DATA MINING 2
Dimensionality Reduction

Riccardo Guidotti

a.a. 2019/2020

UNIVERSITÀ DI PISA

Dimensionality Reduction

- Dimensionality reduction is the process of reducing the number of random variables under consideration by obtaining a set of principal variables.
- Approaches can be divided into **feature selection** and **feature projection**.

X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
1.1	10	0.3	0.5	A	1	C	15	1.3	a
1.2	12	0.3	0.7	A	0	D	19	1.8	P
…	…	…	…	…	…	…	…	…	…

X_A	X_B
1.8	5.4
1.9	6.3
…	…

Feature Selection

- Select a subset of the features according to different strategies:
    - the **filter** strategy (e.g. information gain),
    - the **wrapper** strategy (e.g. search guided by accuracy),
    - the **embedded** strategy (selected features add or are removed while building the model based on prediction errors).

- Classification and/or regression or can be done in the reduced space more accurately than in the original space.

Feature Selection

- **Variance Threshold.** 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.
- **Univariate Feature Selection.** It selects the best features based on univariate statistical tests. For instance it removes all but the k highest scoring features. An example of statistical test is the ANOVA F-value between label/feature.
- F-value = $\sum_{i=1}^{K} n_i(\bar{Y}_{i\cdot} - \bar{Y})^2/(K-1) / \sum_{i=1}^{K}\sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i\cdot})^2/(N-K),$
- where $\bar{Y}_{i\cdot}$ denotes the sample mean in the i^{th} group, n_i is the number of observations in the i^{th} group, $\bar{Y}$ denotes the overall mean of the data, Y_{ij} is the j^{th} observation in the i^{th} out of K groups, K denotes the number of groups, N the overall sample size.
- F-value is large if the numeartor is large, which is unlikely to happen if the population means of the groups all have the same value.

Recursive Feature Elimination (RFE)

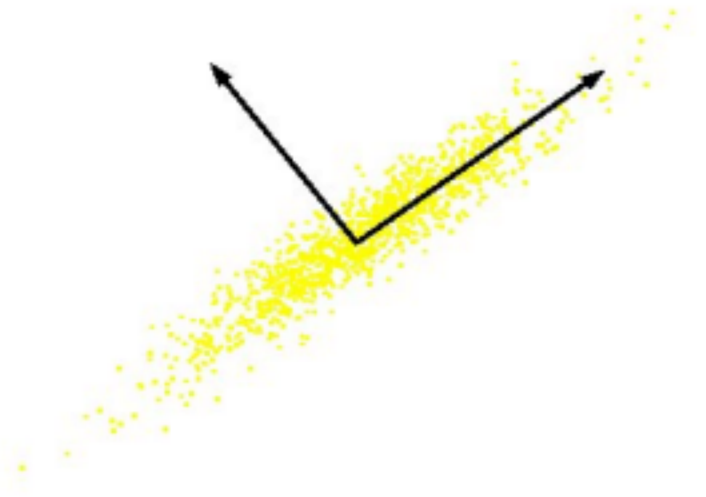
- Given an external estimator that assigns weights to features (e.g., the coefficients of a linear model, or feature importance of decision tree), RFE selects features by recursively considering smaller and smaller sets of features.
- First, the estimator is trained on the initial set of features and the importance of each feature is obtained.
- Then, the least important features are pruned from current set of features.
- That procedure is recursively repeated on the pruned set until the desired number of features to select is eventually reached.

Feature Projection (a.k.a Feature Extraction)

- It transforms the data in the high-dimensional space to a space of fewer dimensions.
- The data transformation may be linear, or nonlinear.
- Approaches:
    - Principal Component Analysis (PCA)
    - Singular Value Decomposition (SVD)
    - Non-negative matrix factorization (NMF)
    - Linear Discriminant Analysis (LDA)
    - Autoencoder

Principal Component Analysis

- The goal of PCA is to find a new set of dimensions (attributes or features) that better captures the variability of the data.
- The first dimension is chosen to capture as much of the variability as possible.
- The second dimension is orthogonal to the first and, subject to that constraint, captures as much of the remaining variability as possible, and so on.



Covariance

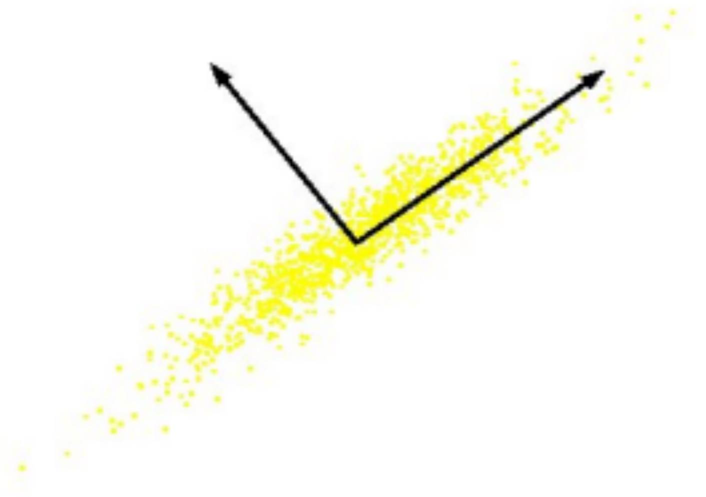
- The covariance of two attributes is a measure of how strongly the attributes vary together.

$$\text{covariance}(\mathbf{x}, \mathbf{y}) = s_{xy} = \frac{1}{n-1} \sum_{k=1}^{n} (x_k - \overline{x})(y_k - \overline{y})$$

- PCA calculates the covariance matrix of all pairs of attributes.
- Given matrix A, remove the mean of each column from the column vectors to get the centered matrix C
- The matrix V = C^T C is the covariance matrix of the row vectors of A.

Eigenvalue and Eigenvectors

- Eigenvector of matrix *A*: a vector *v* such that $Av = \lambda v$
- λ: eigenvalue of eigenvector v
- A square matrix *A* of rank *r*, has *r* orthonormal eigenvectors $v_1, v_2, \ldots, v_r$ with eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_r$.
- Eigenvectors define an orthonormal basis for the column space of *A*

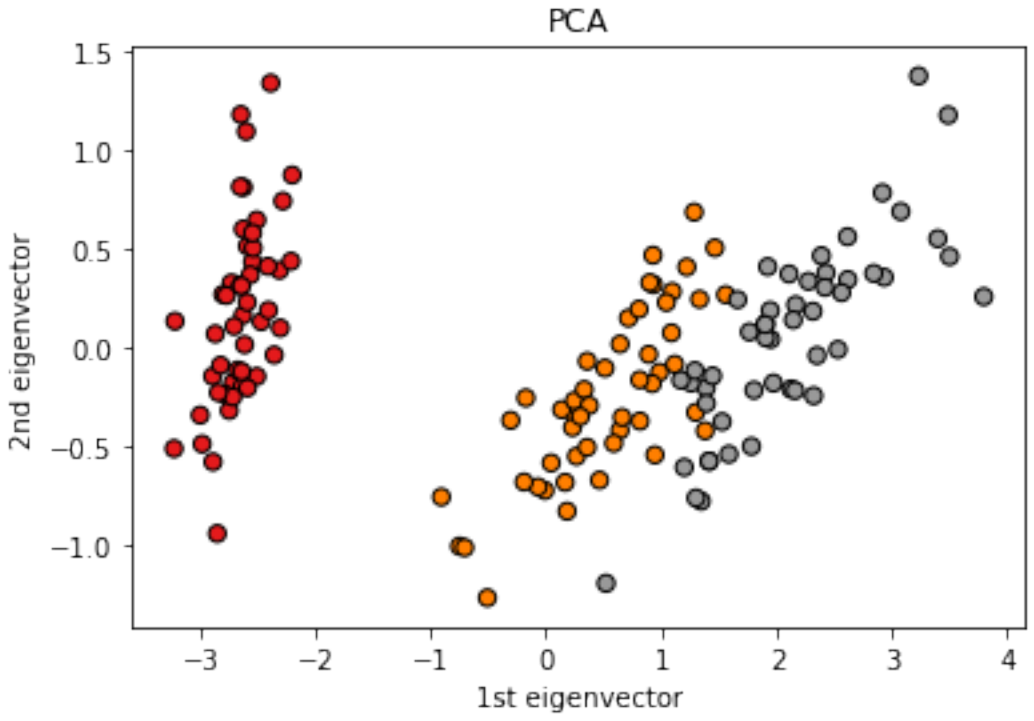


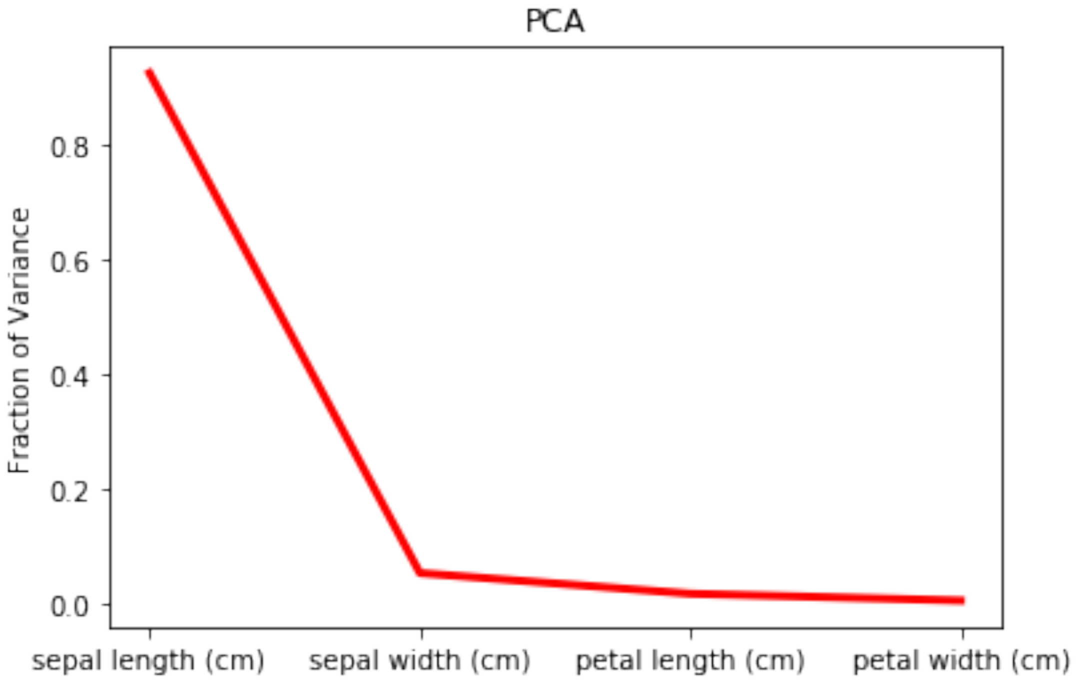
PCA Algorithm

- We finds the **eigenvalues** and **eigenvectors** of the covariance matrix (a positive semidefinite matrix with non-negative eigenvalues).
- The principal components are the eigenvectors with the largest eigenvalues and correspond to the dimensions that have the strongest correlation in the dataset.
- The new attributes have zero covariance to each other (they are orthogonal) and each attribute captures the most remaining variance in the data.
- The first attribute should capture the most variance in the data

Example

- Iris Dataset





References

- Dimensionality Reduction. Appendix B. Introduction to Data Mining.