Milestone 2 & 3

Dr. Mohamed Ehsan Ashour
Eng. Youssef Tawfilis
Eng. Nour Amin

Milestone Two: Anomaly Detection using Z-Score and Probability Density Function (PDF) Analysis

In this milestone, we will focus on predicting anomalies using z-score thresholds and analyzing feature distributions to enhance the accuracy and reliability of our anomaly detection model. This milestone comprises two primary tasks: evaluating model performance with various metrics and identifying optimal data distributions.

Task 1: Performance Evaluation and Anomaly Prediction

1. Data Preparation

- **Column Selection**: Exclude the `class` column to ensure predictions are based on input features alone.
- **Dataset Division**: Split the dataset into training (70%) and testing (30%) sets for effective model validation.

2. Z-Score Calculation and Anomaly Prediction

- **Z-Score Computation**: Calculate z-scores using the formula:

$$z = \frac{(X - \mu)}{\sigma}$$

  where X is the value, μ is the mean, and σ is the standard deviation.

- **Threshold Tuning**: Experiment with different z-score thresholds to optimize anomaly detection across feature combinations.

3. Performance Metrics

To evaluate the model on the Test set, we will use three key metrics: **accuracy**, **precision**, and **recall**, each providing unique insights for anomaly detection.

- **Accuracy**:

$$\text{Accuracy} = \frac{\text{True Positives (TP)} + \text{True Negatives (TN)}}{\text{Total Number of Samples}}$$

- **Precision**:

$$\text{Precision} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}}$$

- **Recall**:

$$\text{Recall} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}}$$

4. Key Metric Selection

You need to decide which metric is most important for this project, Provide your Reasoning.

Task 2: Distribution Analysis (PDF and PMF)

This task involves identifying the best-fitting probability density functions (PDFs) for numerical features and probability mass functions (PMFs) for categorical features.

1. Numerical Columns PDF Fitting

- For each numerical column, identify the PDF that best fits the data, such as uniform, Pareto, or exponential distributions (not only those, try all possible ons).
- Calculate PDFs for each column alone, conditioned on an anomaly, and conditioned on no anomaly (normal).
- Evaluate the best PDF fit using mean square error (MSE) and select the PDF with the lowest MSE. Reference code for fitting the best PDF is available https://stackoverflow.com/a/37616966.
- Carefully adjust the calculation range for PDF estimation to obtain accurate data fits. (Very Important)

2. Categorical Columns PMF Storage

- Store the probability mass functions (PMFs) calculated in the previous assignment for each categorical column.
- Include conditional PMFs for each column based on anomaly and no anomaly (normal) in a structured format for easy reference.

3. Result Documentation

- Summarize this analysis in a list containing each field name, best-fit PDF, and associated parameters for each numerical column.
- Organize PMF data for each categorical column for future reference.

Deadline for Milestone Two is on 15 November 2024 11:59 PM

Submission is through the following link https://forms.gle/DtfPojDsDd3m656q9

Milestone 3: Naïve Bayes Estimation

Task 1: From Scratch Estimation Across Test Dataset

The Naïve Bayes estimation for predicting whether an anomaly occurs based on a single row of data is expressed using the following equation:

$$\Pr(\text{Anomaly}|\text{row}) =$$

$$\frac{\text{pdf}_{A|\text{Anomaly}}(a) \cdot \text{pdf}_{B|\text{Anomaly}}(b) \cdot \text{pdf}_{C|\text{Anomaly}}(c) \cdots \cdot \text{PMF}_{Q|\text{Anomaly}}(q) \cdot \text{PMF}_{R|\text{Anomaly}}(r) \cdots \cdot \Pr(\text{Anomaly})}{\text{pdf}_{A}(a) \cdot \text{pdf}_{B}(b) \cdots \cdot \text{PMF}_{Q}(q) \cdot \text{PMF}_{R}(r) \cdots}$$

Where:
- $\text{pdf}_{A|\text{Anomaly}}(a)$ is the fitted conditional PDF of column A evaluated at the value for this row.
- $\text{pdf}_{B|\text{Anomaly}}(b)$ is the conditional PDF of column B for the same row.
- You can include as many numerical columns as appropriate for the analysis.
- $\text{PMF}_{Q|\text{Anomaly}}(q)$ and $\text{PMF}_{R|\text{Anomaly}}(r)$ represent the fitted conditional PMFs for categorical columns Q and R, respectively.

Note that the above rule assumes that the columns $A, B, C, \ldots, Q, R$ are independent. The quality of the estimation may degrade if any of these features are dependent.

Additionally, you may use the same equation to calculate the probability of no Anomaly:

$$\Pr(\text{no Anomaly}|\text{row})$$

to assess the quality of your predictions.

The calculation described is performed for each row in the test dataset, which represents individual customer connections. It is preferable for the implementation to allow easy inclusion or removal of columns in the calculation.

1. Repeat the above equation for all rows in the test dataset and compare the results to the actual labels in the test data.
2. Use the three metrics (Accuracy, Precision, Recall) to evaluate the model

Task 2: Categorical Feature Encoding and Machine Learning Model

In this task, we will encode categorical features and evaluate various Naïve Bayes models.

1. Use one-hot encoding to encode all categorical features in the dataset.
2. Utilize the split train-test set to train the following models from Scikit-Learn:
   - Gaussian Naïve Bayes (GaussianNB)
   - Multinomial Naïve Bayes (MultinomialNB)
   - Bernoulli Naïve Bayes (BernoulliNB)
3. For each model, assess performance using the three evaluation metrics: accuracy, precision, and recall. Determine which model provides better results and justify your findings.

Deadline for Milestone Three is on 17 December 2024 11:59 PM

Submission is through the following link https://forms.gle/DtfPojDsDd3m656q9