CSC 546/746

Assignment 6

(25 points)

1. (1 point) Create a new Jupyter Notebook project and name it as “hw06.ipynb”.
2. (2 points) Read data from the “hw06_data1.csv” and standardize the data.
3. (5 points) Run the KNN classifier on the dataset and evaluate:
    a. You will use features other than “Target Class” to predict “Target Class”.
    b. Splitting the dataset into the Training set (70%) and Test set (30%).
    c. Use “Euclidean distance” for the KNN model and set the “K” value you choose.
    d. Output the confusion matrix & classification report based on test dataset.
4. (5 points) Use the elbow method to pick a good K Value:
    a. Use Scikit-Learn’s K-fold cross-validation feature to calculate the accuracy rate with different K values. Use 10-fold cross-validation and test K ranging from 1 to 40.
    b. Visualize the result.
    c. What do you think the best range of K values are? Why? (Answer this question in the hw06.ipynb either in a markdown cell or in the comments)
5. Implement a Naïve Bayes model to classify Spam and Ham short messages. (**Hint:** You can follow the lab06 NB model structure to implement this model.)
    a. (2 points) Load data from the “hw06_data2.txt”.

    The dataset is real Short Message Service (SMS) Spam Collection dataset, corpus of mobile SMS labeled Spam/Legitimate messages from the UCI Machine Learning Repository website (https://archive.ics.uci.edu/ml/datasets/SMS+Spam+Collection). Each line has one SMS message including two instances: the first instance indicates label if the SMS message is spam or legitimate (ham), and the second one is the content of the message (i.e., raw text). The instances are separated by tabs (‘\t’).

    b. (1 point) Splitting the dataset into the Training set (80%) and Test set (20%).
    c. (3 points) Setting up the Naïve Bayes model.
    d. (2 points) Train the Naïve Bayes model with the training set.
    e. (2 points) Evaluate the model with the test set. Print out the confusion matrix and classification report. (**Hint:** You don’t need to use Seaborn. Just print out the confusion matrix since we only have 2 categories: Spam and Ham)
    f. (2 points) Analyze the performance of the model based on the report. (Answer this question in the hw06.ipynb either in a markdown cell or in the comments)
6. Submit “hw06.ipynb” to the Blackboard.