Partial order embedding with multiple kernels

Brian McFee and Gert Lanckriet

University of California, San Diego
Goal

Embed a set of objects into a Euclidean space such that:

1. Distances conform to human perception
2. Multiple feature modalities are integrated coherently
3. We can extend to unseen data

Motivation: leverage existing technologies for Euclidean data
Example

Musicians

Target space
Example

Musicians

tags, acoustics, social data, ...

• Features may not match human perception
Example

Musicians

tags, acoustics, social data, ...

Target space

• Features may not match human perception
• Use human input to guide the embedding
Human input

• Binary similarity can be *ambiguous* in multi-media data
• Example:

  Is *Oasis* similar to *The Beatles*, or not?

• *Quantifying* similarity may also be difficult... how similar are they?
Relative comparisons

[Schultz and Joachims, 2004, Agarwal et al., 2007]

- Instead, we ask which of two pairs is more similar:

  \[(i, j) \text{ or } (k, \ell)\]?

(Oasis, Beatles, Oasis, Metallica)

- Learn a map \(g\) from the data set \(\mathcal{X}\) to a Euclidean space

- For each \((i, j, k, \ell)\),

\[
\|g(i) - g(j)\| < \|g(k) - g(\ell)\|
\]
Partial order

More similar

\[ \begin{align*}
ij & \rightarrow jk \\
ik & \rightarrow kl \\
il & \rightarrow ik \\
\end{align*} \]

- Relative comparisons should exhibit \textit{global structure}.
- Collect comparisons into a directed graph $\mathcal{C}$
- \textit{Cycles} must be broken by any embedding
  - Comparisons should describe a partial order over $\mathcal{X} \times \mathcal{X}$. 

Less similar
Constraint graphs

• Force margins between distances:
  
  \[ \| g(i) - g(j) \|^2 + e_{ijk\ell} \leq \| g(k) - g(\ell) \|^2 \]

• Represent \( e_{ijk\ell} \) as edge weights

• Graph representation lets us
  • detect inconsistencies (\textit{cycles})
  • prune redundancies by \textit{transitive reduction}
  • simplify: focus on meaningful constraints
Constraint simplification
Constraint simplification
Margin-preserving embeddings

- **Claim**: There exists \( g : \mathcal{X} \to \mathbb{R}^{n-1} \) such that all margins are preserved, and for all \( i \neq j \):

\[
1 \leq \|g(i) - g(j)\| \leq \sqrt{(4n + 1)(\text{diam}(C) + 1)}
\]

- Reduction via constant-shift embedding [Roth et al., 2003]
- *Constraint diameter* bounds *embedding diameter*
- May produce artificially high-dimensional embeddings
Dimensionality reduction

- We show that it’s **NP-hard** to minimize dimensionality for POE

- Instead, optimize a convex objective that prefers low-dimensional solutions

- Assume objects are **dissimilar**, unless otherwise informed

- Adapt MVU [Weinberger et al., 2004]:
  - Maximize all distances
  - Diameter bound ensures that a solution exists
  - Respect all partial order constraints
Partial Order Embedding (SDP)

- **Input:** \( n \) objects \( \mathcal{X} \), margin-weighted constraints \( \mathcal{C} \)
- **Output:** \( g : \mathcal{X} \rightarrow \mathbb{R}^n \)

\[
\max_{A \succeq 0} \quad \text{Tr}(A)
\]

\[
\begin{align*}
&d(i, j) \leq O(n \cdot \text{diam}(\mathcal{C})) & \text{(Variance)} \\
&d(i, j) + e_{ijk\ell} \leq d(k, \ell) & \text{(Diameter)} \\
&\sum_{i,j} A_{ij} = 0 & \text{(Centering)} \\
&d(i, j) = A_{ii} + A_{jj} - 2A_{ij} & \text{(Distance\(^2\))}
\end{align*}
\]

- Decompose \( A = V \Lambda V^T \) \( \Rightarrow \) \( g(i) = (\Lambda^{1/2} V^T)_i \)
Out-of-sample extension: kernels

• How can we extend embeddings to unseen data?
• Learn a linear projection from a feature space

Parameterization:

\[ g(x) = NK_x \]

\( (K_x = x \text{ column of } K) \)

• Learn \( N \) by solving an SDP over \( W = N^TN \succeq 0 \)

• PO constraints may be impossible to satisfy:
  • Soften ordering constraints
Multi-kernel embedding

• Concatenate linear projections from \( m \) feature spaces:

\[ g(x) = \begin{bmatrix} N^{(1)} K_x^{(1)} \\ N^{(2)} K_x^{(2)} \\ \vdots \\ N^{(m)} K_x^{(m)} \end{bmatrix} \]

• \( N^{(\cdot)} \)'s are jointly optimized by SDP to form the space
MK-POE

\[
\max_{w \geq 0, \xi \geq 0} \quad \sum_{p=1}^{m} \text{Tr} \left( K^{(p)} W^{(p)} K^{(p)} \right) - \gamma \text{Tr} \left( W^{(p)} K^{(p)} \right) - \beta \sum_{c} \xi_{ijk\ell}
\]

s.t.
\[
\forall i, j \in X \quad d(i, j) \leq O(n \cdot \text{diam}(C))
\]
\[
\forall (i, j, k, \ell) \in C \quad d(i, j) + e_{ijk\ell} \leq d(k, \ell) + \xi_{ijk\ell}
\]

\[
d(i,j) \doteq \sum_{p=1}^{m} \left( K_i^{(p)} - K_j^{(p)} \right)^{T} W^{(p)} \left( K_i^{(p)} - K_j^{(p)} \right)
\]
Experiment 1: Human perception

Data [Agarwal et al., 2007]
- 55 images of 3D rabbits with varying surface reflectance
- 13049 human perception measurements: \((i, j, i, k)\)

Constraint processing
- Random sampling to achieve a maximal DAG
- Transitive reduction to eliminate redundancies
  \(13000 \rightarrow 9000\) constraints

Final constraint graph
- Unit margins
- Diameter = 55
Experiment 1 results

POE (Top 2 PCA)

Glare

Luminance
Experiment 2: Multi-kernel

Data [Geusebroek et al., 2005]

- 10 classes from ALOI
- 10 images from each class, varying out-of-plane rotation
- Constraints generated by a label taxonomy

Kernels

- Grayscale dot product
- RBF of R,G,B, and grayscale histograms

Diagonally-constrained $N$: SDP $\Rightarrow$ LP
Experiment 2 results

Sum-kernel space

Learned embedding

- Shoe
- Hat
- White shoe
- Lemon
- Pear
- Orange

Xmas bear
Pink animal
R/Y block
Smurf

Learned weights

Dot
Red
Green
Blue
Gray

Training set
## Experiment 2 kernel comparison

<table>
<thead>
<tr>
<th>Kernel</th>
<th>Native</th>
<th>Optimized</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dot product</td>
<td>0.83</td>
<td>0.85</td>
</tr>
<tr>
<td>Red</td>
<td>0.63</td>
<td>0.63</td>
</tr>
<tr>
<td>Green</td>
<td>0.65</td>
<td>0.67</td>
</tr>
<tr>
<td>Blue</td>
<td>0.77</td>
<td>0.83</td>
</tr>
<tr>
<td>Gray</td>
<td>0.68</td>
<td>0.69</td>
</tr>
<tr>
<td>Unweighted sum</td>
<td>0.76</td>
<td>0.77</td>
</tr>
<tr>
<td>Multi</td>
<td>—</td>
<td>0.95</td>
</tr>
</tbody>
</table>

% Constraints satisfied
Experiment 3: Out-of-sample

Goal
- Predict comparisons \((i, j, i, k)\) with \(i\) out of sample

Data
- 412 popular artists (\textit{aset400}) [Ellis et al., 2002]
- 10-fold cross-validation
- \(\approx 6300\) human-derived training constraints
- Mean diameter \(\approx 30\) (over CV folds)

Features: TFIDF/cosine kernels
- \textit{Tags}: 7737 words (e.g., \textit{rock}, \textit{piano}, \textit{female vocals})
- \textit{Biographies}: 16753 words
Experiment 3 results

Prediction accuracy

<table>
<thead>
<tr>
<th>Tags</th>
<th>Native</th>
<th>Optimized</th>
<th>Biography</th>
<th>Native</th>
<th>Optimized</th>
<th>Tags+Bio</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.705</td>
<td>0.776</td>
<td>0.705</td>
<td>0.514</td>
<td>0.705</td>
<td>0.640</td>
</tr>
<tr>
<td></td>
<td>0.790</td>
<td></td>
<td></td>
<td>0.705</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: test comparisons are not internally consistent
Conclusion

• We developed the **partial order** embedding framework
  • Simplifies relative comparison embeddings
  • Enables more careful constraint processing
    • Graph manipulations can increase embedding robustness

• Derived a novel **multiple kernel learning** technique
  • Widely applicable to metric learning problems
Thanks!

Questions?


