Instituto Superior de Engenharia de Lisboa
Mestrado em Engenharia Informática e de Computadores
Mestrado em Engenharia Informática e Multimédia
Big data mining (MDLE)

Laboratory Class #2 — Data Representation and Dimensionality Reduction
2^{nd} semester, 2022/2023 (March, 29)

Code and Report about the highlighted blue text questions/comments are due by April, 19

PART I. MATERIALS AND METHODS

1. **Datasets**
For this laboratory class, we will consider two datasets, as described in Table 1. More details on the datasets are available at:

   - `https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database`
   - `https://archive.ics.uci.edu/ml/datasets/Influenza+outbreak+event+prediction+via+Twitter+data`

Table 1: Datasets with d features, c classes, n instances and the corresponding problem/task to solve.

Dataset	d	c	n	**Problem/Task**
Diabetes (Pima)	8	2	768	Detect if a patient shows signs of diabetes
Influenza Outbreak	545	2	1095 (train) + 485 (test)	To predict an influenza outbreak

2. **Software Tools**
In this laboratory class, we will use the following tools to explore the concepts lectured in this module, for low and medium-dimensional datasets:

   (i) Part II - Waikato Environment for Knowledge Analysis (WEKA), version 3.8, available at `http://www.cs.waikato.ac.nz/ml/weka`
   (ii) Part III - Orange data mining, version 3.34, available at `https://orangedatamining.com`

   In Part IV, we will also explore programming with R, using R Studio, which together with Spark scales for high-dimensional (big) data.

3. **Laboratory Class Setup/Preparation - Software Tools**
As a preparation to the lab, please perform the following actions:

   (i) Install the WEKA GUI, available at `https://www.cs.waikato.ac.nz/ml/weka/downloading.html`, and check for its proper functioning. You may use some datasets, available on the WEKA installation folder `C:\Program Files\Weka-3-8-4\data`
   (ii) Install Orange data mining software, available at `https://orangedatamining.com`, and check for its proper functioning.

4. **Laboratory Class Setup/Preparation - Data Inspection and Datasets Available on the Web**
As a preparation to the lab, please perform the following actions:

   (i) For the datasets described in Table 1, identify the number of features with: integer values; real values.
   (ii) Take a look at some of the public domain dataset repositories available on the Web
      (1) University of California at Irvine (UCI) `https://archive.ics.uci.edu/ml/datasets/`
      (2) Knowledge Extraction based on Evolutionary Learning (KEEL) repository, available at `https://sci2s.ugr.es/keel/datasets.php`
      (3) Science Data Bank, `https://www.scidb.cn/en`
      (4) Lx Data Lab, `https://lisboainteligente.cm-lisboa.pt/lxdatalab`

PART II. WEKA

1. **Data representation formats and data inspection and visualization tools**

   (a) Run the WEKA GUI Chooser and select the Explorer application, located at the top of the `Applications` tab. Load the Diabetes dataset (diabetes.arff file located on the `data` subfolder) into memory using the `Open File` button.
   Identify the list of supported file formats/types for datasets.

   (b) Check the functionalities available under the `Preprocess` tab. Apply these functionalities to:

   (i) Export the Ionosphere dataset into files of the CSV, JSON, XRFF, and XRFF.GZ formats.
   Describe the key aspects of the data representation for each of these formats.

   (ii) For the three datasets, analyze their key parameters as well as some statistics of its features (attributes).
   Identify: the number of patterns, features, classes, and the number of patterns per class; the feature with the largest number of distinct values; the feature with the largest variance.

   (c) Check the functionalities available under the `Visualize` tab. With the Diabetes dataset, use these functionalities to compute the scatter plot: of feature 2 against the class label; of feature 5 against the class label.
   Show a screen-shot of each scatter plot. What seems to be more relevant: feature 2 or feature 5? Justify your answer.

2. **Dimensionality reduction and evaluation**

   (a) Check the functionalities available under the `Select attributes` tab.
   Identify the techniques available under the "Attribute Evaluator", "Search Method", and "Attribute Selection Mode" tabs.

   (b) With one dataset at your choice, apply the `CfsSubsetEval` method with `Best First` search using the `Full training set` option. Repeat this process 5 times.
   State the number of features selected by the method and identify the subset of the selected features. The chosen subset is always the same? Why or why not?

   (c) Repeat (b) using the `Cross-validation` option with 10-folds, to check for the consequences of data sampling.
   State the number of features selected by the method and identify the subset of the selected features. The chosen subset is always the same? Why or why not?

   (d) With one dataset at your choice, check the functionalities available under the `Classify` tab. Learn a decision tree classifier (`J48` algorithm) using the 10-fold cross-validation procedure.
   Show the details/rules of the learned tree and report its classification accuracy. Show the Java source code of the classifier. What is the most relevant attribute/feature?

PART III. ORANGE

1. **Orange environment and analysis of existing examples**

   (a) Run the Orange application and select the `Examples` (Example Workflows) option to see some examples and demos of the use of the software, as depicted in Figure 1.

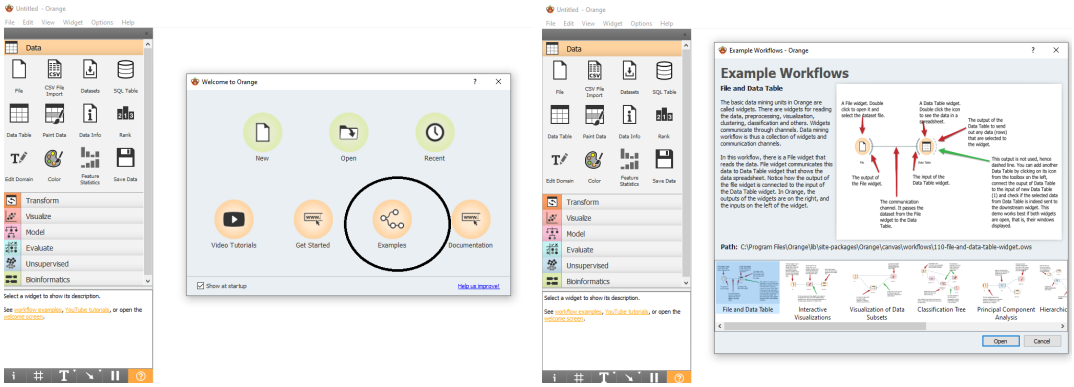

Figure 1: Orange software and its example workflows

   (b) Run the `File and Data Table` example and check its key functionalities, using the (default) Iris dataset. Add a `Feature Statistics` Widget and analyze the four features in the dataset.
   Identify the list of supported file formats/types for datasets. Show a screen-shot of the statistical analysis.

   (c) Run the `Interactive Visualizations` example. Add a `Rank` widget and report the Information Gain (IG) and FCBF relevance measures for all the features.
   Explain the purpose of this example. Identify the information provided by the Data Info widget. What are the most relevant features with the IG and FCBF criteria?

2. **Feature ranking and selection**

   (a) Run the `Feature Ranking` example (also available on the Web, `https://orangedatamining.com/workflows/Feature-Selection`) with the Iris dataset. On the `Rank` widget, try all the available scoring methods and look for the most relevant feature.
   What seems to be the most relevant feature?

   (b) On the `Feature Ranking` example with the Iris dataset, use the `Scatter Plot` widget to identify the most relevant feature.
   Show some screen-shots that of your analysis to find the most relevant feature and justify your answer.

3. **Feature reduction with principal component analysis and discretization**

   (a) Run the `Principal Component Analysis` example, `https://orangedatamining.com/widget-catalog/unsupervised/PCA`, with the default Brown-Selected dataset.
   Explain the key actions of this demo and find an adequate number of reduced dimensions.

   (b) Modify the example to discretize the data with the EFD technique, in the reduced dimensionality space. Save the discretized data into a file.
   Show the Orange widget that performs these actions as well as the resulting file.

PART IV. R STUDIO

1. **Feature Selection**
   For both datasets:

   (a) Compute the (unsupervised) relevance of each feature, using variance and mean-median, as the relevance measures.
   Plot the sorted relevance values in decreasing order. Comment on the resulting plot. Compare on the smallest and the largest relevance value.

   (b) For the relevance values found in (a), compute an adequate number of features, m, by the cumulative sorted relevance criterion, with three different thresholds.
   State the value of the considered thresholds as well as the corresponding values of m.

2. **Feature Reduction**
   For both datasets:

   (a) Compute the PCA decomposition.
   Plot the corresponding eigenvalues sorted in decreasing order. What would be an adequate number of reduced dimensions, m, for this dataset?

   (b) Compute the SVD decomposition.
   Plot the corresponding singular values sorted in decreasing order. What would be an adequate number of reduced dimensions, m, for this dataset?

   (c) Using the decomposition results of (a) and (b), compute the dimensionality-reduced versions of both datasets.
   Explain how you perform the dimensionality reduction. State the number of features of the reduced datasets.

3. **Feature Discretization**
   For one dataset of your choice, compute a discretized version with one unsupervised technique and with another supervised technique, at your choice.
   State the chosen discretization technique as well as the number of discretization intervals for each feature. Comment on the results.