Lab Course Machine Learning

Exercise Sheet 9

Prof. Dr. Dr. Lars Schmidt-Thieme
Jung Min Choi
HiWi: Harish Malik

Submission deadline : January 5, 2025

1 Data Imputation using PCA/Probabilstic PCA (15 points)

You will implement data imputation using Principal Component Analysis (PCA) and Probabilistic PCA (PPCA).

1. **[5 Points]** Implement PCA for data imputation from scratch without using pre-built libraries (e.g., `sklearn`). Use the following steps:
   - Replace missing values with the column mean (mean value imputation).
   - Perform low-rank approximation using Singular Value Decomposition (SVD from numpy.linalg.svd).
   - Use the low-rank approximation to reconstruct the dataset with imputed values.

   Low-rank approximation:
   $$X_{\text{approx}} = U_d \Sigma_d V_d^T,$$
   where U_d, Σ_d, V_d are the truncated matrices obtained from SVD.
2. **[6 Points]** Implement Probabilistic PCA (PPCA) for data imputation using the Coordinate Descent Expectation-Maximization (EM) algorithm.
3. **[1 Point]** Compute the Mean Squared Error (MSE) between the original data and the imputed data for both PCA and PPCA methods. Report the results.
4. **[3 Points]** Visualize the original data and the imputed data (from PCA and PPCA) separately in the space of the first two principal components. Use scatter plots for the visualization, and clearly label each plot.

2 Dimensionality Reduction using Kernel PCA (5 points)

You will implement Kernel PCA from scratch to perform non-linear dimensionality reduction. Kernel PCA extends PCA by using a kernel function to implicitly map the data into a higher-dimensional space where linear methods can capture non-linear relationships.

1. **[1 Point]** Use a synthetic dataset with non-linear structure and visualize the data.
   ```
   from sklearn.datasets import make_circles
   ```
2. **[3 Points]** Implement Kernel PCA:
   - Define the Radial Basis Function (RBF) kernel to compute pairwise similarities:
     $$k(x, y) = \exp\left(-\frac{\|x - y\|^2}{2\sigma^2}\right).$$
     Experiment with different values of the hyperparameter σ.
   - Center the kernel matrix K to ensure it has zero mean.
     $$K' = HKH, \quad H = I - \frac{1}{N}\mathbf{1},$$
     where $\mathbf{1}$ is an $N \times N$ matrix of ones.

- Perform eigenvalue decomposition on the centered kernel matrix K'. Extract the top d eigenvalues and corresponding eigenvectors.
- Project the data into the reduced-dimensional space.

$$\text{Projections} = K' \cdot \text{Eigenvectors}.$$

3. **[1 Point]** Create a scatter plot to visualize the transformed dataset in the 2D space of the first two components. Color the points by their class labels and write observations on the effectiveness of Kernel PCA for non-linear data.