A Theoretical Analysis of Metric Hypothesis Transfer Learning

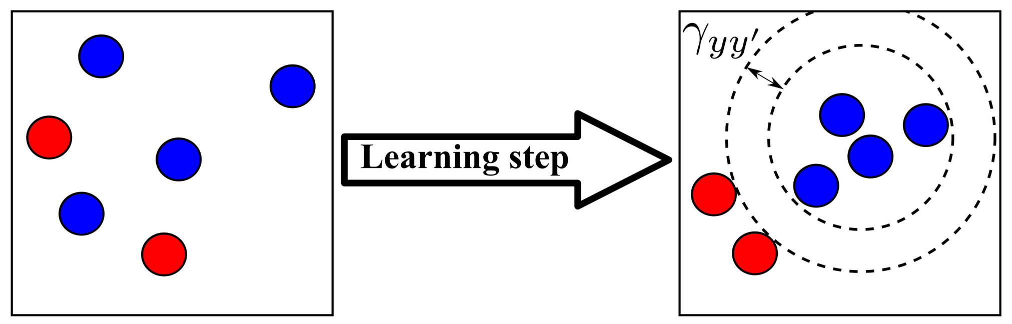
Michaël Perrot and Amaury Habrard

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

Objectives: Provide a theoretical analysis of biased regularized metric learning and propose an efficient way to reweight the source metric.

Metric Learning

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:

$$d_{\mathbf{M}}(\mathbf{x},\mathbf{x}') = \sqrt{(\mathbf{x}-\mathbf{x}')^T\mathbf{M}(\mathbf{x}-\mathbf{x}')},\ \mathbf{M} \text{ a PSD matrix.}$$

Link with some well-known distances:

- Euclidean Distance: $\mathbf{M} = \mathbf{I}$
- Original Mahalanobis Distance: $\mathbf{M} = \mathbf{\Sigma}^{-1}$
- Zero Distance: $\mathbf{M} = \mathbf{0}$

Biased Regularized Metric Learning

Let $\|\cdot\|_{\mathcal{F}}$ be the Frobenius norm, $\mathbf{M}_{\mathcal{S}}$ is a fixed metric biasing the regularization, we consider the following optimization problem w.r.t. a learning sample $T = \{z_i = (\mathbf{x}_i, y_i)\}_{i=1}^n \subset (\mathcal{X}\times\mathcal{Y})^n$:

$$\operatorname*{arg\,min}_{\mathbf{M}\succeq 0} L_T(\mathbf{M}) + \lambda\|\mathbf{M}-\mathbf{M}_{\mathcal{S}}\|_{\mathcal{F}}^2 \qquad (1)$$

where $L_T(\mathbf{M}) = \sum_{\mathbf{z},\mathbf{z}'\in T} l(\mathbf{M},\mathbf{z},\mathbf{z}')$ stands for the empirical risk with $l(\mathbf{M},\mathbf{z},\mathbf{z}')$ a convex, (σ, m)-admissible and k-lipschitz loss.
(σ, m)-admissibility: $|l(\mathbf{M},\mathbf{z}_1,\mathbf{z}_2) - l(\mathbf{M},\mathbf{z}_3,\mathbf{z}_4)| \leq \sigma\,|y_1y_2 - y_3y_4| + m$
k-lipschitz continuity: $|l(\mathbf{M},\mathbf{z},\mathbf{z}') - l(\mathbf{M}',\mathbf{z},\mathbf{z}')| \leq k\|\mathbf{M}-\mathbf{M}'\|_{\mathcal{F}}$

We use the following loss in the experiments:

$$l(\mathbf{M},\mathbf{z},\mathbf{z}') = \left[yy'((\mathbf{x}-\mathbf{x}')^T\mathbf{M}(\mathbf{x}-\mathbf{x}') - \gamma_{yy'})\right]_+ \qquad (2)$$

where $[\cdot]_+$ is the hinge loss, $yy' = 1$ for examples of the same class and -1 otherwise and $\gamma_{yy'}$ is the chosen margin.

References

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

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

[SSBD14] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms.* Cambridge University Press, 2014.

On Average Analysis: The learned metric is better than the source one

Definition 1 (On-average-replace-two-stability). *Let $\epsilon : \mathbb{N} \to \mathbb{R}$ be monotonically decreasing and let $U(n)$ be the uniform distribution over $\{1\ldots n\}$. A metric learning algorithm is on-average-replace-two-stable with rate $\epsilon(n)$ if for every distribution $\mathcal{D}_{\mathcal{T}}$:*

$$\mathop{\mathbb{E}}_{\substack{T\sim\mathcal{D}_{\mathcal{T}}{}^n\\ i,j\sim U(n)\\ \mathbf{z}_1,\mathbf{z}_2\sim\mathcal{D}_{\mathcal{T}}}} \left[l(\mathbf{M}^{ij^*},\mathbf{z}^i,\mathbf{z}^j) - l(\mathbf{M}^*,\mathbf{z}^i,\mathbf{z}^j)\right] \leq \epsilon(n)$$

where $\mathbf{M}^*$, respectively $\mathbf{M}^{ij^*}$, is the optimal solution when learning with the training set T, respectively T^{ij}. T^{ij} is obtained by replacing $\mathbf{z}^i$, the i^{th} example of T, by $\mathbf{z}_1$ to get a training set T^i and then by replacing $\mathbf{z}^j$, the j^{th} example of T^i, by $\mathbf{z}_2$.

Theorem 1 (On-average-replace-two-stability). *Given a training sample T of size n drawn i.i.d. from $\mathcal{D}_{\mathcal{T}}$, an algorithm solving optimization problem (1) is on-average-replace-two-stable with $\epsilon(n) = \frac{8k^2}{\lambda n}$.*

Theorem 2 (On average bound). *For any convex, k-lipschitz loss, we have:*

$$\mathbb{E}_{T\sim\mathcal{D}_{\mathcal{T}}{}^n}\left[L_{\mathcal{D}_{\mathcal{T}}}(\mathbf{M}^*)\right] \leq L_{\mathcal{D}_{\mathcal{T}}}(\mathbf{M}_{\mathcal{S}}) + \frac{8k^2}{\lambda n}$$

where the expected value is taken over size-n training sets.

Uniform Stability Analysis: An algorithm solving problem (1) is consistent

Definition 2 (Uniform stability [JWZ09]). *An algorithm has a uniform stability in $\epsilon(n)$ if $\forall i$,*

$$\sup_{\mathbf{z},\mathbf{z}'\sim\mathcal{D}_{\mathcal{T}}} \left| l(\mathbf{M}^*,\mathbf{z},\mathbf{z}') - l(\mathbf{M}^{i^*},\mathbf{z},\mathbf{z}')\right| \leq \epsilon(n)$$

where $\mathbf{M}^*$ is the matrix learned on the training set T and $\mathbf{M}^{i^*}$ is the matrix learned on the training set T^i obtained by replacing the i^{th} example of T by a new independent one.

Theorem 3 (Uniform stability). *Given a training sample T of n examples drawn i.i.d. from $\mathcal{D}_{\mathcal{T}}$, an algorithm solving optimization problem (1) has a uniform stability in $\epsilon(n) = \frac{4k^2}{\lambda n}$.*

Theorem 4 (Generalization bound). *With probability $1-\delta$, for any matrix $\mathbf{M}^*$ learned with an $\epsilon(n)$ uniformly stable algorithm and for any convex, k-lipschitz and (σ, m)-admissible loss, we have:*

$$L_{\mathcal{D}_{\mathcal{T}}}(\mathbf{M}^*) \leq L_T(\mathbf{M}^*) + (4\sigma + 2m + c)\sqrt{\frac{\ln\frac{2}{\delta}}{2n}} + \mathcal{O}\left(\frac{1}{n}\right)$$

where c is a constant linked to the k-lipschitz property of the loss and $\epsilon(n)$ appears in $\mathcal{O}\left(\frac{1}{n}\right)$.

Specific Loss Analysis

Theorem 5 (Generalization bound). *With probability $1-\delta$ for any matrix $\mathbf{M}^*$ learned by an algorithm solving optimization problem (1) with loss (2), we have:*

$$L_{\mathcal{D}_{\mathcal{T}}}(\mathbf{M}^*) \leq L_T(\mathbf{M}^*) + 4\left(\sqrt{\frac{L_T(\mathbf{M}_{\mathcal{S}})}{\lambda}} + \|\mathbf{M}_{\mathcal{S}}\|_{\mathcal{F}} + c_\gamma\right)\sqrt{\frac{\ln\frac{2}{\delta}}{2n}} + \mathcal{O}\left(\frac{1}{n}\right)$$

where c_γ is a constant linked to the k-lipschitz property of the loss and the chosen margins.

Optimizing the influence of the source by reweighting

Let $C(\mathbf{M}_{\mathcal{S}}) = \sqrt{\frac{L_T(\mathbf{M}_{\mathcal{S}})}{\lambda}} + \|\mathbf{M}_{\mathcal{S}}\|_{\mathcal{F}}$, let $\mathbf{M}_{\mathcal{S}} = \beta\mathbf{M}_{\mathbf{SOURCE}}$ we search β such that:

$$\beta^* = \operatorname*{arg\,min}_{\beta} C(\beta\mathbf{M}_{\mathbf{SOURCE}}) \qquad (3)$$

The goal is to minimize the right hand side of the bound, i.e. to choose the best matrix to transfer.

Experiments

Interest of **optimizing** β (UCI datasets):

	Baselines		Solving optimization problem (1) with loss (2)			
Dataset	1-NN	ITML	$\mathbf{M}_{\mathcal{S}} = \beta\mathbf{I}$	$\mathbf{M}_{\mathcal{S}} = \mathbf{I}$	$\mathbf{M}_{\mathcal{S}} = \beta\mathbf{\Sigma}^{-1}$	$\mathbf{M}_{\mathcal{S}} = \mathbf{\Sigma}^{-1}$
Breast	95.31 ± 1.11	95.40 ± 1.37	**96.06 ± 0.77**	95.75 ± 0.87	95.71 ± 0.84	94.76 ± 1.38
Pima	67.92 ± 1.95	68.13 ± 1.86	67.87 ± 1.57	67.54 ± 1.99	**68.37 ± 2.00**	66.31 ± 2.37
Scale	78.73 ± 1.69	**87.31 ± 2.35**	80.98 ± 1.51	80.82 ± 1.27	81.35 ± 1.17	80.88 ± 1.43
Wine	93.40 ± 2.70	93.82 ± 2.63	**95.42 ± 1.71**	95.07 ± 1.68	94.31 ± 2.01	80.56 ± 5.75

Application to a **transfer learning task** (Office-Caltech dataset):

	Baselines (using source examples)			Solving optimization problem (1) with loss (2) (using the source metric but no source examples)		
Task	1-NN$_{\mathcal{S}}$	MMDT	GFK	$\mathbf{M}_{\mathcal{S}} = \beta\mathbf{\Sigma}^{-1}$	$\mathbf{M}_{\mathcal{S}} = \beta\mathbf{M}_{\mathbf{ITML}}$	$\mathbf{M}_{\mathcal{S}} = \beta\mathbf{M}_{\mathbf{LMNN}}$
A → C	35.95 ± 1.30	**39.76 ± 2.25**	37.81 ± 1.85	32.65 ± 3.76	32.93 ± 4.60	34.66 ± 3.66
A → D	33.58 ± 4.37	54.25 ± 4.32	51.54 ± 3.55	54.69 ± 3.96	51.54 ± 4.03	**54.72 ± 5.00**
A → W	33.68 ± 3.60	64.91 ± 5.71	59.36 ± 4.30	67.11 ± 5.11	64.09 ± 5.20	**67.62 ± 5.18**
C → A	37.37 ± 2.95	**51.05 ± 3.38**	46.36 ± 2.94	50.15 ± 4.87	49.89 ± 5.25	50.36 ± 4.67
C → D	31.89 ± 5.77	52.80 ± 4.84	**58.07 ± 3.90**	56.77 ± 4.63	53.78 ± 7.23	57.44 ± 4.48
C → W	28.60 ± 6.13	62.75 ± 5.19	63.26 ± 5.89	64.64 ± 6.44	64.00 ± 6.08	**65.11 ± 5.25**
D → A	33.59 ± 1.77	**50.39 ± 3.40**	40.77 ± 2.55	49.48 ± 4.41	49.11 ± 4.09	49.67 ± 4.00
D → C	31.16 ± 1.19	**35.70 ± 3.25**	30.64 ± 1.98	32.90 ± 3.14	32.99 ± 3.58	33.84 ± 2.99
D → W	**76.92 ± 2.18**	74.43 ± 3.10	74.98 ± 2.89	65.57 ± 4.52	66.38 ± 6.04	69.72 ± 3.78
W → A	32.19 ± 3.04	50.56 ± 3.66	43.26 ± 2.34	50.80 ± 3.63	50.16 ± 4.32	**50.92 ± 4.00**
W → C	27.67 ± 2.58	**34.86 ± 3.62**	29.95 ± 3.05	31.54 ± 3.60	31.40 ± 4.29	32.64 ± 3.52
W → D	64.61 ± 4.30	62.52 ± 4.40	**71.93 ± 4.07**	57.17 ± 6.50	56.85 ± 5.51	61.14 ± 5.78
Mean	38.93 ± 3.26	**52.83 ± 3.93**	50.66 ± 3.28	51.12 ± 4.55	50.26 ± 5.02	52.32 ± 4.36