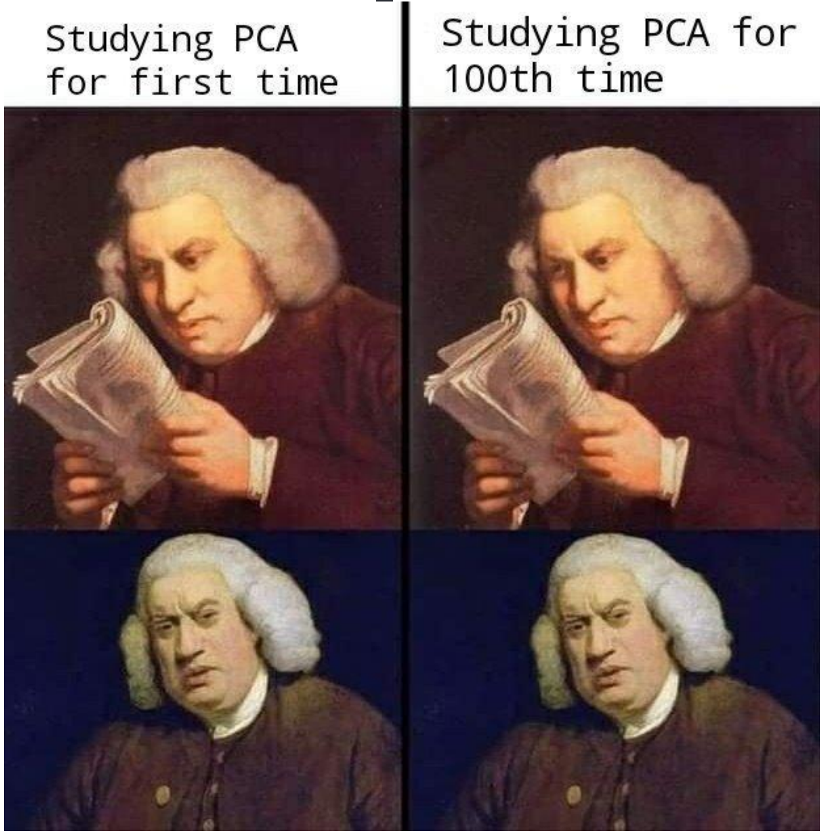
Principal Component Analysis!



Welcome to Week 10 Lecture 1!

Data Science in Python &
Machine Learning

Last Week: Clustering

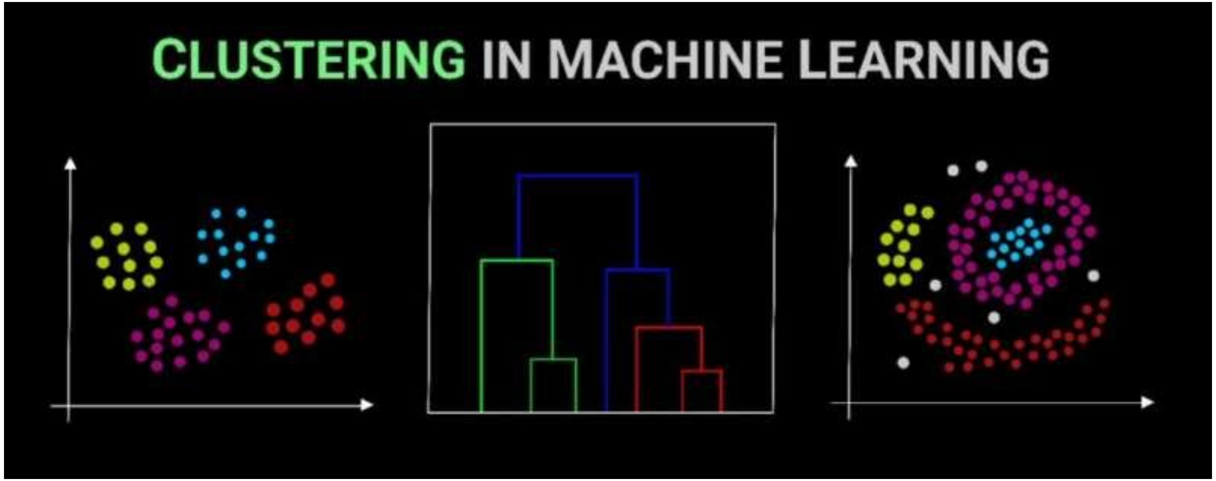


Image Source

Clustering groups similar samples together.

An unsupervised model defines What ‘similar’ means!

Learning Objectives

- ❑ List the pros and cons of dimensionality reduction.
- ❑ Explain how principal component analysis reduces the dimensionality of data while retaining maximum information.
- ❑ Apply PCA to reduce the dimensionality of a set of features to prepare them for supervised learning without leaking data.

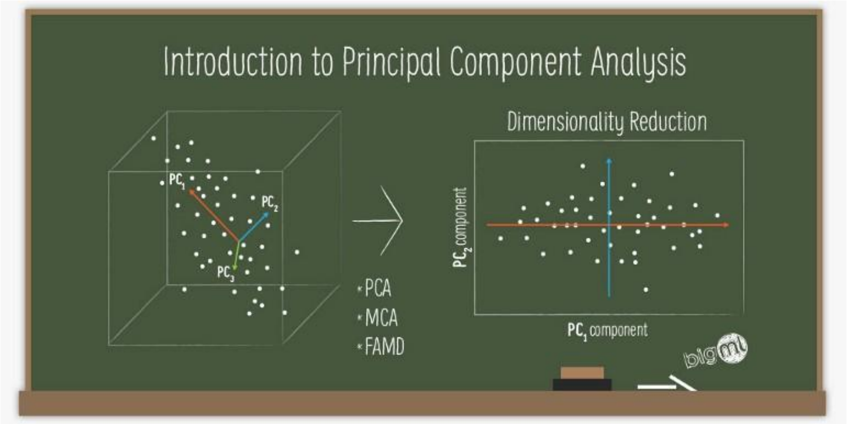


Image courtesy of [Mthanraj Sharma]

Types of Unsupervised Learning

Clustering	Dimensionality Reduction
Groups Data Together	Combines and Changes Features
Analysis	Reduces Number of Features
Feature Extraction	Feature Engineering

Feature Engineering:

- Make new features from old features
- Transform features
- Combine features
- Improve model's ability to make predictions.

Photo by Christopher Burns on Unsplash

Why Dimensionality Reduction?

- **<u>Dimensions</u> are features (columns in the dataset)**

- Machine learning datasets can have a huge number of features (even in the millions!)

    - Too many features slow training and/or predicting
        - Certain algorithm training or predicting times are *especially* sensitive to more features

    - “Curse of Dimensionality”
        - Clustering algorithms tend to perform worse with more features: data more ‘spread out’

    - Greater risk of overfitting.
        - Dimensionality reduction can be regularization by reducing complexity

Why Dimensionality Reduction Quiz

Principal Component Analysis

- Combines all features into new features called Principal Components
  - **These are NOT the same as the original features!!!**

- Principal Components are ordered from most informative to least.
  - i.e. first PC explains the most variance, second PC explains the next most…

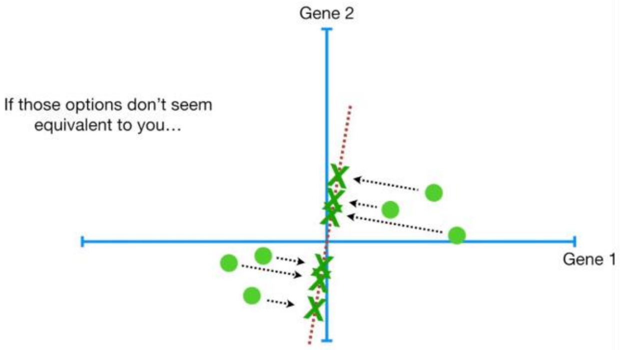


Image Source

Visualizing dimensionality reduction

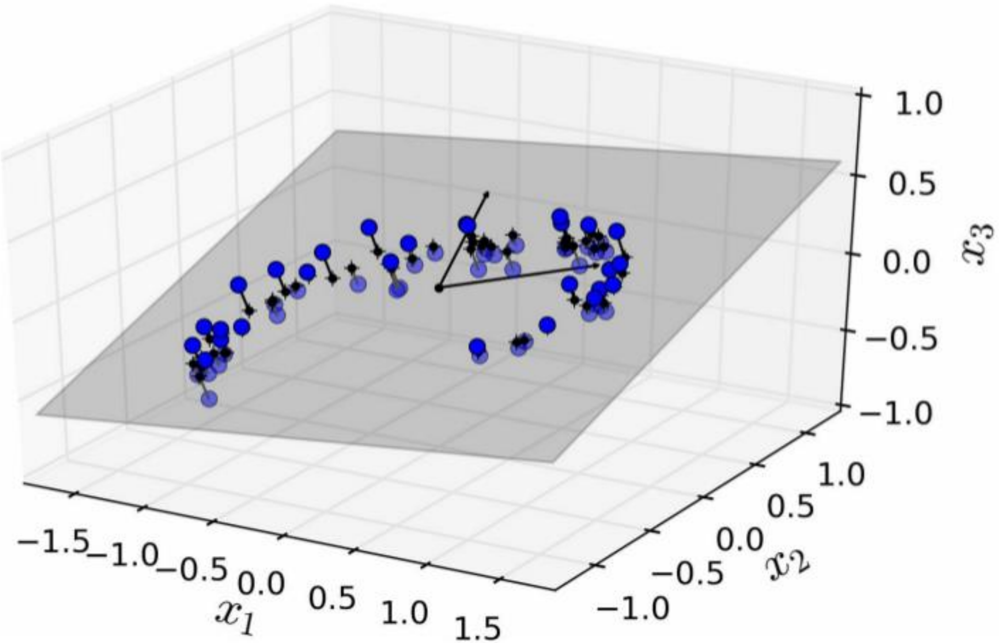


Figure 8-2. A 3D dataset lying close to a 2D subspace

- NOT just dropping columns

  Example
- This dataset is on a 3d plane (it has 3 features or dimensions)
- Data is usually NOT spread out evenly across all dimensions
- 2d plane captures most of the variance in the data
- Now, we can “pull out” that 2d plane and that becomes a 2d graph!

Source: Hand-On Machine Learning with Scikit-Learn, Keras & Tensorflow by Aurelien Geron

<u>3 dimensions</u> becomes <u>2 dimensions</u>

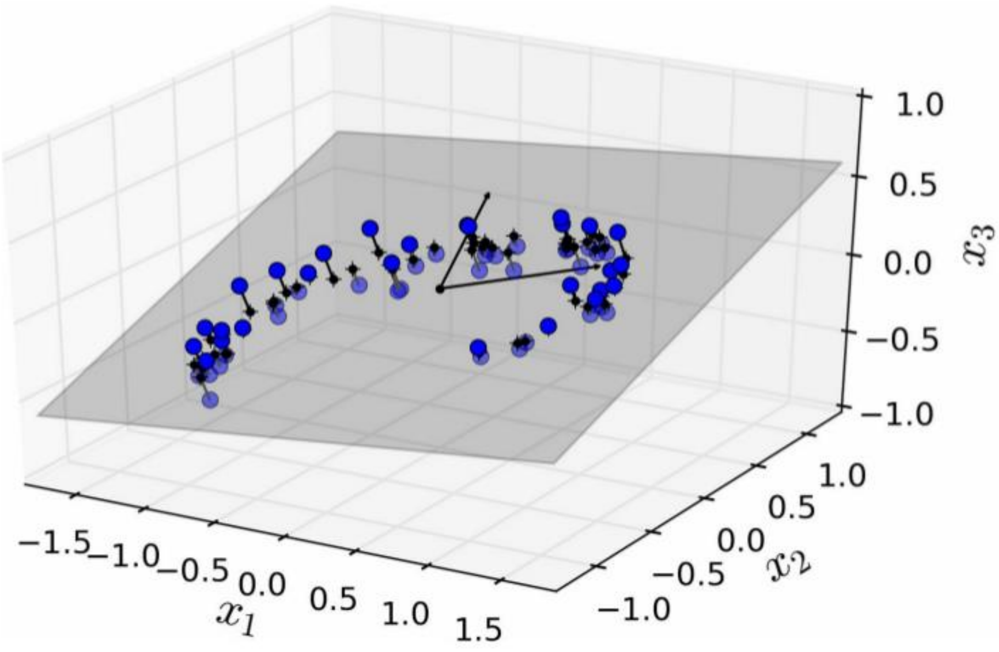


Figure 8-2. A 3D dataset lying close to a 2D subspace

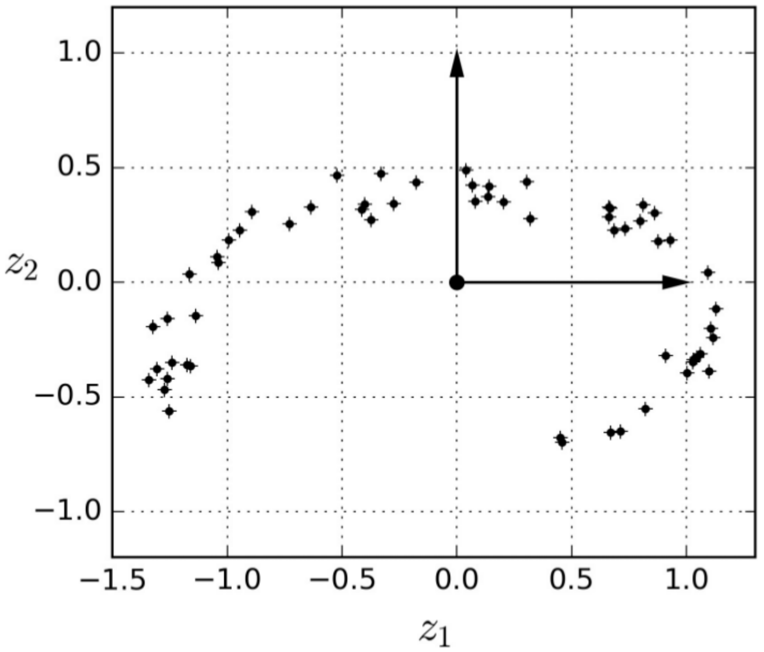


Figure 8-3. The new 2D dataset after projection

- This method is called <u>projection.</u>

- We now have decreased from 3 to 2 dimensions.
- Note: We have completely lost interpretability of the dimensions

How Are Principal Components Defined?

Each new component is defined as a combination of the original features, for example:

If we are reducing a dataset with 3 features, X1, X2, and X3

Into a new dataset with 2 features, Z1 and Z2,

The new features might be defined as:

Z1 = (X1 * 0.7) + (X2 * 1.3) + (X3 * -0.9)

Z2 = (X1 * 1.2) + (X2 * 1.5) + (X3 * -0.2)

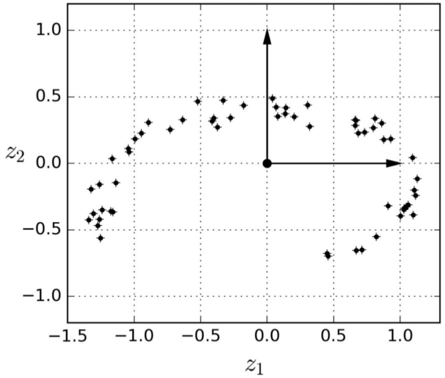


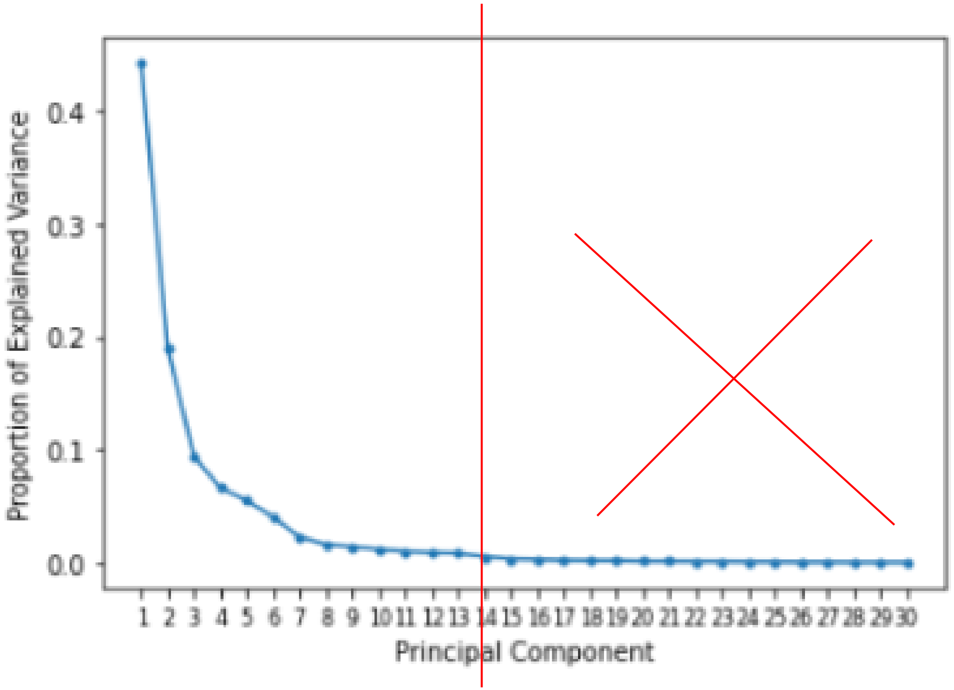
Figure 8-3. The new 2D dataset after projection

For each of the 3 features of each data point in the original dataset we would use the above formulae to convert them to the 2 features of the new dataset.

[PCA Step by Step]

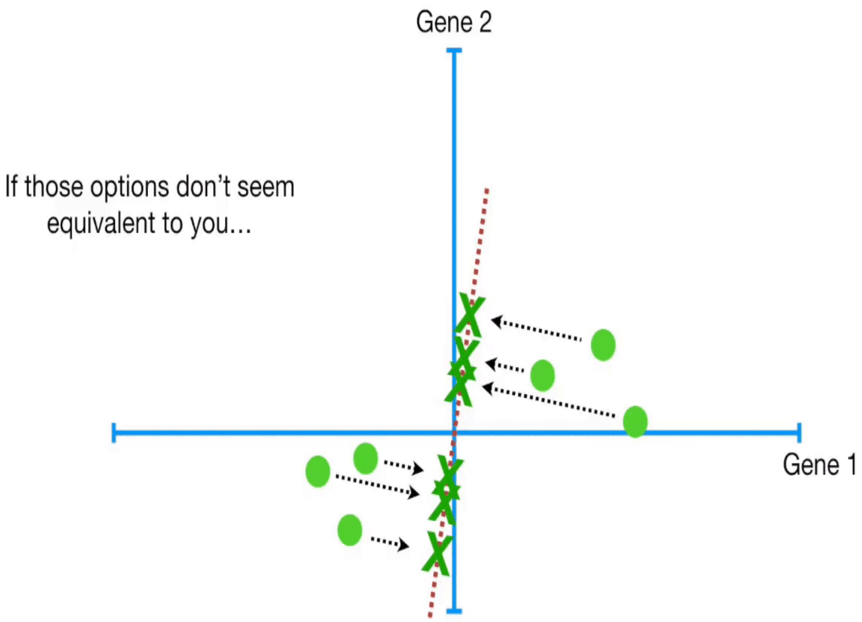
Dimensionality Reduction

- Principal Components are ordered, most explained variance to least.
- Reduce dimensionality with minimal information loss by removing later PCs



Principal Component Analysis Review

- Resulting features are called ‘Principal Components’
- Some information is always lost if you drop any principal components
- Principal components are NOT interpretable.
- Components are ordered, each explains more variance than the next.



Pros and Cons of PCA

Pros:

- It speeds up training for huge datasets
- Reduces “curse of dimensionality”
- Can reduce overfitting
- It allows us to visualize higher dimensional data on a 2d or 3d plot

Cons:

- Lose information (variance)
- Lose interpretability
- Transformation is computationally expensive

Breakout Discussion: 3 minutes

In your own words, how does PCA reduce the dimensionality of the data while losing minimal information?

Choose a reporter that will report the group's discussion.

PCA in Python

Import PCA

```
from sklearn.decomposition import PCA
```

Instantiate PCA

To return 20 components:

```
pca = PCA(n_components=20)
```

To capture 50% of variance:

```
pca = PCA(n_components=0.5)
```

PCA as a Transformer in a Model Pipeline

Do other preprocessing: OHE, Ordinal Encoding, Scaling, Imputing, etc BEFORE PCA,

Then PCA transform ALL features.

```
knn_pipe = make_pipeline(preprocessor, pca, model)
```

```
knn_pipe.fit(X_train, y_train)
```

Today’s Challenge

[Today’s Data: Identify defects in motors]

- “Column 49” is the target, note that 1 is the normal condition, and the others are various types of defects
- Goal: Predict the condition of the motor with the highest overall accuracy
- This dataset is a great candidate for PCA because it has a lot of features
- Also the prediction task is not focused on interpreting the features: We just need to identify what type of defect it is.

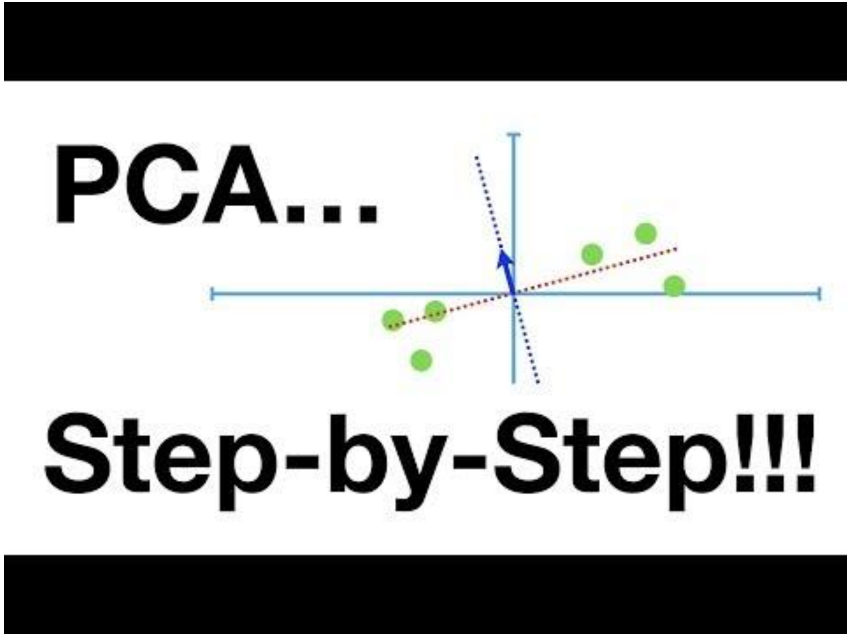
Challenge Notebook

Click on : Open in Colab

Original Source of Data

More Information:

Check out this awesome video by StatQuest!



Assignments This Week:

1. **PCA Exercise**:
   a. MNIST dataset: Image classification!
   b. 2 different ways to load the dataset are shown in assignment
      i. *Option 1: fetch_openml*
         1. mnist.data = features
         2. mnist.target = target
      ii. *Option 2: keras.datasets.mnist*
         1. Comes already split
         2. Must be reshaped for traditional ML (See code)

Assignments This Week (cont):

1. **Feature Engineering Exercise**
    a. Be sure to carefully read ALL directions!
    b. Please use a Lambda function to convert temperatures
        i. Double check your conversion formula: are the results sensible? Would many people likely go bike riding in 104 degree heat?

Assignments This Week (cont):

1. **Project 2 Part 4**
    a. This is big assignment! Set aside plenty of time!
    b. You are finishing your project
        i. PCA
        ii. Model development
        iii. Finishing touches (code comments, clean code, section headings, professional quality ‘final draft’ notebook)
            1. 1 notebook with all parts
            2. Or 2 notebooks: analysis and modeling
            3. Don’t name sections ‘part 1, part 2, etc’
        iv. README: Final draft! Visually appealing with proper, readable, English.
            1. If writing in English is not a strength, ask someone to read it over for you for grammar and clarity.
            2. Employers will judge you on your ability to communicate about analysis and modeling!

Study Next: Feature Engineering

To prepare for next lecture:

Please Read:

1. Feature Engineering: Overloaded Operators
2. Feature Engineering: Strings
3. Feature Engineering: Datetime
4. Feature Engineering: Functions

Daily Schedule