Applied Machine Learning

Principal Coordinate Analysis

Principal Coordinate Analysis

- Overview of PCoA
- Distances in a high-dimensional dataset
- Low-rank approximation of the high-dimensional dataset
- Summary

Principal Coordinate Analysis - PCoA

- Transform high-dimensional dataset into low-dimensional representation
- Suitable to understand relationships
  - visualizations
  - spot blobs in original dataset
  - spot emerging patterns
- common choices of dimensions: $r \in [2,3]$
- Aim to preserve ratios of distances

Distances in High-Dimensional Set

- Dataset $\{X\}$, mean$(\{X\}) = 0$
- Squared distance between items within set
  $D_{i,j}(\mathbf{x}) = (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^\top$
  $$D(X) = \begin{bmatrix} D_{1,1} & D_{1,2} & \dots & D_{1,N} \\ D_{2,1} & D_{2,2} & \dots & D_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ D_{N,1} & D_{N,2} & \dots & D_{N,N} \end{bmatrix}$$
- $\mathbf{x}_i \in \{X\} \mapsto \mathbf{y}_i \in \{Y\}$
  - minimize pairwise distances:
    - $\sum_{i,j} (D_{i,j}(\mathbf{x}) - D_{i,j}(\mathbf{y}))^2$

- $A = \left(I_{\{N\times N\}} - \frac{1}{N}\mathbf{1}_{\{N\times 1\}}\mathbf{1}^\top_{\{N\times 1\}} \right)$
  - $-\frac{1}{2}AD(X)A^\top = XX^\top$
- $D(Y)$ close to $D(X)$ by
  - making $YY^\top$ close to $XX^\top$
  - through low rank approximation of $XX^\top$

Low-Rank Approximation

- $XX^\top$
  - SVD of $X = USV^\top$
- Full rank SVD of $XX^\top$

$$
\begin{aligned}
XX^\top &= (USV^\top)(USV^\top)^\top \\
&= USV^\top(V^\top)^\top S^\top U^\top \\
&= USS^\top U^\top \\
&= US^2U^\top
\end{aligned}
$$

-

- The low rank SVD of $X_rX_r^\top$

$$
\begin{aligned}
X_rX_r^\top &= (U_rS_rV_r^\top)(U_rS_rV_r^\top)^\top \\
&= U_rS_rV_r^\top(V_r^\top)^\top S_r^\top U_r^\top \\
&= U_rS_rS_r^\top U_r^\top \\
&= U_rS_r(U_rS_r)^\top \\
&= YY^\top
\end{aligned}
$$

-

- $Y = U_rS_r$

Principal Coordinate Analysis

- Low-dimensional representation of dataset $\{X\}$, mean$(\{X\}) = 0$

$$XX^\top = -\frac{1}{2}AD(x)A^\top$$

- $$W = -\frac{1}{2}AD(x)A^\top$$
  - $A = \left(I_{\{N\times N\}} - \frac{1}{N}\mathbf{1}_{\{N\times 1\}}\mathbf{1}^\top_{\{N\times 1\}}\right)$
  - we only need $D(X)$ even if X is not known directly
- W: eigenvectors U, eigenvalues Λ
  - $WU = U\Lambda$

- U_r: first r columns of U
- S_r from upper left submatrix of S
  - $S = \Lambda^{\frac{1}{2}}$, from $XX^\top = US^2U^\top$
- $Y = U_rS_r$

Principal Coordinate Analysis - PCoA

- Overview of PCoA
- Distances in a high-dimensional dataset
- Low-rank approximation of the high-dimensional dataset
- Summary

Applied Machine Learning

Principal Coordinate Analysis