Sample Complexity of Learning Mahalanobis Distance Metrics

Nakul Verma **Kristin Branson**

Janelia Research Campus, HHMI

{verman, bransonk}@janelia.hhmi.org

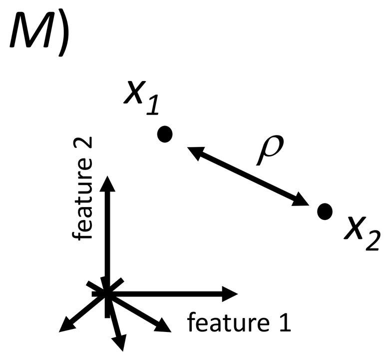
Want To Study

- ***Sample complexity*** rates for Metric Learning (ML).
- What key factors of input data determine the rate?
  - Representation dimension, noise levels, etc.
  - Are these factors *necessary*? (lower bounds)
- Is it possible to ***adapt*** the rates to the intrinsic complexity or information content of input data?
  - How do we *quantify* information content in data?
  - Is it possible to design algorithms that achieve error rates proportional to the information content *without* any a priori knowledge?
- How does the theory fare in practice?

Metric Learning

Mahalanobis distance metric (with weighting M)

$$\rho_M(x_1, x_2) = \|M(x_1 - x_2)\|^2$$



Goal: Learn M, that improves a prediction task

Distance Based Learning:
Given some distance based loss function:

$$\rho_M^{ij} = \rho_M(x_i, x_j) \qquad Y_{ij} = \mathbf{1}[y_i = y_j]$$

$$\phi(\rho_M^{ij}, Y_{ij}) = \begin{cases} (\rho_M^{ij} - U)_+ & \text{if } Y_{ij} = 1 \\ (L - \rho_M^{ij})_+ & \text{otherwise} \end{cases}$$

(e.g. losses MMC, ITML, LMNN)

Find M that *minimizes* the loss $\quad \mathrm{err}(M) := \mathbb{E}_{\substack{(x_i,y_i)\\(x_j,y_j)}} [\phi(\rho_M^{ij}, Y_{ij})]$

Classifier Based Learning:
Given a (real-valued) hypothesis class $\mathcal{H} := \{X \to [0,1]\}$

Find M that *minimizes* $\quad \mathrm{err}(M) := \inf_h \mathbb{E}_{(x,y)} [|h(Mx) - y| > 1/2]$

Sample Complexity

Statistical sample complexity of Metric Learning

$$M^* = \mathrm{argmin}_M \ \mathrm{err}(M) \qquad M_m^* = \mathrm{argmin}_M \ \mathrm{err}_{S_m}(M)$$

*(best **generalization** error)* *(best **sample** error with m samples)*

- at what rate does $\mathrm{err}(M_m^*) \to \mathrm{err}(M^*)$ as $m \to \infty$?
- what key factors affect the rate?
  (data dimension, noise levels, etc.)

What We Show

For data that resides in a D-dimensional feature space:

Upper Bounds

Th.1: For any λ-Lipschitz loss ϕ, and any sample size m,

$$\mathrm{err}(M_m^*) - \mathrm{err}(M^*) \le O\left(\lambda\sqrt{\frac{D\ln(1/\delta)}{m}}\right)$$

Distance based

Th.2: For any λ-Lipschitz hypoth. class, and any sample size m,

$$\mathrm{err}(M_m^*) - \mathrm{err}(M^*) \le O\left(\sqrt{\frac{(D^2 + \mathrm{Fat}_{\gamma/16}(\mathcal{H}))\ln(\lambda/\gamma\delta)}{m}}\right)$$

Classifier based

$\mathrm{Fat}_{\gamma/16}(\mathcal{H})$ *is the Fat-shattering dimension at margin* $\gamma/16$.

w.p. ≥ 1 - δ over the draw of m size sample

Lower Bounds

For *any* ML alg. A that minimizes sample error (on sample S_m).

Th.3: There exists a λ-Lipschitz loss function ϕ, s.t. for all ϵ, δ, if sample size $m \le O(D/\epsilon^2)$ then

$$P_{S_m}[\mathrm{err}(A(S_m)) - \mathrm{err}(M^*) > \epsilon] > \delta$$

Distance based

Th.4: There exists real-valued hypothesis class $\mathcal{H}$
if sample size $m \le O((D^2 + \mathrm{Fat}_{768\gamma}(\mathcal{H}))/(\epsilon^2 \ln 1/\gamma^2))$ then

$$P_{S_m}[\mathrm{err}(A(S_m)) - \mathrm{err}(M^*) > \epsilon] > \delta$$

Classifier based

what if most data has high representation dimension but low intrinsic complexity?

Quantifying Intrinsic Complexity

Observation: not all features are created equal.
(each features has a *different information content* for the prediction task)

Fix a prediction task T, and
let $\mathbf{M}$ the optimal feature weighting for T for a given dataset.
Define: ***metric learning complexity***

$$d^* := \|\mathbf{M}^\mathsf{T}\mathbf{M}\|_F^2$$

d is unknown a priori*

Question: Is it possible to achieve error rates that automatically adapt to d^*, without any prior knowledge about it?

Refined Rates

Th.5: For a prediction task T with (unknown) metric learning complexity d^*

$$\mathrm{err}(M_m^{\mathrm{reg}}) - \mathrm{err}(M^*) \le O\left(\sqrt{\frac{d^* \ln(D)\ln(1/\delta)}{m}}\right)$$

Distance & classifier based

where

$$M_m^{\mathrm{reg}} = \mathrm{argmin}_M \left[\mathrm{err}_{S_m}(M) + \Lambda\|M^\mathsf{T}M\|_F\right] \qquad \Lambda \approx \lambda\sqrt{\ln(D/\delta)/m}$$

w.p. ≥ 1 - δ over the draw of m size sample S_m

norm-regularization helps adapt to unknown intrinsic complexity of a given dataset in metric learning

Previous Theoretical Analysis

Distance based Learning (upper-bounds):

- (Jin et al. 2009) norm-regularized convex loss for stable algs.
- (Bian & Tao 2011) thresholds on bounded convex losses.
- (Cao et al. 2013) thresholds on hinge loss with norm reg.
- (Bellet & Habrard 2012) robust algs. with stable partitions.

Classifier based Learning (upper-bounds):

- (Balcan et al. 2008; Bellet et al. 2012) learn weighting metrics that best assist linear classifiers.

Experiments

Given a dataset with small metric learning complexity (d^*), but high representation dimension (D). How do regularized vs. unregularized Metric Learning algs. fare?

Approach

- pick benchmark datasets of low dimensionality (d)

UCI dataset	dim (d)
Iris	4
Wine	13
Ionosphere	34

- augment each dataset with large (D dim.) corr. noise

$$\Sigma_D \sim \mathrm{Wishart}(\text{unit-scale}) \qquad x_\sigma \sim N(0, \Sigma_D)$$

for each sample x_i, create augmented sample $x_i = [x_i\ x_\sigma]$

- study the prediction accuracy as a function of noise dim.