Introduction to Machine Learning
Homework 3

Lu Hongyu Student ID: 2023202269

2025.4.30

Question 1

Answer

Algorithm

Motivated by *Robust Principal Component Analysis* (RPCA), I turn the original optimal function for *Robust NMF* into this form:

$$\min_{U,V,S} \frac{1}{2} \left\| X_{\text{noisy}} - UV^\top - S \right\|_F^2,$$

$$\text{s.t. } \|S\|_0 \le ND\rho_{\text{nz}},\ U \in [0,\infty)^{N\times r},\ V \in [0,\infty)^{D\times r},$$

The problem is tackled by alternating between a classical multiplicative-update NMF subproblem (fixing S) and a hard-thresholding step that refines S (fixing U, V). The complete procedure is summarised below.

Algorithm 1 Non-negative Matrix Factorisation (NMF)

Require: non-negative matrix $X \in \mathbb{R}^{N\times D}$; target rank r; iteration number `num_iter`; seed; small constant $\varepsilon > 0$

Ensure: $U \in \mathbb{R}_{\ge 0}^{N\times r}$, $V \in \mathbb{R}_{\ge 0}^{D\times r}$, low-rank estimate $\hat{L} = UV^\top$

1: Initialise U, V with i.i.d. non-negative random numbers using *seed*.
2: **for** $t = 1$ **to** `num_iter` **do**
3: $\quad V \leftarrow V \odot \dfrac{X^\top U}{V(U^\top U) + \varepsilon}$
4: $\quad U \leftarrow U \odot \dfrac{XV}{U(V^\top V) + \varepsilon}$
5: **end for**
6: **return** U, V, $\hat{L} = UV^\top$

Algorithm 2 Robust Non-negative Matrix Factorisation (Robust NMF)

Require: matrix $X \in \mathbb{R}^{N\times D}$; rank r; inner NMF iterations `num_iter`; outer alternations `n_alt`; sparsity ratio ρ_{nz}; seed; $\varepsilon > 0$

Ensure: U, V, low-rank part $\hat{L}$, sparse part $\hat{S}$

1: Initialise $S \leftarrow 0$
2: $(U, V, \hat{L}) \leftarrow \text{NMF}(\max(X, 0), r, \texttt{num\_iter}, \textit{seed}, \varepsilon)$
3: **for** $k = 1$ **to** `n_alt` **do**
4: $\quad X_{\text{target}} \leftarrow X - S$
5: $\quad (U, V, \hat{L}) \leftarrow \text{NMF}(X_{\text{target}}, r, \texttt{num\_iter}, \textit{seed} + k, \varepsilon)$
6: $\quad R \leftarrow X - \hat{L}$ ▷ current residual
7: $\quad S \leftarrow \textsc{HardThreshold}(R, \rho_{\text{nz}})$
8: **end for**
9: **return** U, V, $\hat{L}$, $\hat{S} = S$

Explanation

- **Low-rank step (Lines 3–6).** With the current sparse estimate fixed, we perform standard NMF (Alg. 1) on the residue $X - S$, yielding an updated non-negative low-rank factorisation $UV^\top$.
- **Sparse step (Lines 7–8).** The new residue $R = X - UV^\top$ is hard-thresholded: only the largest magnitude entries whose count does not exceed $\rho_{\text{nz}}ND$ are kept, producing an updated sparse matrix S. This step is equivalent to an ℓ_0 projection and is inexpensive.
- **Alternation.** Repeating these two steps refines U, V and S simultaneously.

Question 2

Answer

- **ISOMAP**

$$\Phi = -\tfrac{1}{2}C(D^{\text{geo}} \circ D^{\text{geo}})C, \quad \Omega = \{Z \mid Z^\top Z = I\},$$

  where D^{geo} is the matrix of graph-based geodesic distances and $C = I - \frac{1}{N}\mathbf{1}\mathbf{1}^\top$ is the centering matrix.

  Proof: ISOMAP minimizes the classical MDS loss:

$$\|ZZ^\top - K\|_F^2, \quad \text{with } K = -\tfrac{1}{2}C(D^{\text{geo}} \circ D^{\text{geo}})C.$$

  Expanding gives:

$$\text{tr}(ZZ^\top ZZ^\top) - 2\text{tr}(Z^\top KZ) + \text{tr}(K^2),$$

  and minimizing reduces to minimizing $\text{tr}(Z^\top(-K)Z)$, since the first and last terms are constants under the orthonormality constraint.

- **Locally Linear Embedding (LLE)**

$$\Phi = (I - W)^\top (I - W), \quad \Omega = \{Z \mid Z^\top Z = I,\, Z^\top \mathbf{1} = 0\},$$

where W is the sparse weight matrix learned from local linear reconstruction.

Proof: The objective function is:

$$\|Z - WZ\|_F^2 = \mathrm{tr}[(Z - WZ)^\top (Z - WZ)] = \mathrm{tr}(Z^\top (I - W)^\top (I - W)Z).$$

The constraint $Z^\top \mathbf{1} = 0$ removes the trivial translation mode.

- **Kernel PCA**

$$\Phi = -K, \quad \Omega = \{Z \mid Z^\top Z = I\},$$

where K is the centered Gram matrix in the RKHS induced by the kernel.

Proof: Kernel PCA maximizes variance in feature space:

$$\max_{Z^\top Z = I} \mathrm{tr}(Z^\top K Z) \quad \Longleftrightarrow \quad \min_{Z^\top Z = I} \mathrm{tr}(Z^\top (-K) Z).$$

Hence, it fits the trace-minimization framework with $\Phi = -K$.

- **Laplacian Eigenmap**

$$\Phi = L = D - A, \quad \Omega = \{Z \mid Z^\top D Z = I,\, Z^\top D \mathbf{1} = 0\},$$

where A is the adjacency matrix and $D = \mathrm{diag}(A\mathbf{1})$ is the degree matrix.

Proof: The objective function is:

$$\sum_{i,j} A_{ij} \|z_i - z_j\|^2 = \mathrm{tr}(Z^\top L Z),$$

because:

$$\sum_{i,j} A_{ij} \|z_i - z_j\|^2 = \sum_{i,j} A_{ij} (z_i^\top z_i - 2 z_i^\top z_j + z_j^\top z_j) = 2\,\mathrm{tr}(Z^\top L Z).$$

The generalized constraints avoid degenerate solutions and enforce orthogonality in the graph metric.