Pattern Recognition

Workshop 2. PCA

Prof. Dr. Alina Nechyporenko

Biosystemtechnik/Bioinformatik BBM24

Content

- PCA fundamentals
- Scikit-learn (sklearn) package
- Implementation of PCA

Goal of workshop

You will…

… learn how to implement PCA

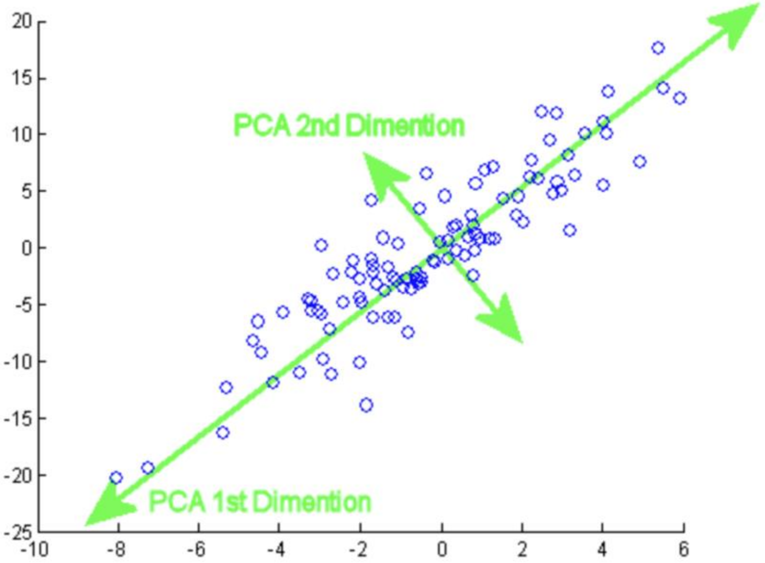
... you will get skills in Scikit-learn, Numpy, Matplotlib

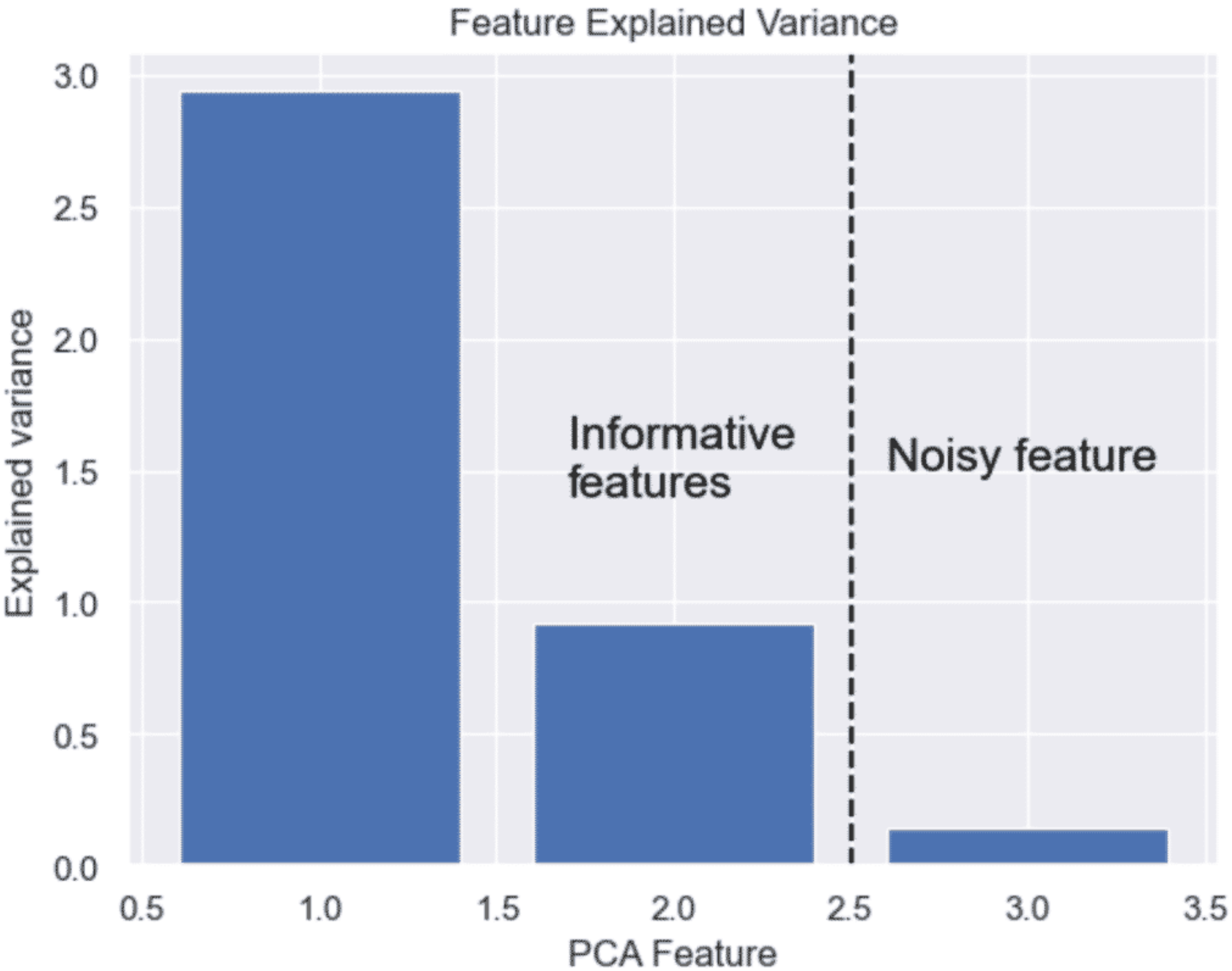
…how to integrate PCA in ML tasks

Principal component analysis

Refresh your mind

- With PCA, we can find a low-dimensional representation of the dataset that contains **as much as possible of the variation**. Therefore, we only get the most *interesting* features, because they are responsible for the **majority of the variance**




Feature Explained Variance
Explained variance
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Informative features
Noisy feature
0.5
1.0
1.5
2.0
2.5
3.0
3.5
PCA Feature

- linear mapping of the data to a lower-dimensional space is performed in a way that maximizes the variance of the data
- PCA assumes that features with low variance are irrelevant and features with high variance are informative

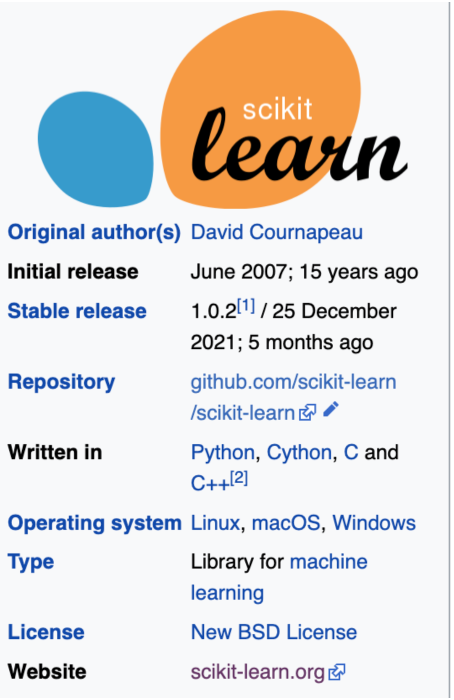
PCA step by Step

- Standardization
- Covariance matrix computation
- Computation of eigenvectors and eigenvalues
- Constraction of a feature vector
- Recast the data along the principlal components axes

What is…

Scikit-learn (sklearn)?

- Scikit-learn is **an open source Machine Learning Python package that offers functionality supporting supervised and unsupervised learning**. Additionally, it provides tools for model development, selection and evaluation as well as many other utilities including data pre-processing functionality



Original author(s)	David Cournapeau
Initial release	June 2007; 15 years ago
Stable release	1.0.2[1] / 25 December 2021; 5 months ago
Repository	github.com/scikit-learn/scikit-learn
Written in	Python, Cython, C and C++[2]
Operating system	Linux, macOS, Windows
Type	Library for machine learning
License	New BSD License
Website	scikit-learn.org

Implementation

- Scikit-learn is largely written in Python, and uses NumPy extensively for high-performance linear algebra and array operations. Furthermore, some core algorithms are written in Cython to improve performance. Support vector machines are implemented by a Cython wrapper around LIBSVM; logistic regression and linear support vector machines by a similar wrapper around LIBLINEAR. In such cases, extending these methods with Python may not be possible

- Scikit-learn integrates well with many other Python libraries, such as Matplotlib and plotly for plotting, NumPy for array vectorization, Pandas dataframes, SciPy, and many more

https://scikit-learn.org/stable/install.html

Task

Goal

- Goal is to implement principal component analysis in order to reduce an amount of data for classification model

- We will follow the classic machine learning pipeline where we will first import libraries and dataset, perform exploratory data analysis and preprocessing, and finally train our models, make predictions and evaluate accuracies

Data set description

- *Iris* flower data set
- https://en.wikipedia.org/wiki/Iris_flower_data_set
- https://archive.ics.uci.edu/ml/datasets/iris

What to do

- Create a project folder
- Create jupyter notebook
- Import python libraries and packages
- Download the data or import the dataset from https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data
- Preprocessing
  - divide the dataset into a feature set and corresponding labels
  - divide data into training and test sets
  - perform standard scalar normalization to normalize feature set
- Apply PCA

Performing PCA using Scikit-Learn

- Initialize the PCA class by passing the number of components to the constructor
- Call the fit and then transform methods by passing the feature set to these methods. The transform method returns the specified number of principal components
- Find principal components
- The PCA class contains explained_variance_ratio_ which returns the variance caused by each of the principal components

Application of PCA in the problem of data classification

Training and Making Predictions

- use random forest classification for making the predictions
- perform evaluation of model
- check how principal components affect a performance of model
- try to implement other classification models

Workflow

- Load Dataset
- Standardize the Data
- Apply PCA
- Apply classification model

Take home message

- PCA fundamentals
- Scikit-learn (sklearn) package
- Implementations of PCA

Vielen Dank für Ihre Aufmerksamkeit !

To be installed

Python https://www.python.org/downloads/
- Jupyter https://jupyter.org/install
- VS Code https://code.visualstudio.com/download
- Docker https://www.docker.com/products/docker-desktop/