Ludwig-Maximilians-Universität München
Institut für Informatik
Dr. Eirini Ntoutsi
PD Dr. Matthias Schubert

Knowledge Discovery in Databases II

WS 2015/2016

Übungsblatt 2: Feature Reduction

Aufgabe 2-1 Subspace Selection by Inconsistency

Determine the most informative subspace using Branch-and-Bound in combination with the inconsistency criterium.

ID	attribute X	attribute Y	attribute Z	class
A	2	red	yes	1
B	3	red	yes	1
C	3	green	yes	1
D	4	green	yes	2
E	1	red	yes	2
F	1	green	yes	2

Aufgabe 2-2 Potential of inconsistencies in different domains

Given attributes $A_i \in \mathbb{N}$, attributes $B_i \in \{\text{red}, \text{green}, \text{blue}\}$, and attributes $C_i \in 0, 1$.

Is it possible for all n elements in a data set to be mutually distinct, when considering a feature space consisting of the following attributes:

- A_1
- B_1
- C_1
- $C_1 \times C_2 \times C_3$
- $B_1 \times C_2$
- $B_i^k \times C_j^l$
- $B_1 \times C_2 \times A_3$

Aufgabe 2-3 Image Compression with RCA and PCA

In this exercise you will try out the application of PCA and SVD to compress image contents. To help you with the implementation download the template *PCA_template.py* from the homepage. The template reads an image and compresses it using the methods *pca_decomposition* and *svd_decomposition*. Both methods receive a grayscale image given as a numpy matrix, a number of target dimensions *vals* and return a reconstruction of the image based on the $vals$ most variant inner dimensions.

(a) Implement the method *svd_decomposition* using the method svd from the $numpy.linalg$ package. The methods should use the input image as matrix, should decompose it, should delete the $d - vals$ smallest dimensions and should rebuild a reconstruction of img based on the $vals$ strongest singular values.

(b) Implement the method *pca_decomposition* using the method $eigh$ from the $numpy.linalg$ package. The methods should use the input image as matrix, should apply PCA, should delete the $d - vals$ dimensions with small eigenvalues and should rebuild a reconstruction of img based on the $vals$ strongest eigenvalues.

(c) Modify a template to check whether both methods come to the same result?

Aufgabe 2-4 Relevant Component Analysis

In this exercise, you will compare the result of PCA to the result of RCA as preprocessing step for a kNN classifier. To help you with the implementation download the template *RCA_template.py* from the homepage. The template reads an ARFF-file and performs a cross-validation test for a kNN classifier on the original data. Additionally, it applies PCA and RCA to the data set for all subspaces.

(a) Implement the method pca to reduce the data set to its i principal components.

(b) Implement the method rca to reduce the data set to its i most separating dimensions.

(c) Run the template and compare the resulting classification accuracies.