應用機器學習

Brian Chan 陳醒凡

課程目標

1. 了解基本的數據分析
2. 了解基本的機器學習(Machine Learning)方法
3. 掌握Python的基本操作和一些有用的package
4. 處理及從網上下載數據
5. 在Python上應用機器學習

今天課堂概要

Dimension reduction

1. Curse of dimensionality
2. Principal component analysis (PCA)
3. Examples: MNIST and Wine datasets

DIMENSION REDUCTION

Motivations:

1. Speed up training algorithm
2. Remove noise and redundant features
3. Help visualize data and have better insight on important features
4. Compress memory space

DIMENSION REDUCTION

Drawbacks:

1. Some information is lost. Hence, subsequent algorithm trained may be degraded.
2. May be computationally expensive.
3. Transformed feature may be hard to interpret.

CURSE OF DIMENSIONALITY

The curse of dimensionality refers to various phenomena that arise when analyzing and organizing data in high-dimensional spaces (often with hundreds or thousands of dimensions) that do not occur in low-dimensional settings such as the three-dimensional physical space of everyday experience.

e.g.

- Sparsity of data

- Computational complexity

PRINCIPAL COMPONENT ANALYSIS (PCA)

PCA is a technique that seeks a r-dimensional basis that best captures the variance in the data.

The direction with the largest projected variance is called the fist principal components.

IDEA OF PCA

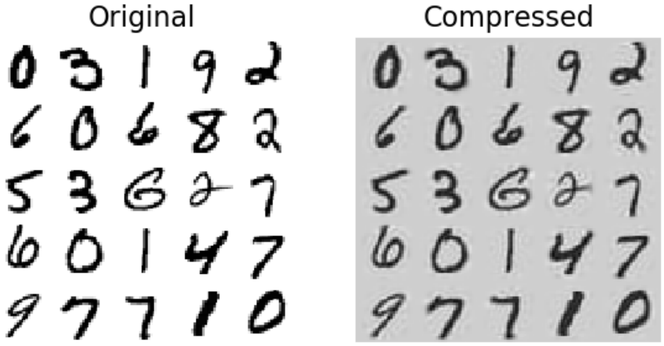
Suppose there is a dataset which dimension of 1,000.

Scenario 1: The dataset is composed of data points that are almost perfectly aligned.

Scenario 2: The dataset is composed of perfectly random points, scattered all around 1,000 dimensions.

How many dimensions are needed to preserve 95% of the variance?

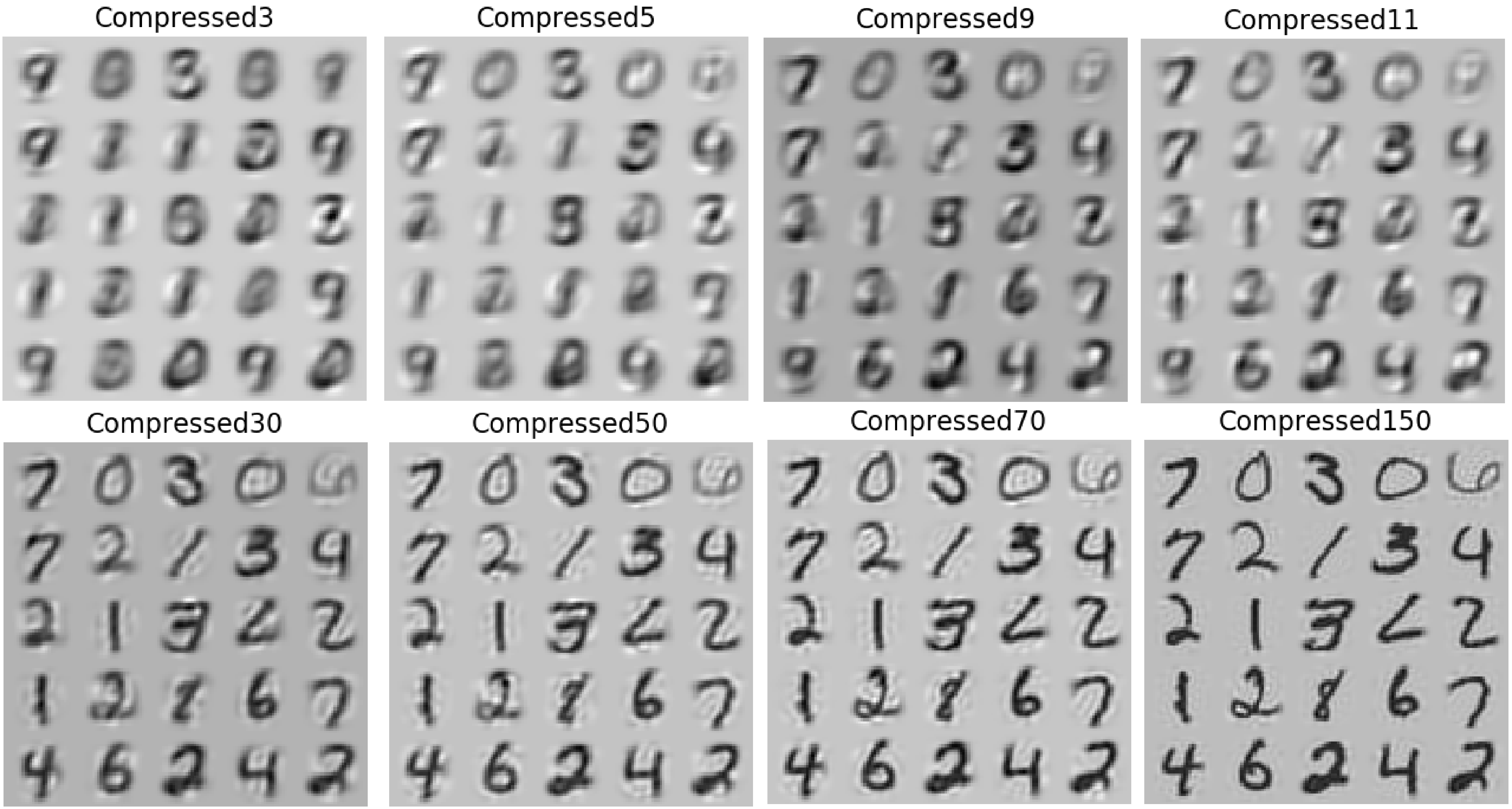
EXAMPLE - MNIST DATASET



Obviously after dimensionality reduction, the training set takes up much less space.

For example, try applying PCA to the MNIST dataset while preserving 95% of its variance. You should find that each instance will have just over 150 features, instead of the original 784 features. So while most of the variance is preserved, the dataset is now less than 20% of its original size!

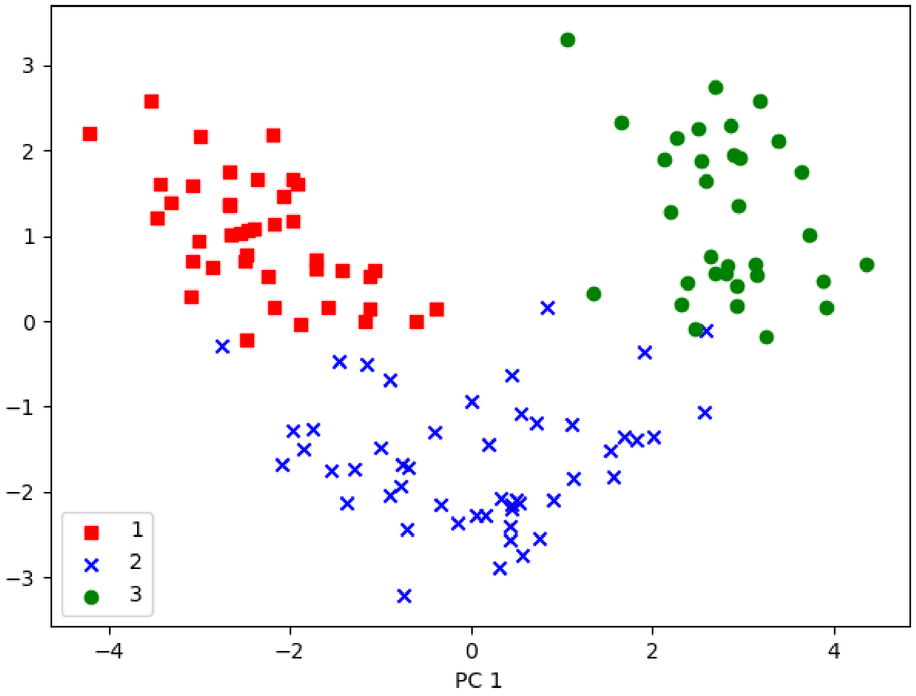
This is a reasonable compression ratio, and you can see how this can speed up a classification algorithm (such as an SVM classifier) tremendously.

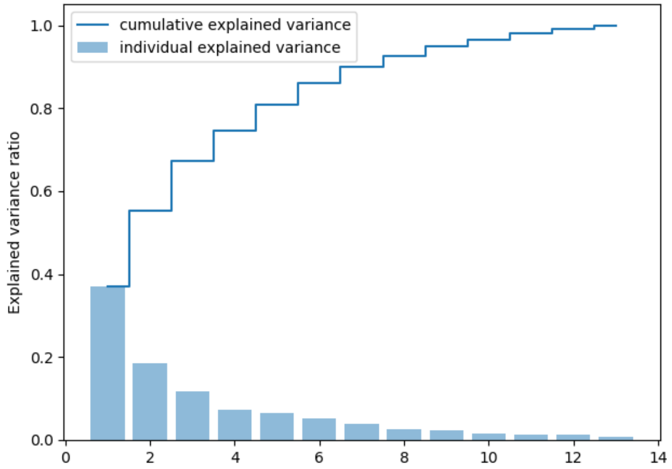

Compressed3
Compressed5
Compressed9
Compressed11
Compressed30
Compressed50
Compressed70
Compressed150

EXAMPLE – WINE DATA

In this example, we will tackle the first four steps of a PCA:

1. Standardizing the data.
2. Constructing the covariance matrix.
3. Obtaining the eigenvalues and eigenvectors of the covariance matrix.
4. Sorting the eigenvalues by decreasing order to rank the eigenvectors.





SOME ALTERNATIVE DC METHODS

1. Multidimensional Scaling (MDS)
2. Isomap
3. t- Distributed Stochastic Neighbor Embedding (t-SNE)
4. Linear Discriminant Analysis (LDA)

REFERENCE

PCA methodology:

https://machinelearningmedium.com/2018/04/22/principal-component-analysis/

https://medium.com/@kyasar.mail/pca-principal-component-analysis-729068e28ec8

https://www.kaggle.com/akhileshrai/pca-for-visualisation-classification/comments

Sklearn documentation:

https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html

今天課堂概要

Dimension reduction

1. Purpose of dimension reduction
2. PCA
3. Example

下一課…

Evaluation methods