Department of Mathematics
University of Calabria

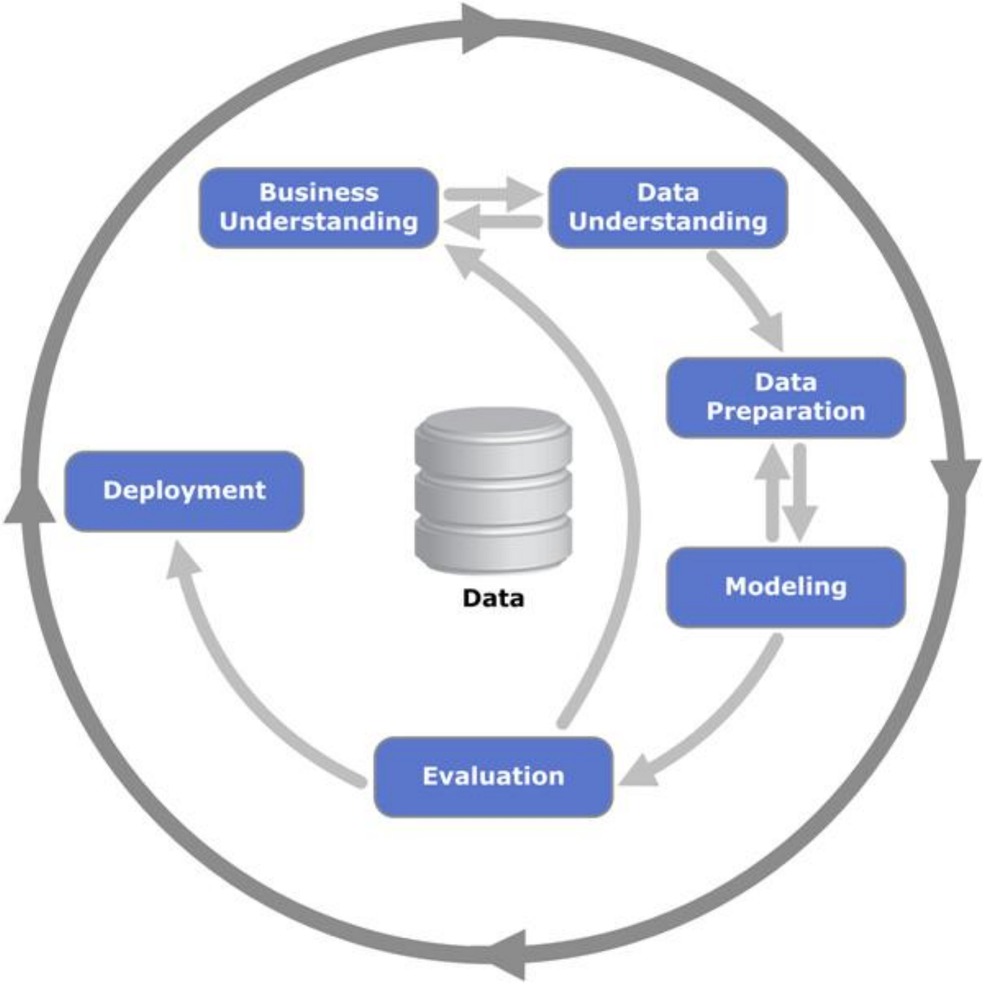
Data Warehouse and Data Mining

Module II – Data Mining

Study case – Image segmentation

Ph.D. Ettore Ritacco

CRISP-DM


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Data

About this lesson

- Exam simulation:
  - You got a data set with data explanation
  - You got a goal (Business Understanding phase)
  - Proceed with the CRISP-DM Methodology

Business Understanding

- We have a dataset whose instances were drawn randomly from a database of 7 outdoor images.
- The images were handsegmented to create a classification for every pixel.
- Each instance is a 3x3 region.
- You have to build a mining model for classifying the instances into the 7 outdoor image classes

Data Understanding

- Image data described by high-level numeric-valued attributes, 7 classes
- Data Set Characteristics: Multivariate
- Number of Instances: 2310
- Attribute Characteristics: Real
- Number of Attributes: 20
- Missing Values? No

Data Understanding

- Attribute Information (1/2):
  1. region-centroid-col: the column of the center pixel of the region.
  2. region-centroid-row: the row of the center pixel of the region.
  3. region-pixel-count: the number of pixels in a region = 9.
  4. short-line-density-5: the results of a line extractoin algorithm that counts how many lines of length 5 (any orientation) with low contrast, less than or equal to 5, go through the region.
  5. short-line-density-2: same as short-line-density-5 but counts lines of high contrast, greater than 5.
  6. vedge-mean: measure the contrast of horizontally adjacent pixels in the region. There are 6, the mean and standard deviation are given. This attribute is used as a vertical edge detector.
  7. vegde-sd: (see 6, 1/2)
  8. hedge-mean: measures the contrast of vertically adjacent pixels. Used for horizontal line detection.
  9. hedge-sd: (see 8, 1/2).
  10. intensity-mean: the average over the region of (R + G + B)/3

Data Understanding

- Attribute information (2/2):
  1. rawred-mean: the average over the region of the R value.
  2. rawblue-mean: the average over the region of the B value.
  3. rawgreen-mean: the average over the region of the G value.
  4. exred-mean: measure the excess red: (2R - (G + B))
  5. exblue-mean: measure the excess blue: (2B - (G + R))
  6. exgreen-mean: measure the excess green: (2G - (R + B))
  7. value-mean: 3-d nonlinear transformation of RGB. (Algorithm can be found in Foley and VanDam, Fundamentals of Interactive Computer Graphics)
  8. saturatoin-mean: (see 7, 2/2)
  9. hue-mean: (see 7, 2/2)
  10. class: target attribute {brickface, sky, foliage, cement, window, path, grass.}

Data Understanding

- Now… it's up to you…