a two steps approach:

- Selection of several landmarks in the training set using a diversity criteria [KJ11].
- Application of Optimal Transport with regularization [CFT14] between the training set and the landmarks. The points obtained after transport are used as virtual points.

Class Based Representation Space Approach The virtual points are defined as unit vectors of a space of dimension the number of classes, i.e. there is one class associated with one virtual point. Each example is associated with a virtual point using its class.



Theoretical Analysis: Bounding the true risk of [JWZ09] by the empirical risk of our approach

Theorem 1. *Let $\mathcal{D}$ be a distribution over $\mathcal{X}\times\mathcal{Y}$. Let $\mathcal{V}\subset\mathbb{R}^{d'}$ be a finite set of virtual points and $f_{\mathbf{v}}$ is defined as $f_{\mathbf{v}}(\mathbf{x}_i,y_i)=\mathbf{v}_i$, $\mathbf{v}_i\in\mathcal{V}$. Let $\|\mathbf{v}\|_2\leq C_{\mathbf{v}}$ for any $\mathbf{v}\in\mathcal{V}$ and $\|\mathbf{x}\|_2\leq C_{\mathbf{x}}$ for any $\mathbf{x}\in\mathcal{X}$. Let $\gamma_1=2\max_{\mathbf{x}_k,\mathbf{x}_l,y_{kl}=1} d^2(\mathbf{v}_k,\mathbf{v}_l)$ and $\gamma_{-1}=\frac{1}{2}\min_{\mathbf{x}_k,\mathbf{x}_l,y_{kl}=-1} d^2(\mathbf{v}_k,\mathbf{v}_l)$, we have:*

$$\mathbb{E}_{(\mathbf{x}_i,y_i)\sim\mathcal{D},(\mathbf{x}_j,y_j)\sim\mathcal{D}}\left[y_{ij}(d^2(\mathbf{L}^T\mathbf{x}_i,\mathbf{L}^T\mathbf{x}_j)-\gamma_{y_{ij}})\right]_+ \leq 8\left(\hat{R}(\mathbf{L})+\frac{8C_{\mathbf{v}}^2C_{\mathbf{x}}^2}{\lambda n}\left(1+\frac{C_{\mathbf{x}}}{\sqrt{\lambda}}\right)^2+\left(\left(\frac{16C_{\mathbf{x}}^2}{\lambda}+1\right)C_{\mathbf{v}}^2\left(1+\frac{C_{\mathbf{x}}}{\sqrt{\lambda}}\right)^2\right)\sqrt{\frac{\ln\frac{1}{\delta}}{2n}}\right).$$

with $y_{ij}=1$ for examples of the same class and -1 otherwise. Note that the margins are expressed w.r.t. the distances between virtual points.

Experiments

	Baselines			Our approach	
Base	1NN	LMNN	SCML	RVML-Lin-OT	RVML-Lin-Class
Amazon	41.51 ± 3.24	65.50 ± 2.28	71.68 ± 1.86	71.62 ± 1.34	**73.09 ± 2.49**
Caltech	18.04 ± 2.20	49.68 ± 2.76	52.84 ± 1.61	52.51 ± 2.41	**55.41 ± 2.55***
DSLR	29.61 ± 4.38	**76.08 ± 4.79**	65.10 ± 9.00	74.71 ± 5.27	75.29 ± 5.08
Isolet	88.97	**95.83**	89.61	91.40	94.61
Letters	94.74 ± 0.27	**96.43 ± 0.28***	96.13 ± 0.20	90.25 ± 0.60	95.51 ± 0.26
Splice	71.17	82.02	**85.43**	84.64	78.44
Svmguide1	**95.12**	95.03	87.38	94.83	85.25
Webcam	42.90 ± 4.19	85.81 ± 3.75	**90.43 ± 2.70**	88.60 ± 3.63	88.60 ± 2.69

	Baselines				Our approach	
Base	1NN-KPCA	LMNN-KPCA	GBLMNN	SCMLLOCAL	RVML-RBF-OT	RVML-RBF-Class
Amazon	20.27 ± 2.42	53.16 ± 3.73	65.53 ± 2.32	69.14 ± 1.74	73.51 ± 0.83	**76.22 ± 2.09***
Caltech	20.82 ± 8.29	29.88 ± 10.89	49.91 ± 2.80	50.56 ± 1.62	54.39 ± 1.89	**57.98 ± 2.22***
DSLR	64.90 ± 5.81	73.92 ± 7.57	76.08 ± 4.79	62.55 ± 6.94	70.39 ± 4.48	**76.67 ± 4.57**
Isolet	68.70	96.28	96.02	91.40	95.96	**96.73**
Letter	95.39 ± 0.27	**97.17*** ± 0.18**	96.51 ± 0.25	96.63 ± 0.26	91.26 ± 0.50	96.09 ± 0.21
Splice	66.99	**88.97**	82.21	87.13	88.51	88.32
Svmguide1	**95.72**	95.60	95.00	87.40	95.67	95.05
Webcam	73.55 ± 4.57	84.52 ± 3.83	85.81 ± 3.75	88.71 ± 2.83	88.71 ± 4.28	**88.92 ± 2.91**

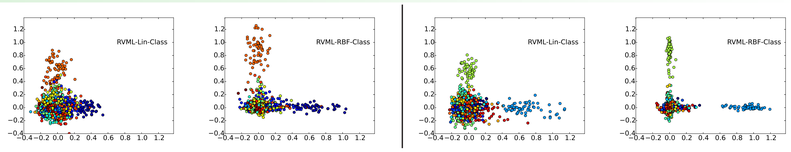


Illustration of the learned metric (2 dimensions on the Isolet dataset).

References

[BHS15] Aurélien Bellet, Amaury Habrard, and Marc Sebban. *Metric Learning*. Morgan & Claypool Publishers, 2015.

[CFT14] Nicolas Courty, Rémi Flamary, and Devis Tuia. Domain adaptation with regularized optimal transport. In *Proc. of ECML*, pages 274–289, 2014.

[JWZ09] Rong Jin, Shijun Wang, and Yang Zhou. Regularized distance metric learning: Theory and algorithm. In *Proc. of NIPS*, pages 862–870, 2009.

[KJ11] Purushottam Kar and Prateek Jain. Similarity-based learning via data driven embeddings. In *Proc. of NIPS*, pages 1998–2006, 2011.