Dimensionality Reduction PCA

ENSF 444

Review Unsupervised Learning

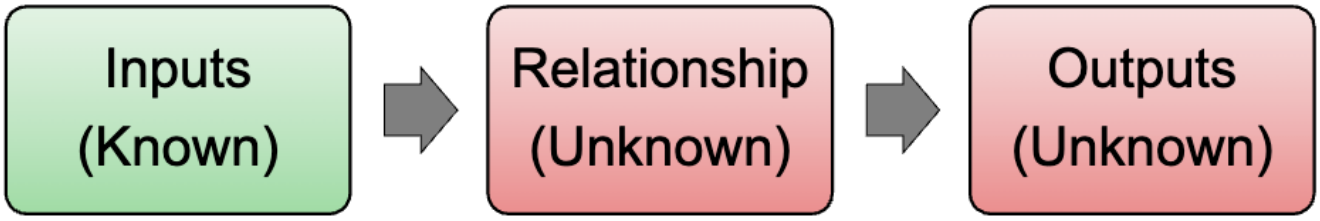
Discuss how this can be applied to reduce the number of columns in the dataset using PCA

Chapter 3.4.1 Machine learning section

Chapter 5 in the Python Data Science Handbook

Up to this point we have been using labeled datasets, now if we dont know the label for each datapoint, we can use **unsupervised learning** to gain insight on different properties of the data

2 Main Applications used is Dimensionality Reduction and Clustering.



Above is unsupervised learning

Dimensionality Reduction

Aims to reduce the number of variables in a dataset while keeping the important information.

2 types of DR(Dimensionality Reduction)

1. **Feature Selection**: Selecting a subset of relevant features and discarding the rest. Requires expertise in the domain and can be done manually or automatically using

statistical or machine learning techniques. This can seen in Recursive Feature Elimination(RFE)
2. **Feature Extraction**: Transforms original features into a new set of features with reduced dimensions using mathematical method. Principal Component Analysis

---

Benefits of Reducing Dimensions in Data

1. **Easier data understanding**: High dimensional data can be overwhelming.
2. **Improved Algorithm Speed**: High number of features can slow down algorithms and increase the need for computational resources.
3. R**eduction of Irrelevant information**: Often High dimensional data includes unnecessary information that can negatively affect the models performance.
4. **Better Data Visualization**: Difficult to visualize data from beyond 3 dimensions

---

Key Points and Hurdles in Reducing Data Dimensions

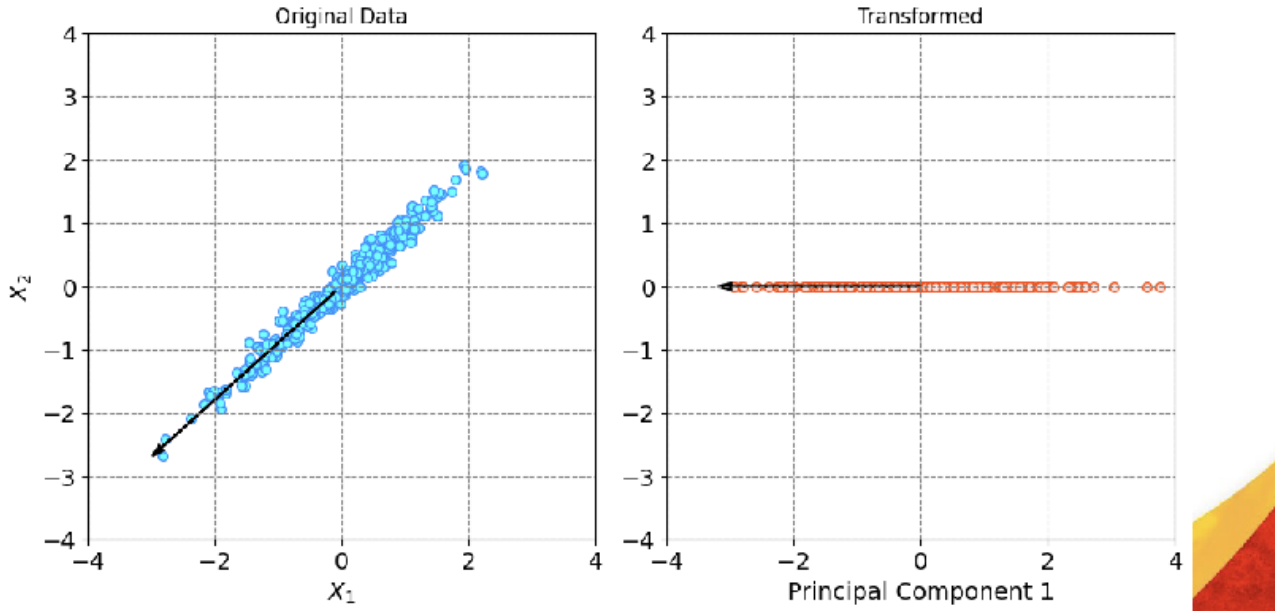
1. **Balancing Complexity and Information Retention**: Balance between simplifying data and keeping valuable information. Important to assess if the information lost is good for the task's goals
2. **Technique Selection**: Choosing the right dimensionality reduction method depends on the dataset
3. **Hyperparameter tuning**: Some methods like PAC require setting hyperparameters, like the. number of components to keep.
4. **Overcoming the curse of Dimensionality**: refers to a phenomena when analyzing high dim data, like increased data sparsity
5. **Preventing Overfitting**: Dimensionality Reduction can help prevent methods from overfitting, but theres a risk of overfitting the reduction process. BY using regularization methods, you can prevent the reduction method from overfitting.

Principal Component Analysis

PCA transforms the original variables into a new set of variables

Core Concepts:

• **Principal Axes**: These are the directions in the feature space that maximize the variance of the data. The data is projected onto these axes to obtain the principal components.
• **Principal Components (Scores)**: These are the new features formed from linear combinations of the original features, aligned with the principal axes.
• **Loadings**: The weights assigned to the original variables that define the principal axes.
• **Explained Variance**: The amount of variance captured by each principal component.
• **Explained Variance Ratio**: The proportion of the dataset's total variance that is explained by each principal component.

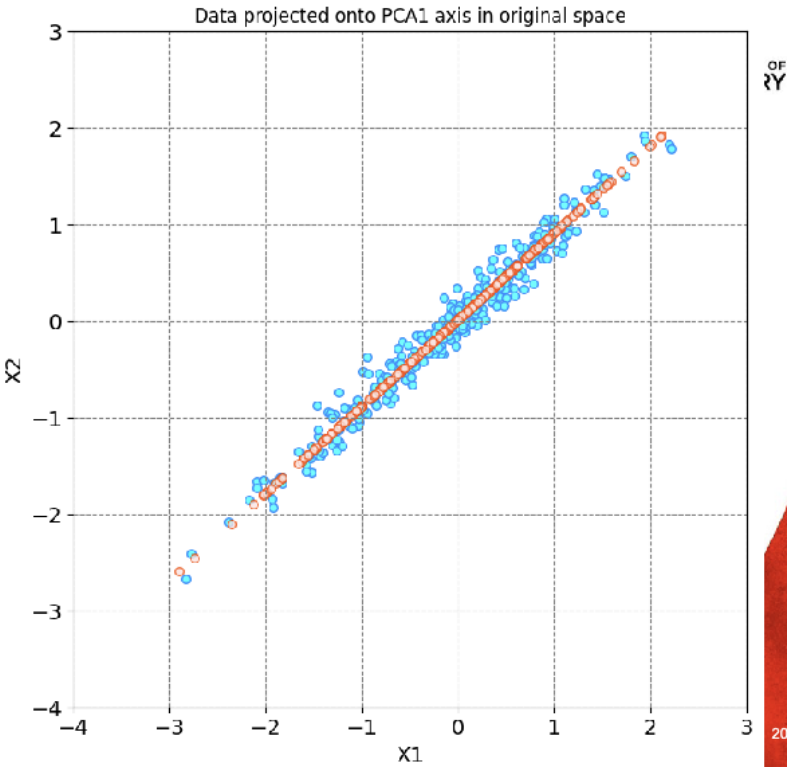


How data can be transformed

Example – PCA (10/)

- The **blue points** represent the original data in its two-dimensional feature space.
- The **orange points** represent the data after it has been transformed to one dimension using PCA and then inversely transformed back into the two-dimensional space.

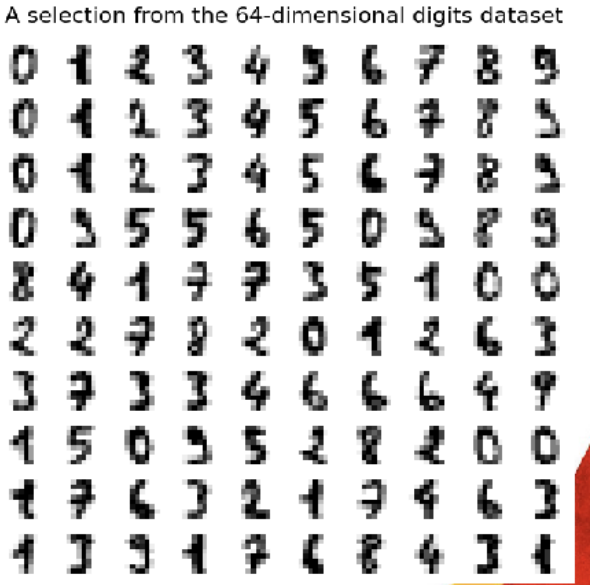
The inverse transformation maps the one-dimensional PCA data back onto the original feature space, but since PCA reduces dimensionality by projecting the data onto the direction of maximum variance (the first principal component), the orange points lie along a line. This line represents the first principal component axis in the original feature space.



Then there is an example of for Optical Recognition of handwritten digits

Data Set Characteristics:

- Number of Instances: 1797
- Number of Attributes: 64
- Attribute Information: 8x8 image of integer pixels in the range 0 - 16.
- Missing Attribute Values: None



Reducing the dims to from 64 to 2