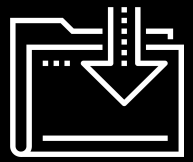
Principal Component Analysis (PCA)

Data Boot Camp

Lesson 19.3



© 2022 Trilogy Education Services, a 2U, Inc. brand. All Rights Reserved.

Class Objectives

By the end of this lesson, you will be able to:

Explain PCA and how it can be used to reduce dimensionality in data.

Conceptualize how PCA relates to K-means and other applications in machine learning.

Use PCA to reduce the number of features in an unsupervised learning setting.


WELCOME

Activity: Warm-Up

In this activity, you will standardize global currency data, and then segment it into three clusters.

Suggested Time:

20 Minutes

Time's Up! Let's Review.

Questions?

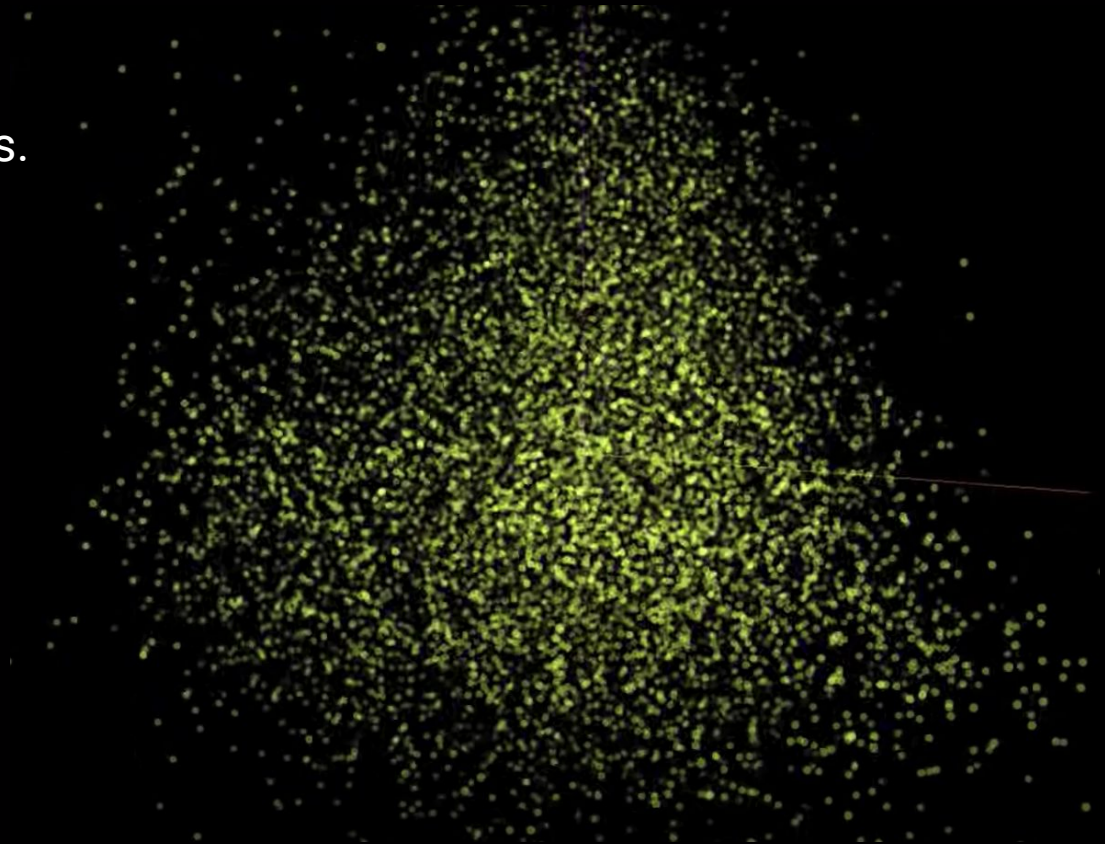
Introduction to PCA


Time For a Quick Video
Visualizing High-Dimensional Space

Introduction to PCA

Some of these datasets displayed are very large, with thousands (or more!) columns.

For some machine learning algorithms, this high dimensionality can become a problem.



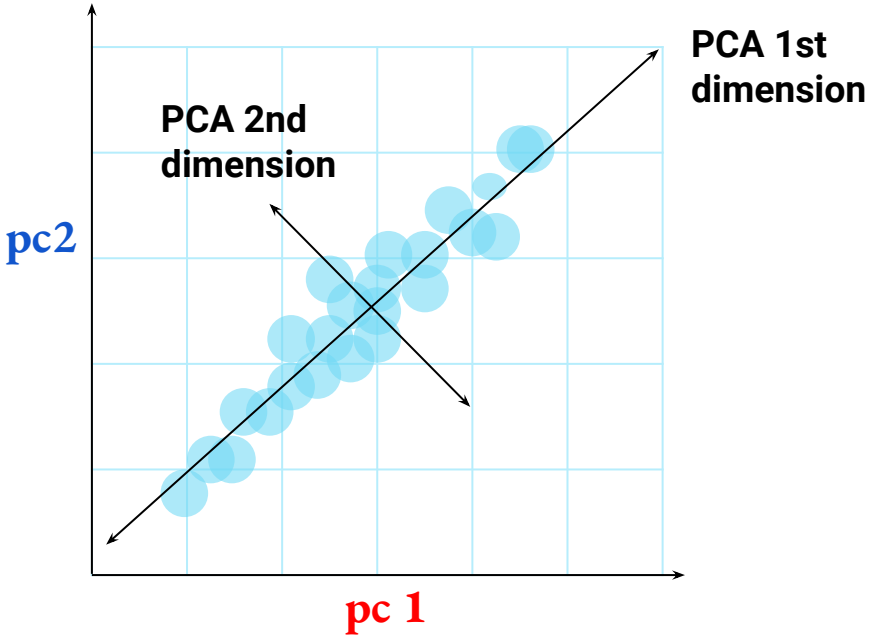
https://experiments.withgoogle.com

What if we could reduce the size of the dataset while preserving as much useful information as possible?

Introduction to PCA

The principal component analysis algorithm, or PCA, works because all data has **variability**; variation contains useful information.

By keeping only the most common variation across the dataset, we can reduce the size of our dataset but still maintain a variety of useful features.



Introduction to PCA

PCA is a statistical technique that we can use to streamline the machine learning process when too many factors exist in the data.

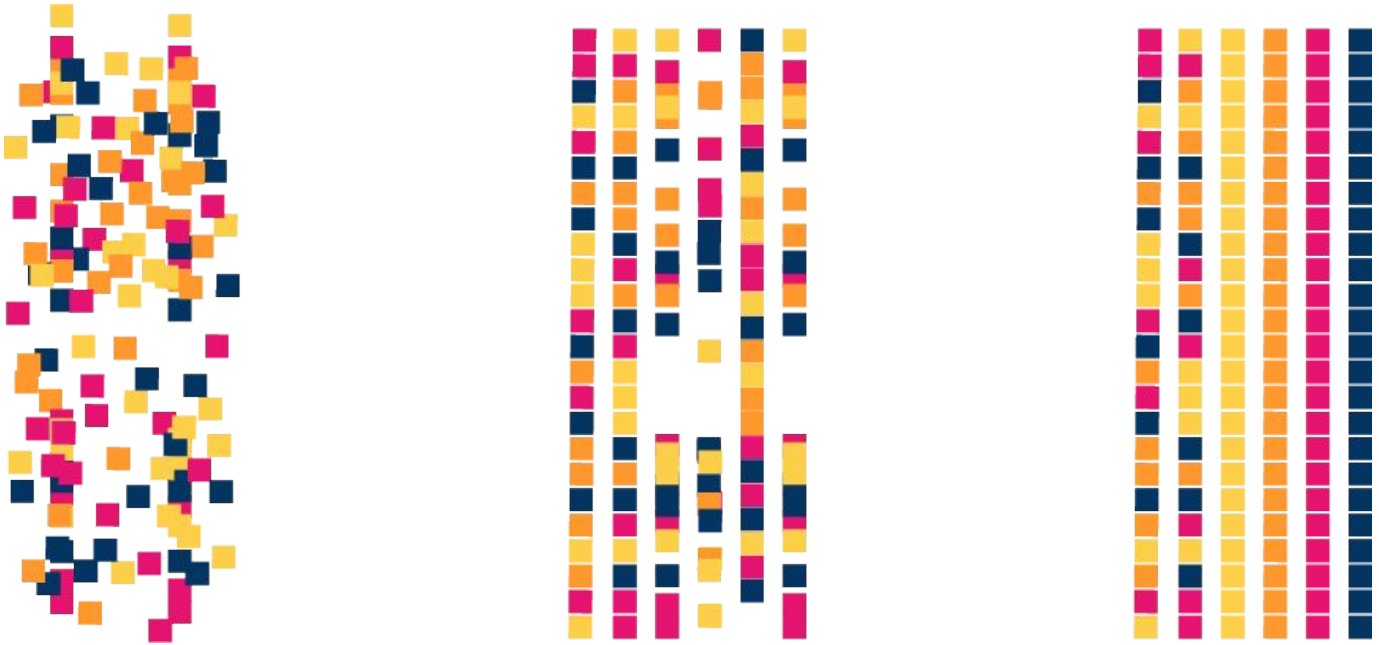
PCA reduces the number of factors by transforming a large set of features into a smaller one that contains MOST of the information of the original, larger dataset.

PCA is a dimensional-reduction method that looks at all the dimensions (or data columns) in a dataset, which:

- Analyzes the weight of their contribution to the variance in the dataset.
- Reduces variables to a smaller set of dimensions that still contains as much of the information, or the maximum variance, of the original dataset as possible.

Introduction to PCA

Reducing the number of factors, or **dimensional reduction**, comes at the expense of some accuracy, but the goal is to trade a little accuracy for simplicity.



The goal of PCA is simple:

Reduce the number of factors in a dataset while preserving as much of the information from the original dataset as possible.

Instructor Demonstration

Introduction to PCA

Questions?

Activity: Segmenting with PCA

In this activity, you will use your knowledge of PCA to reduce the dimensionality of the customers DataFrame from the previous activity, and then compare that result to the segmentation of the data by using all the factors.

Suggested Time:

25 Minutes

Activity: Segmenting with PCA

Using the starter code and the customer data provided, perform the following steps:

- Reduce the factors to only two dimensions by using PCA.
- Determine the optimal value for k by using the PCA DataFrame.
- Segment the data by using the K-means algorithm and the optimal value for k.
- Segment the preprocessed customer DataFrame by using the K-means algorithm and the same value for k.
- Compare the segmentation results.

Time's Up! Let's Review.

Questions?


Break

Activity: Energize Your Stock Clustering

In this activity, you will:

- Strengthen your ability to reduce the number of features in your model by using PCA.
- Cluster both standardized data and principal components into various clustering algorithms, and then compare the results.

Suggested Time:

30 Minutes

Time's Up! Let's Review.

Questions?

Structured Review


The End