EDA - Session 2

Exploring the Data II

Content

1. Exploring the data II
    - Continuous values
    - Correlation bi/multivariate
2. Feature selection
    - Methods
    - Alternative: PCA

3. Hands-On

From Exploring the data I

What should you do when encountering a new data set? In general:

- Answer the basic questions about, context, data set size, fields meaning
- Summary statistics
- Pairwise correlations
- Class breakdowns
- Plots of distributions

From Exploring the data I

What should you do when encountering a new data set? In general:

- Answer the basic questions about, context, data set size, fields meaning
- Summary statistics
- **Pairwise correlations**
- Class breakdowns
- Plots of distributions

Exploring the data II

Pairwise correlations measure relationships:

- Between independent variables
- Between independent variables and the outcome

Exploring the data II

Pairwise correlations are important for:

- Selection of relevant features
- Observing the predictive power of variables

Exploring the data II

Different measures of correlations (statistical framework) depends on the type of variables involved, for example:

- Pearson or spearman for continuous variables
- Phi's coefficient or Cramers' V for categorical variables
- Kendall's Tau for ordinal variables

Feature selection

At least two categories of selection can be described:

- Filter methods (Univariate)
  - f-test (ANOVA) Categorical output
  - chi2 categorical output
  - mutual information
  - Other (pearson, spearman, kendall's tau) Continuous input and continuous output

- Model based
  - Feature importance
  - Recursive Feature Elimination

Feature selection

An alternative is to use dimensionality reduction methods such as Principal Components Analysis.

Feature selection

An alternative is to use dimensionality reduction methods such as Principal Components Analysis. Briefly:

- PCA finds the directions of maximum variance in high-dimensional space and projects the data onto a new subspace with less or equal dimensions.

- Eigenvalues of covariance matrices are employed for selecting the k eigenvectors corresponding to the new dimensions.

Feature selection

An alternative is to use dimensionality reduction methods such as Principal Components Analysis. Briefly:

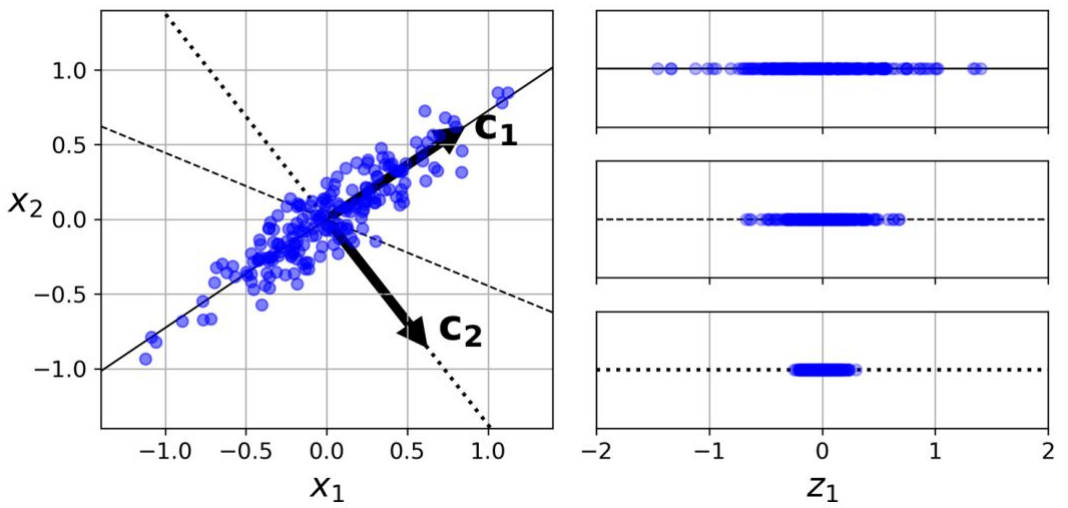
- PCA finds the directions of maximum variance in high-dimensional space and projects the data onto a new subspace with less or equal dimensions.

- Eigenvalues of covariance matrices are employed for selecting the k eigenvectors corresponding to the new dimensions.

The PCA directions are highly sensitive to data scaling, and we need to standardize the features prior to PCA if the features have different scales.

Feature selection

PCA Illustration



Hands-On