Feature Selection with PCA

Yasaman Amannejad,
Department of Mathematics and Computing,
Mount Royal University

Outline

- What is feature selection?
- Feature selection methods in python
- Principle Components Analysis (PCA)
- Feature selection with PCA
- Time analysis of PCA

Feature Selection

- Feature selection is a process where you automatically select those features in your data that contribute most to the prediction variable or output in which you are interested.

- Three benefits of performing feature selection before modeling:
  - **Reduces Overfitting**: Less redundant data means less opportunity to make decisions based on noise.
  - **Improves Accuracy**: Less misleading data means modeling accuracy improves.
  - **Reduces Training Time**: Less data means that algorithms train faster.

Feature Selection Techniques

Automatic Feature Selection

1. Removing features with low variance
2. Univariant statistics
   - SelectKBest
   - SelectPercentile
3. Recursive feature elimination
4. Model-based selection

Scikit Library: sklearn.feature_selection

Removing features with low variance

- **VarianceThreshold** is a simple baseline approach to feature selection. It removes all features whose variance doesn't meet some threshold. By default, it removes all zero-variance features, i.e. features that have the same value in all samples.

from sklearn.feature_selection import VarianceThreshold

Let's see the example

Univariant Statistics

- Determine if statistically significant relationship between each feature and the output.
- Each feature is studied in isolation.

- Univariant statistics
  - SelectKBest: Select features according to the k highest scores.
  - SelectPercentile: Select features according to a percentile of the highest scores.

Recursive Feature Elimination

sklearn.feature_selection.RFE

The goal of recursive feature elimination (RFE) is to select features by recursively considering smaller and smaller sets of features.

Recursive feature elimination with cross-validation

sklearn.feature_selection.RFECV

- Feature ranking with recursive feature elimination and cross-validated selection of the best number of features.

Model-based Feature Selection

Meta-transformer for selecting features based on importance weights.

- RFE removes least significant features over iterations.
- SelectFromModel is a little less robust as it just removes less important features based on a threshold given as a parameter. There is no iteration involved.

Reminder: Standardize the Data

- Feature scaling is an important preprocessing step for many machine learning algorithms.
- Rescale the features such to have the properties of a standard normal distribution with a **mean of zero** and a **standard deviation of one**.
- The standard score of a sample x is calculated as:

$$z = \frac{(x - u)}{z}$$

u: mean of the training data
z: standard deviation of the training sample

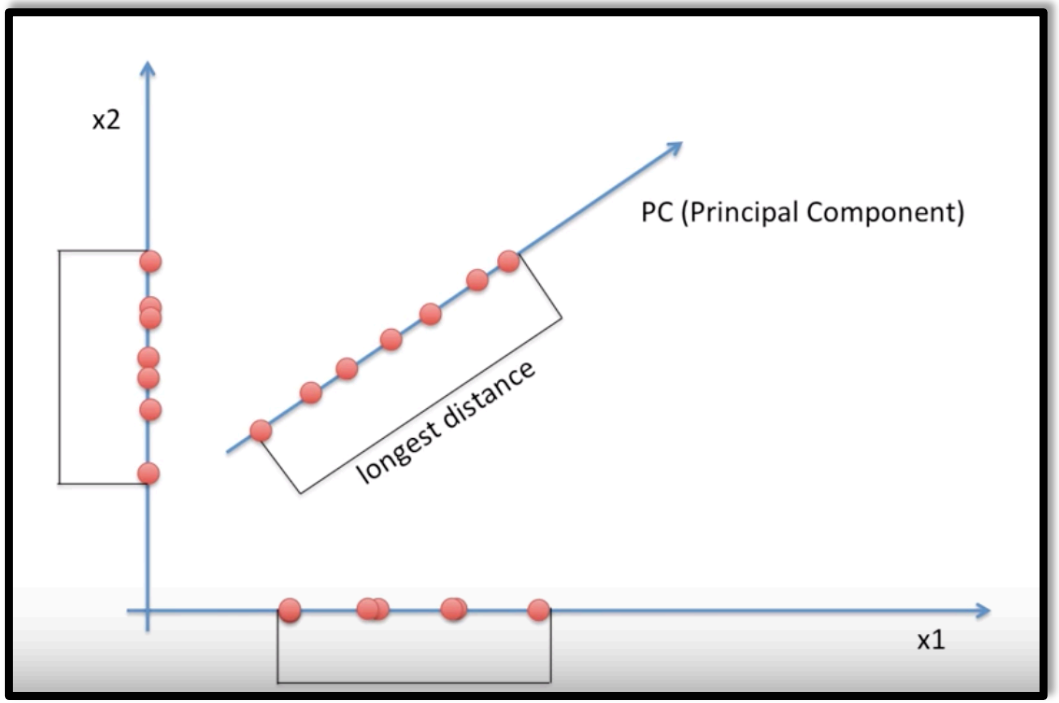
PCA

Principle Component Analysis (PCA)

- PCA is a statistical method for dimensionality reduction and feature extraction that transforms the data from a d-dimensional space into a new coordinate system of dimension p, where p<= d (the worst case would be to have p = d).

PCA: Goal

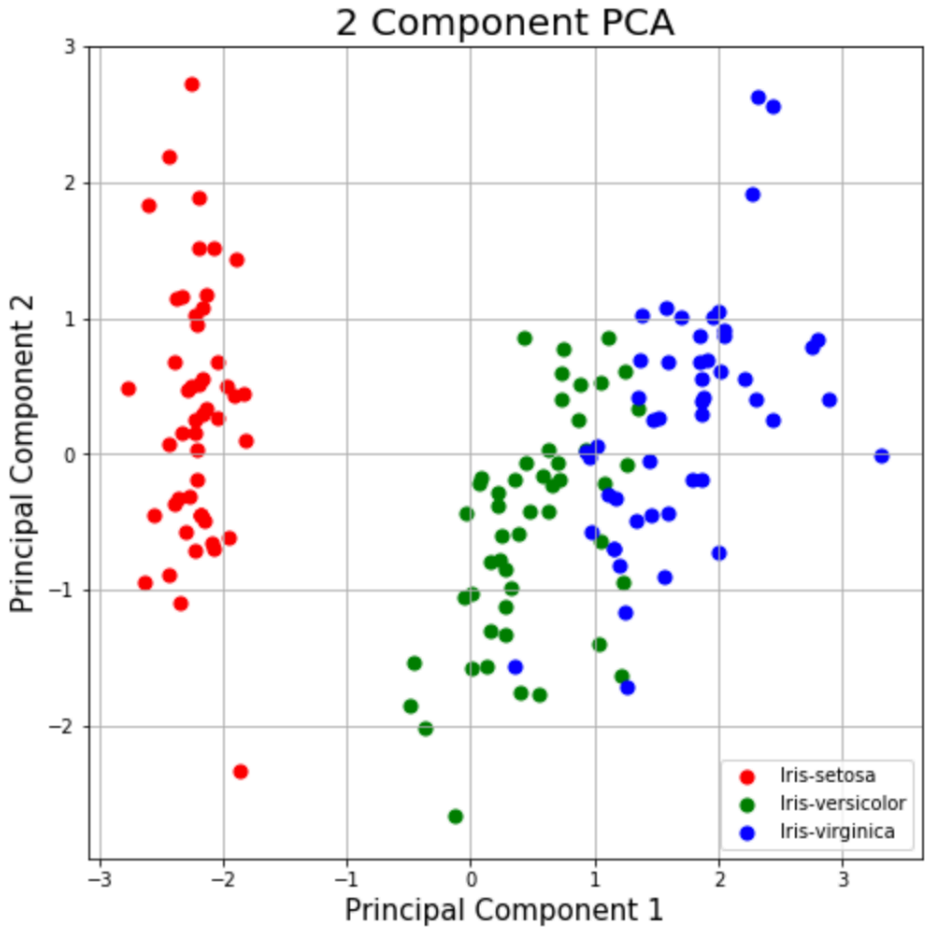
- **Goal**: to preserve as much of the variance in the original data as possible in the new coordinate systems.



Explained Variance

- The amount of variance explained by each of the selected components.

PCA for Visualization



PCA for Speeding

- PCA with Regression

Variance Retained	Number of Components	Time (mSec)	Accuracy
0.99	54	3731	0.968
0.95	40	3,692	0.967
0.90	31	3,571	0.96
0.85	25	2,372	0.94

Yasaman Amannejad, PhD
Assistant Professor, Mount Royal University

Email: yamannejad@mtroyal.ca
Website: mru.ca/amannejad