CS 584 - Data Mining - Fall19 - Logistics - September 5

Harry Wechsler <wechsler@gmu.edu>
Thu 9/5/2019 10:17 AM
To: Harry Wechsler <wechsler@gmu.edu>
Cc: Priya Mani <pmani@masonlive.gmu.edu>

NO LATE HOMEWORK

HW (Data Mining # 1) Due October 24.

DM 1.1

Download Pen digits dataset from UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets/Pen-Based+Recognition+of+Handwritten+Digits to classify hand-written digits into 10 classes corresponding to digits 0-9. Each instance of data consists of 16 attributes, which are the (x,y) co-ordinates of 8 sampled points on a pen's trajectory while drawing a digit.

Use the training data to build a classifier of your choice and report accuracy, precision and recall. Evaluate the model by 5-fold cross-validation. Use the **Model** to predict labels on the test set and report test accuracy, precision and recall.

Use *Data Mining Architecture* including pre-processing, feature selection, Learning / classification method / 5-fold cross-validation / **Model** Selection, and metrics / Prediction / performance evaluation

DM 1.2

Apply the **Model** learned from Pen digits data set to predict labels on a sample of MNIST data (http://yann.lecun.com/exdb/mnist/). MNIST consists of gray-scale images of hand-written digits in a 28-by-28 box. Each image is represented as a vector of pixels [0-784], columns 0 – 783 for pixel values and column 784 for digit label, and the gray-scale values are in the range [0-255]. The csv format of data can be downloaded from https://www.kaggle.com/oddrationale/mnist-in-csv. The images can be visualized using simple code to plot the pixels.

Sample 50 instances from each class and predict on this data.
Construct 16 (x,y) co-ordinate features for the data. For each image, sample 8 regularly spaced (x,y) co-ordinates along the trajectory of a digit.
Apply the classifier learned with Pen to predict on MNIST data using the 16 co-ordinate for features.
Report accuracy, precision and recall.

HINT: The co-ordinates have to be normalized so that they are comparable across the dataset's bounding box.

Upload your homework on Blackboard on or before October 24 at 3:00 pm.
Including short report that details the architecture used and the particulars of your work including software, hardware, and time used to train and test. // DM 1.2 (NIST): Data Capture .. Performance Evaluation.

Prepare Progress Report and have it checked and vetted / signed by TA by October 8.

//

MIDTERM ~ November 7 ~ COVERS Chapters 1 – 3 (see textbook) and class discussion / material (see emails)
90 minutes ~ OPEN BOOKS

//

Term (Team) Project (Data Mining #2) ~ Progress Report ~ October 1

Team Size ~ 2
Prepare Progress Report and have it checked and vetted / signed by TA by October 1.
Progress Report includes team composition, data mining problem / task, data repositories (source, size, features)
(full-fledged) Architecture, Software / Hardware, and future milestones (including schedule and time table).