2025 Data Mining

HW2

Task introduction

- Anomaly Detection
    - The dataset is composed of real-world images with a focus on industrial inspection.
    - Implement machine/deep learning model to do anomaly detection.
    - All packages are available (sklearn, keras, PyTorch etc.).
    - Do not use any pretrained model & Do not use any extra data for training.
- Requirement
    - Upload your submission to Kaggle
    - Submit a report and your source code to E3

- Deadline is 11/5 23:59, no late submission

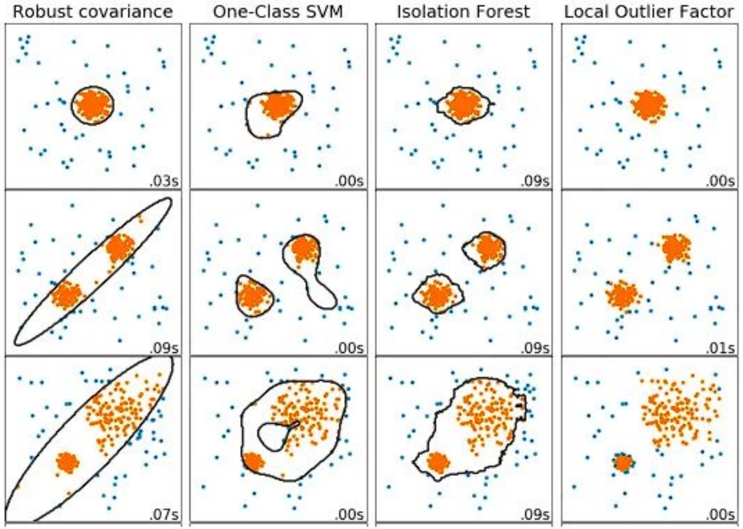
Dataset

The MVTec anomaly detection dataset (MVTec AD)

- Train set
    - Includes 3629 normal images.
- Test set
    - Includes 1725 images.
    - The task is to classify each as either normal or anomalous.

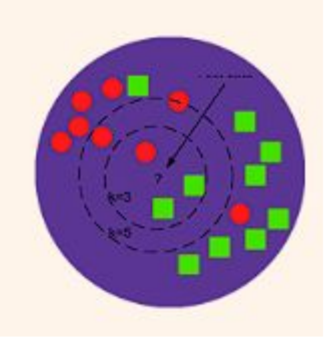
Method 1 - SVM

- Use OneClass SVM to learn a decision boundary.
- Find the suitable kernel space and parameters to fit the data.
- Convert the result of classification to the self-defined value.



Overview of outlier detection methods

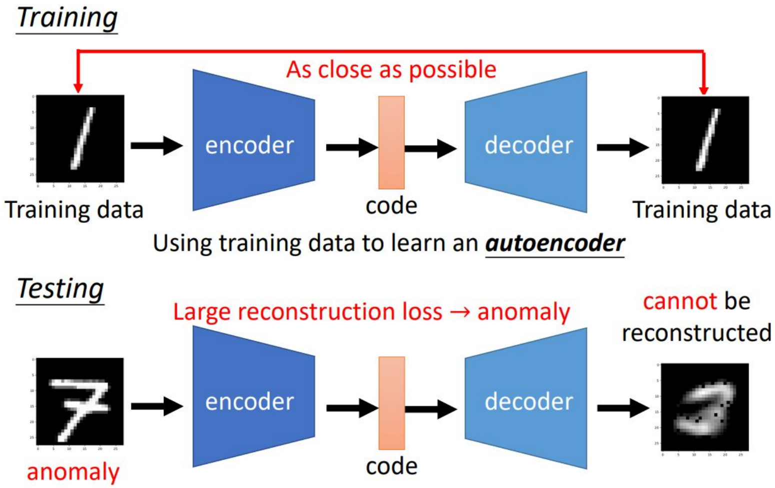
Method 2 - KNN



- Assume that there are n clusters in training data.
- Assume that n is a small value
- Using K-means to calculate the n centroids of training data. Then use these n centroids to cluster the testing data.
- In the same cluster, the distance between inliers to centroid must smaller than the distance between outliers to centroid.
- We can take the distance to centroids as the weight value for prediction.

Method 3 - Autoencoder

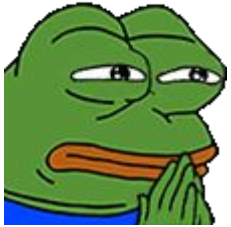
- Using training data to train a AE or VAE
- Because the outliers cannot be reconstructed well, the MSE of outliers must greater than inliers.
- We can take the reconstruction loss as the weight value for prediction.



Methed 4 - Any reasonable way you can think

The key point is to make objects within the same group as similar as possible, and keeping those in different groups to be as dissimilar as possible.

HW3? No way!


No restrictions?

Let me think...

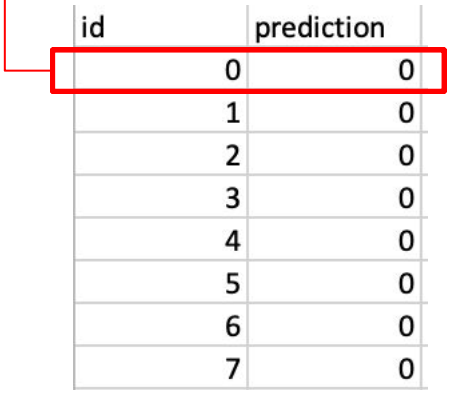
Ah, ha!

Compactness & Separation

Kaggle Submission

- Kaggle link
- Display team name : <student ID>
  - team name error : -5%
- Submission format
  - A 1726*2 .csv file, index starts from 0.
  - Prediction
    - 0: normal
    - 1: anomaly
  - Column name must be id and prediction.
- There are one simple baseline and one strong baseline, beat them to get the higher score.

This row means:
Your prediction for “0.png” is 0 (normal).



id	prediction
0	0
1	0
2	0
3	0
4	0
5	0
6	0
7	0

Kaggle Submission

- You can submit at most 3 times each day.
- The scoring metric is auc score.
- You can choose 3 of the submissions to be considered for the private leaderboard, or will otherwise default to the best public scoring submissions.

Report Submission

Answer the following 3 questions:

1. Explain your implementation which get the best performance in detail.
2. Explain the rationale for using auc score instead of F1 score for binary classification in this homework.
3. Discuss the difference between semi-supervised learning and unsupervised learning.

Please answer the question in detail !

Grading policy

- Kaggle (70%)
  - Basic score: (modified score with *)

    Over strong baseline : 55

    Over simple bassline : 45*

    Over trivial bassline : 35*

    Under trivial baseline : 25
  - Ranking score:

    15-(15/N)*(ranking-1), N=numbers of people
  - Final Kaggle score depend on
    - 30% public leaderboard
    - 70% private leaderboard

- Source code and report (30%)
  - 10% for each question

E3 Submission

Submit your source code and report to E3 before 11/5 23:59.

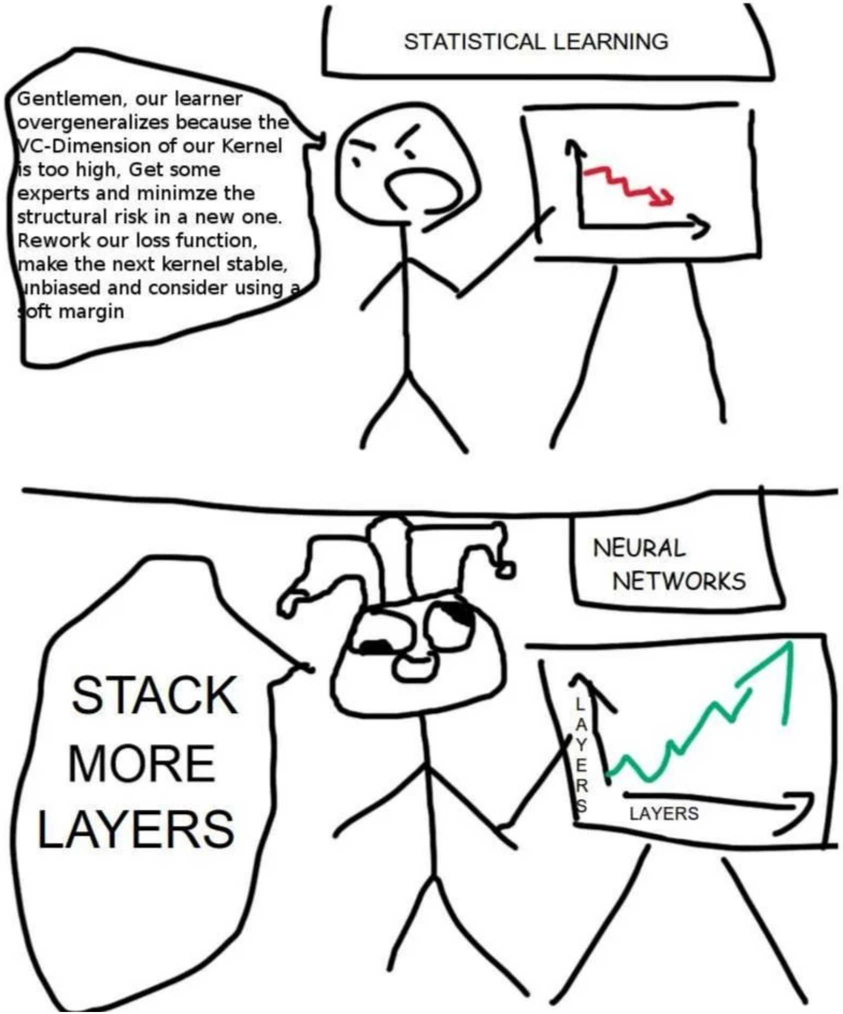
No late submission !

- Format
  - source code : HW2_<student ID>.py or HW2_<student ID>.ipynb
  - report : HW2_<student ID>.pdf

If you have any questions about HW2, please feel free to contact the TA: YU-TUNG CHOU

through email auroch.cs13@nycu.edu.tw

Take Easy


STATISTICAL LEARNING
Gentlemen, our learner overgeneralizes because the VC-Dimension of our Kernel is too high, Get some experts and minimze the structural risk in a new one. Rework our loss function, make the next kernel stable, unbiased and consider using a soft margin
NEURAL NETWORKS
STACK MORE LAYERS
LAYERS
LAYERS