Math 6350 fall 2019 MSDS Robert Azencott

Homework 2

Due date Thursday sept 26 at Midnight

Data Set Information:

Go to the site of

" university of california irvine repository of machine learning datasets "

which contains more that 400 data sets for machine learning

The data set to be used for HW2 consists of digitized images of typed characters

the font choices for these characters is quite large (153 fonts).

Download the fonts.zip file from the following link

https://archive.ics.uci.edu/ml/machine-learning-databases/00417/

fonts.zip contains 153 files in format .csv (= comma delimited files), **one file for each font**

for instance AGENCY.csv, ..., TIMES.csv, etc

Each such XYZ.csv file has a "header row" (= top row) which lists the names of each column, ("attribute names")

EXTRACT from font.zip the three font files

COURIER.csv, CALIBRI.csv, TIMES.csv

Description of a typical table such as TIMES.csv :

the font type "TIMES" is the same throughout the file of TIMES.csv

after the header row , each row of the TIMES.csv file describes one single "example"

there are 12530 such rows = 12530 examples of characters typed in the font TIMES

the number of examples of characters will be **different for each font type**

each "example" corresponds to a digitized image of some specific character (letter or digit essentially) typed in the font TIMES

each image has size 20 x20 pixels

each one of these 400 pixels has a "gray level" which is an integer value between 0 and 255

in each row, the image associated to this row has 400 features (or numerical descriptors) whose values are listed in the 400 columns named r0c0, r0c1,r0c2, ... , r19c18, r19c19

these 400 values represent the gray levels of the 400 pixels of this specific image

"rLcK"= gray level image intensity for pixel in position{Row L, Column K}

the **m_label** column is an arbitrary integer from 33 to 65535 ; in each row this m-label integer is the standardized ID of the character present in the image associated to this row;

the **strength** column lists values either equal to 0.4 or to 0.7; in each row , strength =0.4 for NORMAL character ; strength= 0.7 for BOLD character;

the **italic** column lists values either equal to 0 or to 1; in each row , italic= 1 for ITALIC character ; italic= 0 for NORMAL character ;

preliminary treatment of the data set

in each font file, the header row begins with 12 names

font, fontVariant, m_label, strength, italic, orientation, m_top, m_left, originalH, originalW, h, w

DISCARD the 9 columns listed below :

fontVariant, m_label, orientation, m_top, m_left, originalH, originalW, h, w

KEEP the 3 columns {font, strength, italic}

KEEP the 400 columns named r0c0, r0c1,r0c2, ... , r19c18, r19c19

any row containing missing numerical data will be discarded

define then three CLASSES of images of "normal" characters as follows

CL1 = all rows of COURIER.csv file for which {row # >1 and strength = 0.4 and italic=0}

CL2 = all rows of CALIBRI.csv file for which {row # >1 and strength = 0.4 and italic=0}

CL3 = all rows of TIME.csv file for which {row # >1 and strength = 0.4 and italic=0}

Display their respective sizes n1, n2, n3

The full data set (denoted DATA) for the next questions will be the union of the three classes CL1 , CL2, CL3 and hence has size N= n1 +n2 +n3

Example # i in DATA corresponds to a specific row "i" in the matrix DATA , and will be described by this vector of 400 **features** , namely the 400 numbers listed in row "i"

Recall that these 400 feature values observed for example #i are also viewed as the 400 values taken for this particular example by the random variables

X1 = r0c0, X2 = r0c1, X3 = r0c2, ... , X399 = r19c18, X400 = r19c19

Each such random variable Xj is observed N times, and its N observed values are listed in the column "j" of DATA.

We will use the basic machine learning tools covered in class sofar to attempt a rough automatic classification of DATA into 3 classes CL1 CL2 CL3

PART 0

Compute the means m1 = mean(X1) ....mean(X400) = m400 and the standard deviations

s1 = std(X1) ....s400 = std(X400)

Standardize the features matrix DATA by centering and rescaling each random variable Xj into a new random variable Yj= (Xj - mj) /sj; the matrix DATA becomes a standardized data matrix SDATA, with coefficients given by

SDATA(i,j) = (DATA(i,j) - mj)/sj

SDATA is also called a rescaled data matrix for short. The example # i associated to row "i" of DATA will be from now on described by the vector " Ei "of standardized features defined by row"i" of SDATA

PART 1

1.1) Compute the correlation matrix COR of the 400 random variables Y1,..., Y400

1.2) For the matrix COR , compute its 400 eigenvalues $\lambda 1 > \lambda 2 > ... > \lambda 400 > 0$, and its

400 eigenvectors v1, v2, ..., v400

1.3) Plot the decreasing curve λj versus j for j=1 , 2, ..., 400

1.4) for j =1 ,2, ... ,400 compute the successive percentages Rj given by

$Rj = (\lambda 1 + \lambda 2 + ... + \lambda j)/400$

1.5) Plot the increasing curve Rj versus j for j=1 , 2, ..., 400 and compute the smallest

integer "r" such that Rr >90%

1.6) Explain the relationship between these computations and the PCA analysis of the set DATA

1.7) Implement the PCA analysis of the rescaled data matrix SDATA either in R or in Python, or in Matlab; explain clearly the inputs and the outputs of the pre-existing standard PCA functions you use

1.8) The standardized example # i is described by row "i" of SDATA . After matrix transposition, this row becomes a column vector Ei in R^{400} . Compute the first three "scores" of example "i" by

$scor_1$ (i)= < Ei, v_1> , $scor_2$ (i)= < Ei, v_2> , $scor_3$ (i)= < Ei, v_3>,

These 3 numbers are the coordinates of a 3-dimensional vector Ui in R^3.

The 2 numbers $scor_1$ (i), $scor_2$ (i) define a 2-dimensional vector Wi in R^3. Explain the geometric relationship between Ei, Ui,Wi.

1.9) Display graphically the 2 dimensional scatterplot of all the Wi , i= 1,2,...,N , using 3 colors (1 for each class); for instance red for CL1, blue for CL2 ; green for CL3. Interpret this display visually in terms of separability of the 3 classes

1.10) Display graphically the 3 dimensional scatterplot of all the Ui , i= 1,2,...,N , with the same 3 colors; Interpret visually. To facilitate the visual interpretation generate a similar display but with only the two classes CL1 and CL2. Repeat this operation and visual interpretation for CL1 and CL3, and then for CL2 and CL3

PART 2

2.1) Fix k = 15. Use the standardized data matrix SDATA to apply the k nearest neighbor (kNN) algorithm for the automatic classification of arbitrary examples into one of the three classes CL1 CL2 CL3. Compute the percentage per(15) of correct classifications on the whole data set of N examples

2.2) Repeat the preceding operation for k = 5, 10 , 15, 20, 30, 40, 50 ,100,200; and compute the percentages per(k) of correct classifications on the whole data set of N examples

plot the curve per(k) versus k to try to identify a best range [A < k <B] of values for the integer k

2.3) Repeat the preceding exploration for a few more values of k within the range [A,B]. Conclude by selecting a "best" value k* for the integer k

2.4) Compute and interpret the 3x3 confusion matrix for kNN classification using the "best" k= k*