Math 6350 fall 2019 MSDS Robert Azencott

Homework HW3

Due date Thursday October 10th at Midnight

HW3 Part1 :

Prepare a detailed presentation of all the contents presented in Math6350 class during the lectures on kNN automatic classification. This text has to be typed , with explicit formulas. You can either prepare a .tex file or a .docx file , to be emailed to me as well as a version in .pdf

Make sure that the authors names appear in the name of the file, as well as within the documnt itself

HW3 Part2 : Data Analysis

Data Set Information:

The data set to be used for HW3 is the same data set used for HW2 . After downloading the fonts.zip file from https://archive.ics.uci.edu/ml/machine-learning-databases/00417/ , you had extracted from font.zip the 3 font files COURIER.csv, CALIBRI.csv, TIMES.csv . In these files , each row describes the digitized image of one character (image size = 20 x20 pixels) , and each row provides the 400 original features { r0c0, r0c1,r0c2, ... , r19c18, r19c19 } . The "m_label" in row j provides the ID of the character described by row j; the "strength" value is 0.4 for NORMAL characters and 0.7 for BOLD characters; the "italic" value is 1 for ITALIC characters ; and is 0 for NORMAL characters ;

In HW2 we have extracted 3 CLASSES of images of "normal" characters

CL1 = all normal characters in COURIER.csv

CL2 = all normal characters in CALIBRI.csv

CL3 = all normal characters in TIME.csv

Class Sizes are n1 , n2 n3; data set size : N= n1 + n2 + n3

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

Summary of HW2 notations and HW2computations :

mj = mean(Xj) ....sj = std(Xj) .... standardization Yj= (Xj - mj) /sj ;

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

example #i is described by column vector " Ei "= transpose {row"i" of SDATA}

COR = correlation matrix of Y1,..., Y400

eigenvalues of COR ==> $\lambda 1 > \lambda 2 > ... > \lambda 400 > 0$

400 eigenvectors v1, v2, ..., v400

Rj = proportion of variance explained by first j eigenvectors

Rj = (λ1 + λ2 + ... + λj)/400

HW3 question 1

Compute **a =** smallest integer j such that Rj >35% , and **b =** smallest integer j such that Rj >60%

Fix a training set TRAIN of size NTRA ≃ 80% N and a test set TEST of size NTST≃ 20% N. Explain how you implement the random choice of these 2 sets, to ensure that *within the TEST* set, the sizes m1 m2 m3 of classes CL1 , CL2, CL3 verify mj/NTST ≃ nj /N for j=1,2,3

HW3 question 2

Example # i is originally described by row "i" of SDATA . After matrix transposition, this row becomes a column vector Ei in R^{400} . For each m= 1,2,..., and each i=1,2, the score #m of example #i is defined (and computable) by the formula **$scor_m$ (i)= < Ei, v_m>** . We now describe each example # i from SDATA by the vector Ai ∊ R^a which lists the values of the new features Ai= [$scor_1$ (i), $scor_2$ (i)., ,. $scor_a$ (i)] . Note that dim(Ai) = a. Give a geometric interpretation of Ai in terms of Ei. Fix k = 5 and apply kNN *in the Euclidean space R^a* to implement the automatic classification of all examples in the TEST set , using TRAIN as the training set, and *using the new feature vectors* Ai ∊ R^a . The three classes are CL1, CL2, CL3 , exactly as in HW2. Compute the percentage of successful classifications on TEST and on TRAIN, as well as the confusion matrices on TEST and on TRAIN. Compare to the results already obtained in HW2 for kNN classification with k=5

HW3 question 3

Repeat the preceding automatic classification by kNN with k=5, but based on the vectors Gi ∊ R^{b-a} listing the values of the *(b-a) new features* :

Gi= [$scor_{1+a}$ (i), $scor_{2+a}$ (i)., ,. $scor_b$ (i)] . Note that dim(Gi) = b-a.

Compare these results to the preceding results, and give your interpretation.

HW3 question 4

Use the feature vectors Ai ∊ R^a and apply the unsupervized **Kmean** algorithm in R^a to implement automatic clustering of the TRAIN data into 3 sets H1, H2, H3. Repeat 10 times the implementation of Kmean with different random initializations for the centers of H1 H2 H3. Describe precisely the Cost function Cost(H1,H2,H3) which the Kmean algorithm attempts to minimize. For each implementation of Kmean, compute the terminal value of the Cost(H1,H2,H3). List these 10 terminal Costs and select the clustering result H1 H2 H3 achieving the smallest terminal cost

HW3 question 5

To compare the computed clustering H1 H2 H3 to the "ideal" clustering CL1 CL2 CL3, we first compute Cost(CL1,CL2,CL3) and compare to Cost (H1,H2,H3). To get more concrete information , compute for i=1,2,3 and j=1,2,3 all the percentages

$P_{ij} = \text{size}(H_i \cap CL_j) / \text{size}(CL_j)$ and $Q_{ij} = \text{size}(H_i \cap CL_j) / \text{size}(H_i)$

Interpret these results.