Data Processes: second assignment

Assignment details

The goal of the final assignment is to create a full cognitive system by using the dataset provided and performing all the needed steps with it (descriptive analysis, data exploration, data preparation, model learning and model evaluation).

You are asked to deliver for the assignment the following files:

- Technical report, including the following sections:
  - Scope.
  - Descriptive analysis.
  - Data exploration.
  - Data preparation.
  - Model learning.
  - Model evaluation.
- Any code that you might have used for procedures regarding data management (data preparation, cleaning, integration, etc.) and modelling.
- A README document where you explain what files you delivered and the description of each one.

A ZIP file that contains all these files should be uploaded to the Moodle task for the second assignment. The deadline to upload this file is January 8th, 23:55.

The assignment must be done in groups of 4 people.

Problem definition

The Health Analytics department from a very well-known hospital located in a European country has started a project to improve the results of the diagnosis related to colonoscopies. The project consists of highlighting certain areas of video frames of a colonoscopy that might be considered more relevant for colorectal illnesses due to the texture of the image in them.

To achieve this, the department has extracted some information about the encoded video stream in blocks of 64x64 pixels and they have labelled each block as relevant or not relevant for performing a thorough examination of that part of the frame. It has been possible to obtain 26 features from 16,000 different blocks with their associated class labels.

The system to be developed should aim to classify a frame block as relevant or not relevant in the colonoscopy. We are also asked to identify the most important variables that influence in a block for being relevant or not as, well as any other useful information from the analysis of the dataset.

In the annex is included a codebook that describes the variables contained in the dataset and that would allow to have a basic understanding of the data.

Codebook

quality: a measure of the quality of the recorded video.

bits: number of bits used to encode that block in the video stream.

intra_parts: number sub-blocks inside this block that are not encoded by making use of information in other frames.

skip_parts: number sub-blocks inside this block that are straight-forward copied from another frame.

inter_16x16_parts: number of sub-blocks inside this block making use of information in other frames and whose size is 16x16 pixels.

inter_4x4_parts: number of sub-blocks inside this block making use of information in other frames and whose size is 4x4 pixels.

inter_other_parts: number of sub-blocks inside this block making use of information in other frames and whose size is different from 16x16 and 4x4 pixels.

non_zero_pixels: number of pixels different from 0 after encoding the block.

frame_width: the width of the video frame in pixels.

frame_height: the height of the video frame in pixels.

movement_level: a measure of the level of movement of this frame with respect the previous one.

mean: mean of the pixels of the encoded block.

sub_mean_1: mean of the pixels contained in the first 32x32 sub-bock of the current block.

sub_mean_2: mean of the pixels contained in the second 32x32 sub-bock of the current block.

sub_mean_3: mean of the pixels contained in the third 32x32 sub-bock of the current block.

sub_mean_4: mean of the pixels contained in the fourth 32x32 sub-bock of the current block.

var_sub_blocks: variance of the four previous values.

sobel_h: mean of the pixels of the encoded block after applying the Sobel operator in horizontal direction.

sobel_v: mean of the pixels of the encoded block after applying the Sobel operator in vertical direction.

variance: variance of the pixels of the encoded block.

block_movement_h: a measure of the movement of the current block in the horizontal direction.

block_movement_v: a measure of the movement of the current block in the vertical direction.

var_movement_h: a measure of the variance of the movements inside the current block in the horizontal direction.

var_movement_v: a measure of the variance of the movements inside the current block in the vertical direction.

cost_1: a measure of the cost of encoding this block without partitioning it.

cost_2: a measure of the cost of encoding this block without partitioning it and without considering any movement in it.

relevant: the target variable that indicates whether the current block is relevant (1) or not (0).