Anomaly Detection in Data

Learning Outcomes

Upon successful completion of this workshop, you will have demonstrated the abilities to:

- Applying different anomaly detection techniques
- Compare and see the behaviour of diffrent approaches

Instructions:

1. Read the **tutorial** **(http://www.cse.msu.edu/~ptan/dmbook/tutorials/tutorial9/tutorial9.html)**
2. Download the following data sets (the first two columns are feature and the 3rd column is the class label):
   - **G-data (https://learn.ontariotechu.ca/courses/19275/files/2375095?wrap=1)** ↓ (https://learn.ontariotechu.ca/courses/19275/files/2375095/download?download_frd=1)
   - **compound (https://learn.ontariotechu.ca/courses/19275/files/2375096?wrap=1)** ↓ (https://learn.ontariotechu.ca/courses/19275/files/2375096/download?download_frd=1)
   - **flame (https://learn.ontariotechu.ca/courses/19275/files/2375094?wrap=1)** ↓ (https://learn.ontariotechu.ca/courses/19275/files/2375094/download?download_frd=1)
   - **pathbased (https://learn.ontariotechu.ca/courses/19275/files/2375097?wrap=1)** ↓ (https://learn.ontariotechu.ca/courses/19275/files/2375097/download?download_frd=1)
3. Remove the 3rd column from the compound, flame and pathbased data sets

Part I (Using Parametric Models):

1. For this part use the **G-data** (assume the first column is *x* and the second one is *y*)
2. Use *Mahalanobis* distance between (*x*,*y*) against the mean of *x* and *y* as the anomaly score.
3. Draw an appropriate scatter plot showing the anomaly scores
4. Report the top-5 points that you have detected as the anomaly

Part II (Using Distance-based Models):

1. For this part use **compound**, **flame**, and **pathbased** data sets
2. Use the distance to *k'th* nearest neighbour as the anomaly score (for k =1, 2, 5)
3. Draw appropriate scatter plots showing the anomaly scores
4. Report the top-5 points that you have detected as the anomaly

Part III (Using Density-based Models):

1. For this part use **compound**, **flame**, and **pathbased** data sets
2. Use "relative density" as the anomaly score with the following definition for the density:

    A. Density is the inverse of distance to k’th neighbour (for k = 1, 2, 5)
    B. Density is the inverse of the average distance to k neighbours (for k = 1, 2, 5)

3. Draw appropriate scatter plots showing the anomaly scores
4. Report the top-5 points that you have detected as the anomaly for each method

Report:

1. Your report should have a cover letter including the group member names
2. Organize all your *diagrams* and *interpretations* in your lab report *(PDF format)*
3. Include your code and report in a folder (you can zip the folder) and submit it