Practice Four: Data Preparation Practice 4
Singular Value Decomposition (SVD) with RapidMiner

Learning Outcomes:

✓ Acquiring hands-on on singular value decomposition (SVD) on a real dataset
✓ Getting Started with SVD using Sonar Dataset
✓ Evaluating SVD’s reduction impact using Naïve Bayes and Cross Validation on Spambase Dataset
✓ Testing the PA model using user generated test data
Introduction

RapidMiner offers an operator that can perform dimensionality reduction that is based on Singular Value Decomposition (SVD). The dimension can be reduced either by (1) specifying the required reduced number of dimensions or (2) by specifying the cumulative variance threshold. Once you specify the cumulative variance threshold all columns above the threshold in the resulted matrix will be removed.

For more information about singular value decomposition, please visit Chapter Three from the Predictive Analytics class’ lecture notes.

SVD can be applied to extract the number of important dimensions in your dataset. Through knowing the number of important dimensions in your dataset, you will be able to remove unnecessary features that could be linearly dependent, correlated, redundant or of boring distribution.
Quick Start on Dimensionality Reduction of the Sonar Dataset using SVD

The 'Sonar' data set is loaded using the Retrieve operator. A breakpoint is inserted here so that you can have a look at the ExampleSet. You can see that the ExampleSet has 60 attributes. The Singular Value Decomposition operator is applied on the 'Sonar' data set.

The dimensionality reduction parameter is set to 'fixed number' and the dimensions parameter is set to 10. Thus the resultant ExampleSet will be composed of 10 dimensions (artificial attributes). You can see the resultant ExampleSet in the Results Workspace and verify that it has only 10 attributes. Please note that these attributes are not original attributes of the 'Sonar' data set. These attributes were created using the SVD procedure.

Parameters

- dimensionality_reduction
  This parameter indicates which type of dimensionality reduction (reduction in number of attributes) should be applied.
  - none: if this option is selected, dimensionality reduction is not performed.
  - keep_percentage: if this option is selected, all the components with a cumulative variance greater than the given threshold are removed from the ExampleSet. The threshold is specified by the percentage_threshold parameter.
  - fixed_number: if this option is selected, only a fixed number of components are kept. The number of components to keep is specified by the dimensions parameter.

- percentage_threshold
  This parameter is only available when the dimensionality_reduction parameter is set to 'keep percentage'. All the components with a cumulative variance greater than the percentage_threshold are removed from the ExampleSet.

- dimensions
  This parameter is only available when the dimensionality_reduction parameter is set to 'fixed number'. The number of components to keep is specified by the dimensions parameter.

RapidMiner GmbH 2015. All rights reserved.
Dimensionality Reduction of Spambase dataset from UCI

For this tutorial you will be using Spambase dataset from UCI.
https://archive.ics.uci.edu/ml/datasets/Spambase

The data contains 4601 observations, with 57 real-valued observations. Each observation is tagged with a binary label, indicating if it is spam (1) or not (0). Note that despite stating in the description that there are missing values, you may not see any in the downloaded data set.

1. Download the dataset and retrieve it on RapidMiner
2. Apply SVD
3. Building the validation process with Naïve Bayes as the PA model (as shown in the process above)
4. Report the accuracy of the model on at least 4 variations on SVD parameters (variances (e.g 70%, 80%...), number of reduced dimensions...) Show the confusion matrix and highlight the best accuracy.
5. Prepare test data with same attributes as the original data (you can pick some random rows from the original dataset) and feed it to the process to predict the nature of the email (Spam vs. non-spam)
6. Report your observations and comments