Machine Learning HW3

MLTAs
ntueemlta2025@gmail.com

Links

- [Sample Code](#)
- [Kaggle](#)
- [Programming Report](#)
- [Math problems](#)

Outline

- Task Description - AutoEncoder

- Programming Report

- Requirements & Regulation

- Grading Policy

- Task: Image Classification

 - Given (64, 64, 3) jpg image → which of the 10 classes

AutoEncoder - Data

- Unlabelled: 100000 images

- Labelled: 3750 images

- Test: 6000 images

  - Public: 3000 images

  - Private: 3000 images

AutoEncoder - Methodology

- Semi-supervised Learning
  - Pre-train with unlabelled data (100000 images)
    - AutoEncoder
    - Image reconstruction loss
  - Fine-tune with labelled data (3750 images)
    - Encoder of the pre-trained AutoEncoder
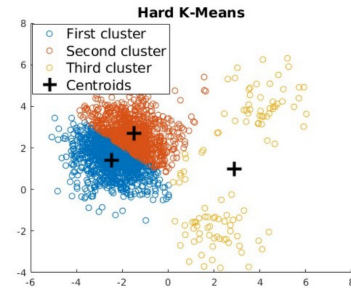   - Classification loss

AutoEncoder - Model 4/7
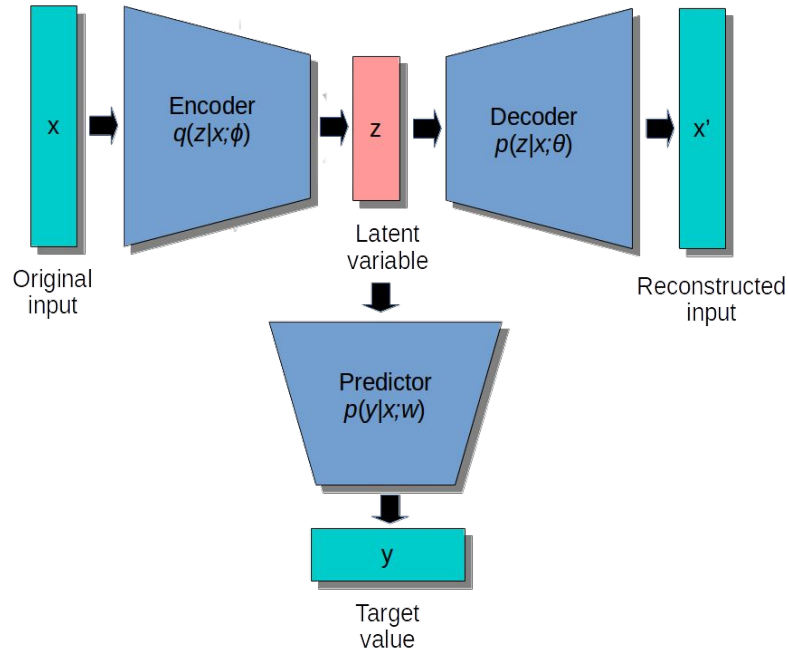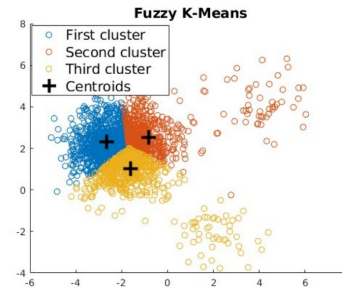
Programming Report - Clustering

- Equilibrium k-means
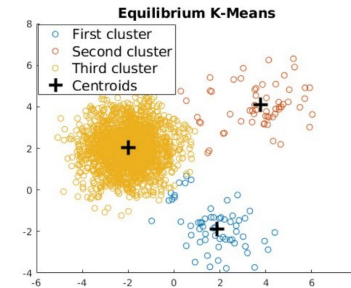 - Imbalanced data
  - Repulsion mechanism

(a)

(b)

(c)

(d)

Ref: https://arxiv.org/pdf/2402.14490
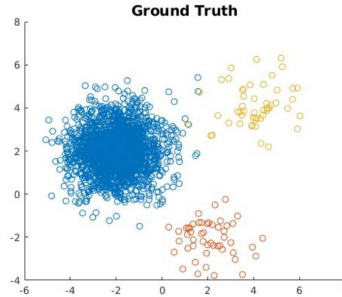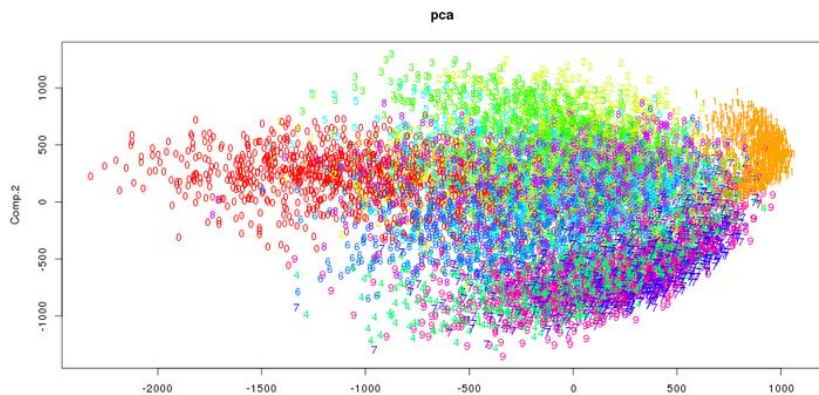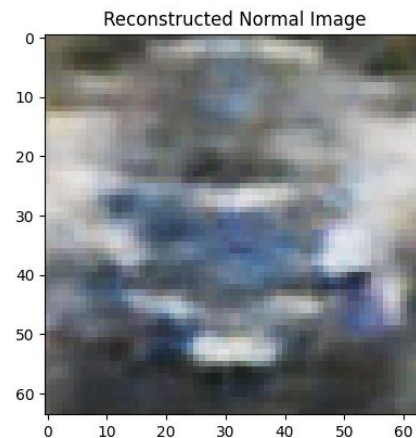
Programming Report - t-SNE

- t-SNE
  - Reduce to 2-dimensional data for visualization

Ref: https://www.biostars.org/p/392321/
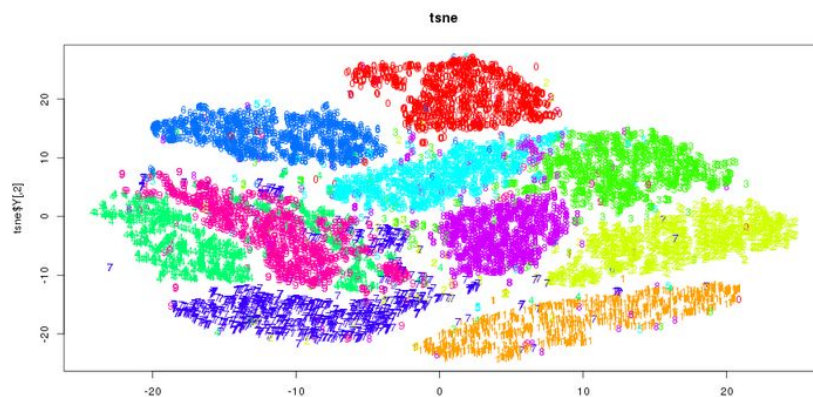
Programming Report - Anomaly Detection

- Distinguish unfamiliar/anomaly data
  - Poor reconstruction with the pre-trained AutoEncoder

Kaggle - Info 1/2
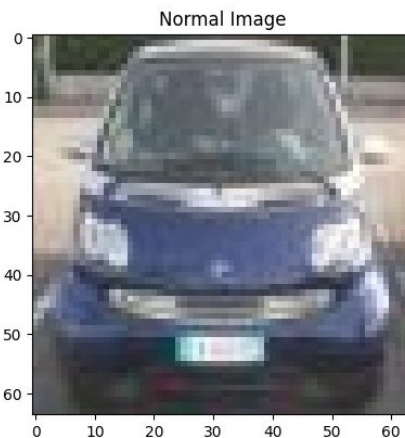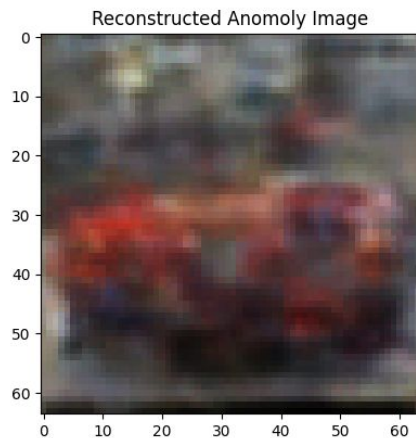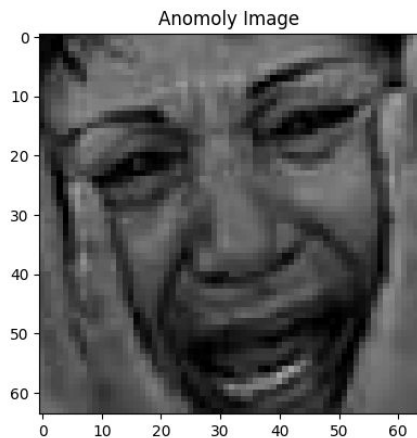
- Kaggle 連結 : https://www.kaggle.com/competitions/ml-2025-fall-hw-3
- 個人進行, 不需組隊
- 隊名:
  - 修課學生 : **學號(底線)任意名稱 (e.g., b12901024_LzTCki)**
  - 旁聽 : 旁聽_任意名稱（旁聽請於期限過後再上傳）
- 每天上傳上限 5 次
- 在Kaggle Deadline前可以選擇2份submission作為private score的評分依據。如果未勾選, 系統會自動選擇Public Leaderboard中表現最佳的兩次。

Kaggle - Format 2/2

請預測test set 6000筆資料並將結果上傳Kaggle

- 上傳格式為csv
- 第一行必須為id,label, 第二行開始為預測結果
- 每行分別為id以及預測的label, 請以逗號分隔
- Evaluation: Accuracy

```
 1   id,label
 2   0,6
 3   1,4
 4   2,0
 5   3,0
 6   4,6
 7   5,1
 8   6,4
 9   7,4
10   8,3
11   9,0
12   10,0
```

Regulation

- 開放使用套件
  - numpy
  - pandas
  - pytorch
  - torchvision
  - cv2
  - pillow
  - sklearn
- 若需使用其他套件，請儘早寄信至助教信箱詢問，並請闡明原因
- No extra data allowed
- No pre-trained model allowed

Grading Policy - Deadline

- Kaggle Deadline: 2025/10/31 23:59:59 (GMT+8)
- Cool Deadline: 2025/10/31 23:59:59 (GMT+8)

Grading Criteria

- Kaggle - 2%
    - 超過public leaderboard的simple baseline分數： **0.5%**
    - 超過private leaderboard的simple baseline分數： **0.5%**
    - 超過public leaderboard的strong baseline分數： **0.5%**
    - 超過private leaderboard的strong baseline分數：**0.5%**
    - [code template](#)
- Programming report - 4%
    - [report template](#)
- Math problem - 6%
    - [math problem](#)
    - 若有和其他修課同學討論，請務必於題號前標明 collaborator（含姓名、學號）

Cool Submissions

在Cool上分別繳交以下檔案：

1. **report.pdf**

2. **math.pdf**

3. **code.ipynb**

Grading Policy - Others

- Lateness
  - Cool 遲交每小時分數*0.95, 兩天後歸0
  - 有特殊原因請找助教

- Runtime Error
  - 當程式錯誤, 造成助教無法順利執行, 請在公告時間內寄信向助教說明, 修好之後重新執行所得kaggle部分分數將x0.5

學術倫理

- Cheating

  - 抄code、抄report（含之前修課同學）

  - 開設kaggle多重分身帳號註冊competition

  - 於訓練過程以任何不限定形式接觸到testing data的正確答案

  - 不得上傳之前的kaggle競賽

  - 教授與助教群保留請同學到辦公室解釋coding作業的權利, 請同學務必自愛