Methods for Dimensionality Reduction

Ramón Roales-Welsch

11 Januar 2018

Contents

Introduction

Overview

1. Multidimensional Scaling (MDS)
   - https://en.wikipedia.org/wiki/Multidimensional_scaling
   - Dimension reduction by using feature-distances between observations
2. Principal Components Analysis (PCA)
   - https://en.wikipedia.org/wiki/Principal_component_analysis
   - Reduce dimensions while maximizing variance of features
   - Slide 88: PCA is not applicable in Big Data (*FastMap* method is proposed.)
3. Feature Selection
   - https://en.wikipedia.org/wiki/Feature_selection
   - Drop weak features and select features with strong explanatory power

Motivation, Goal and Selection Criteria

Wikipedia names three main motivations of dimensionality reduction: Reduce costs (time and storage), reduce multicollinearity and provide visualisation oportunities. In the following the goals and selection criteria for mentioned methods are discussed:

1. MDS
   - Goal: Visualize your dataset in a 2-dimensional space and to show the similarities between objects(observations).
   - Alternative to Factor Analysis: Focus on dissimilarities (distances) between objects rather than similarities between features (via correlation matrices).
2. PCA
   - Goal: Dimension reduction and eliminitaion of correlation between features. (*Bishop 2006*: Dimensionality reduction, lossy data compression, feature extraction and data visualitation.)
   - Projects d features on m features with $m < d$.
3. Feature Selection
   - Goal: Reduce overfitting, improve generalization and speed up computation.
   -

Multidimensional Scaling (MDS)

Classical MDS

To build up the 2-dimensional matrix $D = [d_{i,j}]$, you need to compute the single components of the features x and y. Therefore, you use the subsequent formula according to Wikipedia:

$$d_{i,j} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$$

Lecture Slide 76 gives the formula for $d - dimensions$:

$$d(x_i, x_j)^2 = \sum_{k=1}^{d} (x_{i,k} - x_{j,k})^2$$

Not found on slides In order to evaluate, whether the MDS provides a useful dimensionality reduction, one need to examine the Stress/Strain (refering to the information loss) of the reduction.

Metric MDS

$$Stress_D(x_1, x_2, ..., x_N) = \left(\frac{\sum_{i,j} (d_{i,j} - \|x_i - x_j\|)^2}{\sum_{i,j} d_{i,j}^2} \right)^{1/2}$$

Non-metric MDS

$$Stress = \sqrt{\frac{\sum (f(x) - d)^2}{\sum d^2}}$$

Principal Components Analysis (PCA) - Slides 81-88

PCA projects $x \in \mathbb{R}^d$ to $z \in \mathbb{R}^m$. When $\|w\| = 1$ holds, the following transformation can be applied:

$$z = w^T x$$

The goal is to minimize the information loss in the new m-dimensional projection. Let A be the N x m matrix with N observations and m features. **e** is an all-ones vector of length N. X can then be defined as:

$$X = A + ee^T$$

From which follows Σ as the covariance matrix with m x m dimensions

$$\Sigma = cov(X) = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x}) * (x_i - \bar{x})^T$$

Covariation Matrix

$$\Sigma^{(m)} = E\Bigg[(X^{(m)} - E[X^{(m)}])(X^{(m)} - E[X^{(m)}])^T\Bigg]$$

$$\Sigma^{(m)} = E\Bigg[\big(X^{(m)} - \frac{1}{N}\sum_{i=1}^{N} X_i^{(m)}\big)\big(X^{(m)} - \frac{1}{N}\sum_{i=1}^{N} X_i^{(m)}\big)^T\Bigg]$$

$$\Sigma^{(m)} = \sum_{i=1}^{N} \frac{1}{N}\Bigg[(x_i^{(m)} - \bar{x}^{(m)})(x_i^{(m)} - \bar{x}^{(m)})^T\Bigg]$$

Just Stuff and remaining formulas

The covariance matrix Σ is calculated with

$$\Sigma^{(d)} = cov(X^{(d)}) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x}) * (x_i - \bar{x})^T$$

$$\Sigma_{(1)} = cov(X_{(1)}) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x}) * (x_i - \bar{x})^T$$

Feature Selection

Overview (Slide 89/90)

1. Evaluate the ‘quality’ of a feature.
2. Find best subset of features using the quality criterium.
    - highest quality of featires
    - Lowest multicollinearity in subset

Use greddy algorithm with either bottom-up or top-down approach:

- bottom-up: Start with no feature and add new feature with every iteration. Evaluation via an error function and repeated until the required dimension is achieved.
- top-down: Start with all feature and remove the least explanatory with each step under consideration of an error function and until the required dimension is achieved.

Outlier and Duplicates

Outlier (Slide 101/102)

- A boxplot can reveal outlier.
- Also mean and standard deviation (sd) can be used. Usually outlier can be identified by $mean \pm sd$.
- Use cluster, every observation, which does not belong to a cluster, is a outlier.
- Parametrized processes, for instance mixture of gaussians.

Duplicates (Slide 103)

Use distances to find duplicates or observations which are very similar (very small distance). Reduce the number of observations and only keep one of them.

Technical Implementation

R

Python (sklearn - package)

Useful function/package

sklearn.manifold.MDS(n_components=2, metric=True, n_init=4, max_iter=300, verbose=0, eps=0.001, n_jobs=1, random_state=None, dissimilarity='euclidean')

http://scikit-learn.org/stable/modules/generated/sklearn.manifold.MDS.html

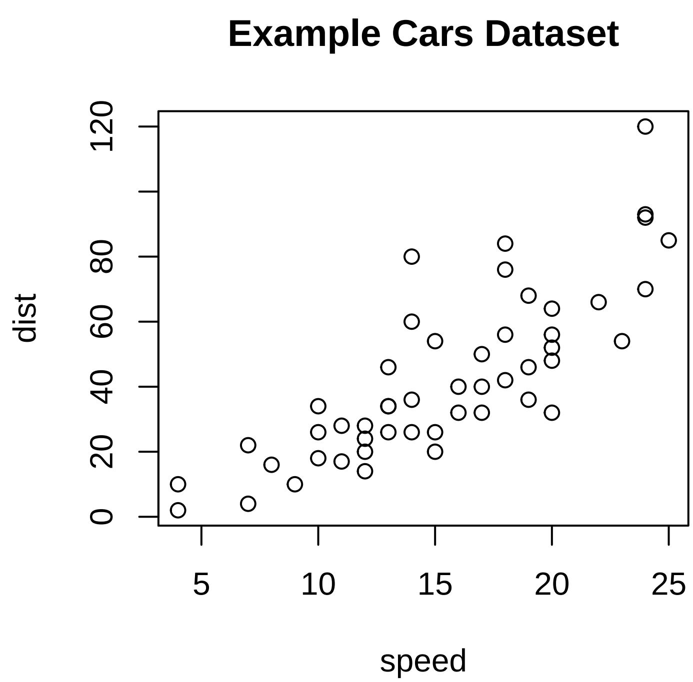
Appendix: Code examples

Example: Table

Table 1: Example Summary Cars Dataset

speed	dist
Min. : 4.0	Min. : 2.00
1st Qu.:12.0	1st Qu.: 26.00
Median :15.0	Median : 36.00
Mean :15.4	Mean : 42.98
3rd Qu.:19.0	3rd Qu.: 56.00
Max. :25.0	Max. :120.00

Example: Plot



Appendix: Data Sample

```
# Code for Data Sample

Mirror the code here. It is executed in a separate document.
```

Appendix: Literature Sources

- Lecture Slides from Machine Learning, Dr. Thomas Fober
- Bishop, C. (2007). *Pattern Recognition and Machine Learning* (Information Science and Statistics), 1st edn. 2006. corr. 2nd printing edn. Springer, New York.
- Wikipedia
  - https://en.wikipedia.org/wiki/Multidimensional_scaling
  - https://en.wikipedia.org/wiki/Principal_component_analysis
  - https://en.wikipedia.org/wiki/Feature_selection
- STUFF
  - STUFF