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?

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

Upper Bounds
Th.1: For any \( \lambda \)-Lipschitz loss \( \phi \), and any sample size \( m \),

\[
\text{err}(M^*_m) - \text{err}(M^*) \leq O\left(\frac{\sqrt{D \ln(1/\delta)}}{m}\right) \tag{Distance based}
\]

where

\[
M^*_m = \arg\min_M \left\{ \text{err}_S(M) + \lambda \|M^T M\|_F \right\} \quad \lambda \approx \rho \sqrt{\ln(D/\delta)/m}
\]

Th.2: For any \( \lambda \)-Lipschitz hypothesis class, and any sample size \( m \),

\[
\text{err}(M^*_m) - \text{err}(M^*) \leq O\left(\frac{(D^2 + \text{Fat},_{1/\delta}(\mathcal{H})) \ln(\lambda/\gamma)}{m}\rho\right) \tag{Classifier based}
\]

Lower Bounds
For any ML alg. \( A \) that minimizes sample error (on sample \( S_m \))

Th.3: There exists a \( \lambda \)-Lipschitz loss function \( \phi \), s.t. for all \( \varepsilon, \delta \), if sample size \( m \leq O(D^2/\varepsilon^2) \) then

\[
P_S(\text{err}(A(S_m)) > \varepsilon) \geq \delta \tag{Distance based}
\]

Th.4: There exists real-valued hypothesis class \( \mathcal{H} \), if sample size \( m \leq O(D^2 + \text{Fat},_{1/\delta}(\mathcal{H}))/\varepsilon^2 \ln(1/\gamma^2) \) then

\[
P_S(\text{err}(A(S_m)) > \varepsilon) \geq \delta \tag{Classifier based}
\]

Quantifying Intrinsic Complexity
Observation: not all features are created equal.
- (each feature has a different information content for the prediction task)

Fix a prediction task \( T \), and let \( M^* \) be the optimal feature weighting for \( T \) for a given dataset.
Define: metric learning complexity

\[
d^* := \|M^T M\|_F
\]

\(d^*\) 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 \( d^* \),

\[
\text{err}(M^*_m) - \text{err}(M^*) \leq O\left(\frac{d^* \ln(1/\delta)}{m}\right)
\]

where

\[
M^*_m = \arg\min_M \left\{ \text{err}_S(M) + \lambda \|M^T M\|_F \right\}
\]

with \( \lambda := \rho \sqrt{\ln(D/\delta)/m} \)

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

Previous Theoretical Analysis
Distance based Learning (upper-bounds):
- (Bian & Tao 2011) thresholds on bounded convex losses.
- (Cao et al. 2013) thresholds on hinge loss with norm reg.
- (Bellet & Habrard 2012) robust alg. 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 algos fare?

Approach
- pick benchmark datasets of low dimensionality \( (d) \)
- augment each dataset with large \( (D \text{dim.}) \) corr. noise \( \Sigma_D \sim \text{Wishart}(\text{unit-scale}) \) for each sample \( x \), create augmented sample \( y = [x, x_n] \)
- study the prediction accuracy as a function of noise dim.

<table>
<thead>
<tr>
<th>UC1 dataset</th>
<th>dim (d)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Iris</td>
<td>4</td>
</tr>
<tr>
<td>Wine</td>
<td>13</td>
</tr>
<tr>
<td>Ionosphere</td>
<td>34</td>
</tr>
</tbody>
</table>

Statistical sample complexity of Metric Learning

\[
M^* = \arg\min_M \text{err}(M) \quad M^*_m = \arg\min_M \text{err}_S(M) \quad \text{(best generalization error)}
\]

\[
M^*_m = \arg\min_M \text{err}_S(M) \quad \text{(best sample error with m samples)}
\]

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

Find that \( \text{err}(M^*_m) \) minimizes the loss \( \text{err}(M) := \mathbb{E}_{(x_i, y_i) \in (S_m)} [\phi(\rho_{ij}^*, Y_{ij})] \)

Find M that minimizes \( \text{err}(M) := \inf_{\mathbb{E}_{(x_i, y_i) \in (S_m)}} [\phi(\rho_{ij}^*, Y_{ij})] \quad \text{(each features has a different information content for the prediction task)} \)

Quantifying Intrinsic Complexity
Observation: not all features are created equal.
- (each feature has a different information content for the prediction task)

Fix a prediction task \( T \), and let \( M \) be the optimal feature weighting for \( T \) for a given dataset.
Define: metric learning complexity

\[
d^* := \|M^T M\|_F
\]

\(d^*\) is unknown a priori

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