Aman Patel

May 1, 2021

CSCI-B 455

Discussed with Taral Shah

Home Assignment 5

1. Problem 1
    a. See attached PDF for A – D
    b. E: LDA is used to find the component axes that maximize between-class variance. While this will maximize class separation, variance in the data is not maximized. To maximize data variance, the principal component can be calculated using PCA.
2. Points: $(1, 1), (1.5, 2), (3, 4), (5, 7), (3.5, 5), (4.5, 5), (3.5, 4.5)$
    a. K-means clustering with initial centroids: $m_1 = (1, 1), m_2 = (5, 7)$
        i. Iteration 1
            1. m_1 points: $\{(1, 1), (1.5, 2)\}$
            2. m_2 points: $\{(3, 4), (5, 7), (3.5, 5), (4.5, 5), (3.5, 4.5)\}$ ($(3, 4)$ is the midpoint of the initial clusters, its cluster is chosen arbitrarily)
            3. $m_1 = \left(\frac{1 + 1.5}{2}, \frac{1 + 2}{2}\right) = (1.25, 1.5)$
            4. $m_2 = \left(\frac{3 + 5 + 3.5 + 4.5 + 3.5}{5}, \frac{4 + 7 + 5 + 5 + 4.5}{5}\right) = (3.9, 5.1)$
        ii. Iteration 2
            1. m_1 points: $\{(1, 1), (1.5, 2)\}$
            2. m_2 points: $\{(3, 4), (5, 7), (3.5, 5), (4.5, 5), (3.5, 4.5)\}$
            3. Points are the same for both clusters, the algorithm has converged.

4. $m_1 = (1.25, 1.5), m_2 = (3.9, 5.1)$

3. SOM and Naïve Bayes to classify intruders.
    a. The SOM training process compares input vectors to the weights within the self-organizing map (SOM). The algorithm finds the neuron whose weights are closest to the input, labelling it as the winning neuron. All other neurons are updated around this neuron, with nearby neurons having a different learning rate than distant neurons. This process is repeated for each data point in the data set.
    b. Naïve Bayes training includes calculating class probabilities and conditional probabilities for each combination of attributes. To find class probabilities, the number of instances of each class is divided by the total number of instances. Conditional probabilities are calculated using the number of instances of a value of an attribute corresponds with a class and the number of instances of the class.
    c. Each value of the input vector would be normalized, with element 3 representing the class of programs most frequently run by the user. The number of nodes in the SOM is typically around $5\sqrt{n}$, where n is the number of data points. To keep the SOM simple, 900 data points can be used, with 150 nodes in the SOM. For the shape of the SOM, hexagonal grids are commonly used because each node in a hexagonal grid has 6 direct neighbors as opposed to 4 when using a rectangular grid.
    d. I believe this method would work well for detecting intruders. However, the input data is not specific to intruders, and many intruders can be misclassified as regular users. I would assume the algorithm would have far more false negatives (predicting regular user when user is not) than false positives because it is easier to identify suspicious activity than it is to identify normal activity.

4. K-means clustering can be used to classify transactions. The algorithm would start with two random transactions, representing the centroids of the classes (fraudulent and legitimate). Each transaction is placed into one of these two groups based on the Euclidean distance between the transactions. The centroids of the clusters are redefined as the mean of each cluster, and the algorithm repeats. It either repeats a fixed number of iterations or it repeats until convergence, which occurs when the clusters are unchanged from one iteration to the next. Once the clusters converge (or the maximum number of iterations is reached), the centroids can be used to predict whether future transactions belong in the fraudulent or legitimate cluster. However, this algorithm may not result in the global optimum. This can be remediated by running the algorithm on a variety of initial centroids and selecting the final centroids that result in the smallest sum of squared error. K-means clustering is effective once the global optimum is found, but it can be misleading if there is an imbalance in the number of instances of each class. For example, if there are 50 examples of Class 1 and only 5 examples of Class 2, the cluster for Class 2 may contain more than 5 data points. This is because the clusters are likely to have some overlap, especially if the classes have similar data. The data can be balanced by removing instances of the more frequent class or by adding more examples of the less frequent class.