Machine Learning: Principal Component Analysis

1 Introduction

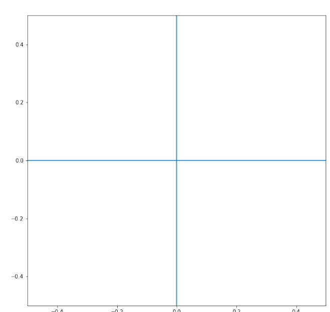
Python libraries required to complete the assignments:

- NumPy
- Matplotlib

2 PCA

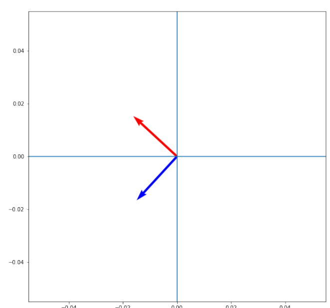
Assignments

1. All figures should be squared (figsize = (10,10), figsize = (5,5), ...).
2. Using NumPy define matrix:
$$X = \begin{bmatrix} 2.5 & 0.5 & 2.2 & 1.9 & 3.1 & 2.3 & 2 & 1 & 1.5 & 1.1 \\ 2.4 & 0.7 & 2.9 & 2.2 & 3 & 2.7 & 1.6 & 1.1 & 1.6 & 0.9 \end{bmatrix}^T$$
3. Number of columns means number of features and number of rows means number of observations.
4. Plot data of X (first column is the x-axis, second is y-axis).
5. Center data points (mean subtraction) and plot centered data with coordinate system lines (below, axvline, axhline from Matplotlib).



6. Calculate the covariance matrix (covariances of features) of the centered data (this matrix should be 2×2).
7. Calculate the eigendecomposition of the covariance matrix (assign eigenvalues and eigenvectors to two variables, for example, D - eigenvalues, V-eigenvectors).
8. Eigenvectors are a matrix and each column corresponds to each eigenvalue (the first eigenvalues correspond to the first column of eigenvector matrix).
9. Make a similar plot to the plot from point 5 (with data points, of course) and with eigenvectors. For plotting eigenvectors, use the following formulas:

```
plt.quiver(V[0,0],V[1,0], color=['r'], scale=5)
plt.quiver(V[0,0],V[1,0], color=['r'], scale=5)
```



10. Assign the eigenvector that corresponds to the largest eigenvalue to the new variable u and the second eigenvector to the variable z. Re-assign u/z as follows to make the matrices from 1D vectors:

```
u = np.expand_dims(u,1)
z = np.expand_dims(z,1)
```

11. Perform the following projections: $Y = X \times u \times u^T$ (projections using the eigenvector corresponding to the largest eigenvalue).
12. Plot projected data similar to the plot at point 9 with coordinate system lines and with eigenvectors.
13. Repeat points 11-12 with projection using the second eigenvector.
14. Add five plots from points 4, 5, 9, 12, and 13 to report.
15. In the report, explain why we project the data using the eigenvector, which corresponds to the largest eigenvalue, write out covariance matrix, eigenvalues, and eigenvectors, describe in which situations PCA is useful in machine learning and what are the pros/cons of PCA and what are the alternatives of PCA.