Data Visualiztion

Laboratory works №9

Dimension reduction with SVD

9.1 Singular Value Decomposition (SVD)

Singular value decomposition (SVD) is a fundamental technique in linear algebra with various applications, including dimensionality reduction.

SVD decomposes a real or complex matrix (A) of dimension (m x n) into three matrices:

- U (m x m): An orthogonal matrix containing the left singular vectors of A.
- Σ (m x n): A diagonal matrix containing the singular values of A on its diagonal, arranged in non-increasing order.
- V^T (n x n): The transpose of another orthogonal matrix V (n x n) containing the right singular vectors of A.

The decomposition is expressed as:

$$A = U \Sigma V^T. \quad (1)$$

The singular values on the diagonal of Σ represent the importance or magnitude of the corresponding singular vectors in U and V.

Larger singular values indicate directions of greater variance in the data captured by A.

Dimensionality reduction with SVD involves selecting a subset of the top k singular values (where k < min(m, n)) and truncating Σ and the corresponding rows/columns of U and V^T. This retains the most significant information from the original data in a lower-dimensional space.

By keeping only the top k singular values and their corresponding vectors, we obtain a lower-dimensional approximation of A:

$$A_k \approx U_k \Sigma_k V_k^T \quad (2)$$

where:

A_k is the reduced-dimensional representation of A.

U_k consists of the first k columns of U.

Σ_k is a diagonal matrix containing the top k singular values.

V_k^T consists of the first k rows of V^T.

Geometric Interpretation of SVD:

In geometric terms, SVD can be viewed as rotating the data (represented by A) such that the principal axes (directions of greatest variance) align with the columns of U. The singular values then determine the scaling along these axes. By keeping only the top k singular values, we project the data onto the subspace spanned by the corresponding principal axes, effectively reducing dimensionality.

The SVD method is implemented by numpy library
https://numpy.org/doc/stable/reference/generated/numpy.linalg.svd.html

Variants of tasks

For the corresponding dataset, according to the option, reduce the dimensionality of the data using PCA and SVD. Datasets are placed in the datasets folder (https://github.com/a-vodka/dv/tree/master/lab/dataset).

1. Using PCA to visualize data in two- and three-dimensional (2D and 3D) spaces. Use PCA class for sklearn library https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html
2. Calculate SVD of your dataset, plot the dependence of the eigenvalues of the matrix on their number. Before plotting, arrange the eigenvalues in descending order.
3. Determine the smallest value of the space size i for which relation (3) is satisfied. Where λ_i are the eigenvalues of the matrix, n is the total number of eigenvalues, 0.8 – is level of data significance.

$$\frac{\sum_{i=0}^{d} \lambda_i}{\sum_{i=0}^{n} \lambda_i} \leq 0.8 \qquad (1)$$

4. Set λ_i to zero for which $d \leq i \leq n$. Perform the reverse transformation and compare the obtained data with the original.
5. Set d = 2 (for 2D) and d = 3 (3D) and perform and plot first d columns of reconstructed data. Compare graph with obtained on step 1.