Applied Data Science

COMS30050/COMS30051/COMSM0055/COMSM0056 (2020 TB-2)

Lab 4: Data Exploration

Week #4 Lab builds on the lectures on data fusion and data exploration and involves exploring the Iris dataset and the MNIST dataset.

Exploratory tasks

- Compute the descriptive statistics for the Iris and the MNIST datasets
    - Central tendency
        - Mean
        - Median
        - Mode
    - Variability
        - Variance
        - Quartiles
        - Max and Min
        - Test the normality of the data
            - Kurtosis and Skewness
                - https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.skew.html#scipy.stats.skew
                - https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.kurtosis.html#scipy.stats.kurtosis
        - More here: https://docs.scipy.org/doc/scipy/reference/stats.html
- In the Iris dataset are there any outliers?
    - Compute the quartiles (Q1, Q2, Q3, and Q4)
    - Compute the Inter Quartile Range (IQR)
        - IQR = Q3 – Q1
    - Are there any values less than Q1-1.5IQR or greater than Q3+1.5IQR?
- Apply dimension reduction algorithm and compare the outputs. Vary the default parameters and observe the changes in the output.
    - PCA
        - https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html
    - t-SNE
        - https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html
    - UMAP
        - https://umap.scikit-tda.org/basic_usage.html
    - Random projections
        - https://scikit-learn.org/stable/modules/random_projection.html

```
from sklearn import random_projection
```

```
min_dim =
random_projection.johnson_lindenstrauss_min_dim(n_samples=X_train.shape[
0], eps=0.9)

grp = random_projection.GaussianRandomProjection(n_components=2)

X_new = grp.fit_transform(Xs)
```

- Additional task: Could we identify clusters automatically?
    - Prediction using clustering
        - K-mean
            - https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html