Anomaly Detection

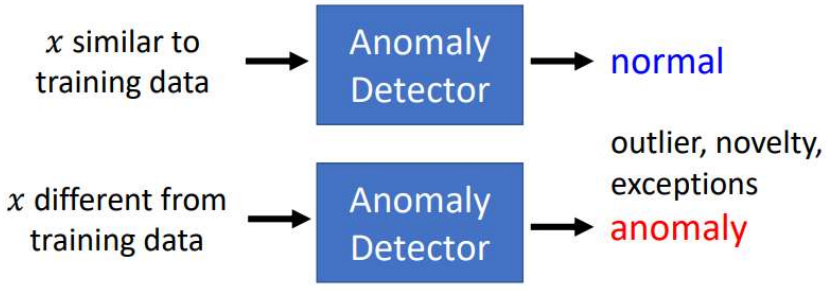
Create at 2022/06/25

- 上課資源 :
  1. Anomaly Detection (1/7) (https://www.youtube.com/watch?v=gDp2LXGnVLQ)
  2. Anomaly Detection (2/7) (https://www.youtube.com/watch?v=cYrNjLxkoXs)
  3. Anomaly Detection (3/7) (https://www.youtube.com/watch?v=ueDIm2FkCnw)
  4. Anomaly Detection (4/7) (https://www.youtube.com/watch?v=XwkHOUPbc0Q)
  5. Anomaly Detection (5/7) (https://www.youtube.com/watch?v=Fh1xFBktRLQ)
  6. Anomaly Detection (6/7) (https://www.youtube.com/watch?v=LmFWzmn2rFY)
  7. Anomaly Detection (7/7) (https://www.youtube.com/watch?v=6W8FqUGYyDo)

目的 : 讓機器知道自己不知道這件事

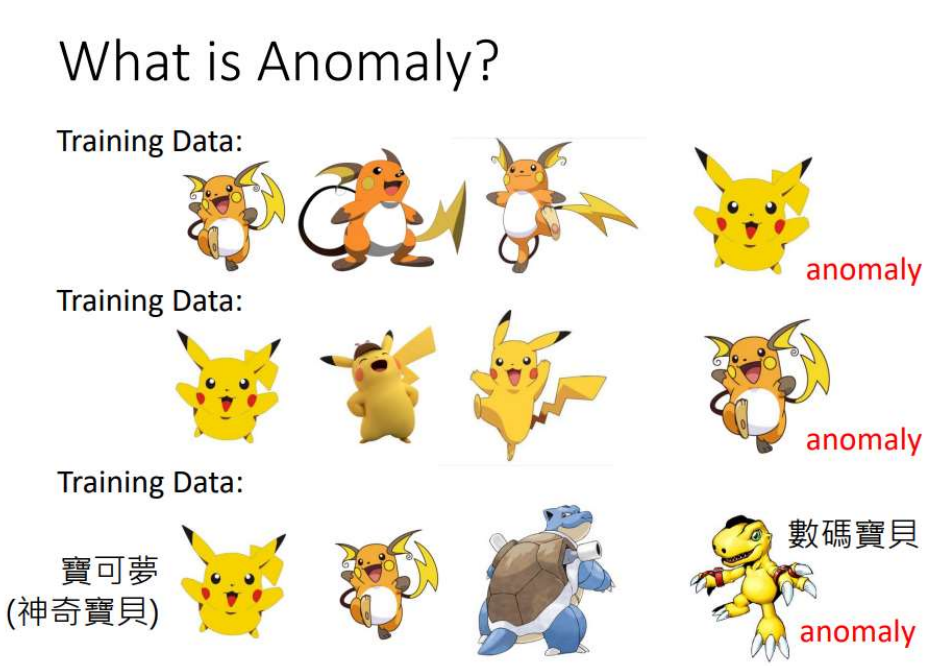
Problem Formulation

- Given a set of training data $\{x^1, x^2, \cdots, x^N\}$
- We want to find a function detecting input x is *similar* to training data or not.



Different approaches use different ways to determine the similarity.

- 在作業會用 auto-encoder 做 Anomaly Detection (異常檢測)
- 假設有一堆訓練資料 $\{x^1, x^2, \ldots, x^n\}$
- 輸入一筆資料，判斷它跟我們之前在訓練資料裡面看過的資料相不相似
- 有一個異常檢測的系統，是透過大量已經看過的資料訓練出來
  - 輸入一筆資料，如果看起來像是訓練資料裡面的 data，輸出 normal
  - 如果看起來不是訓練資料裡面的 data，輸出 anomaly (outlier, novelty, exceptions)



- Example

Applications

- Fraud Detection
  - Training data: 正常刷卡行為, x: 盜刷？
  - Ref: https://www.kaggle.com/ntnu-testimon/paysim1/home
  - Ref: https://www.kaggle.com/mlg-ulb/creditcardfraud/home
- Network Intrusion Detection
  - Training data: 正常連線, x: 攻擊行為？
  - Ref: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html
- Cancer Detection
  - Training data: 正常細胞, x: 癌細胞
  - Ref: https://www.kaggle.com/uciml/breast-cancer-wisconsin-data/home

- 常見應用
  - 詐欺偵測
  - 網路的侵入偵測
  - 癌症檢測

Binary Classification?

- Given normal data $\{x^1, x^2, \cdots, x^N\}$ ➡ Class 1
- Given anomaly $\{\tilde{x}^1, \tilde{x}^2, \cdots, \tilde{x}^N\}$ ➡ Class 2
- Then training a binary classifier ……

Binary Classification?

- 不是一般的分類問題，叫做 one class 分類問題
  - 不容易收集到異常的資料

Binary Classification?

- Given normal data $\{x^1, x^2, \cdots, x^N\}$ ➡ Class 1
- Given anomaly $\{\tilde{x}^1, \tilde{x}^2, \cdots, \tilde{x}^N\}$ ➡ Class 2
- Then training a binary classifier ……

x (Pokémon)

$\tilde{x}$ (NOT Pokémon)

cannot be considered as a class

Even worse, in some cases, it is difficult to find anomaly example ……

- 不能把異常偵測當作二元分類的原因
  - 無法知道 Class 2 整個資料異常的分佈長怎樣
    - 所以無法把整個異常資料視為一個類別，因為他的變化太大了
  - 不太容易蒐集到異常的資料

Categories

Training data: $\{x^1, x^2, \cdots, x^N\}$

With labels: $\{\hat{y}^1, \hat{y}^2, \cdots, \hat{y}^N\}$ ➡ Classifier

The classifier can output "*unknown*" (none of the training data is labelled "unknown")

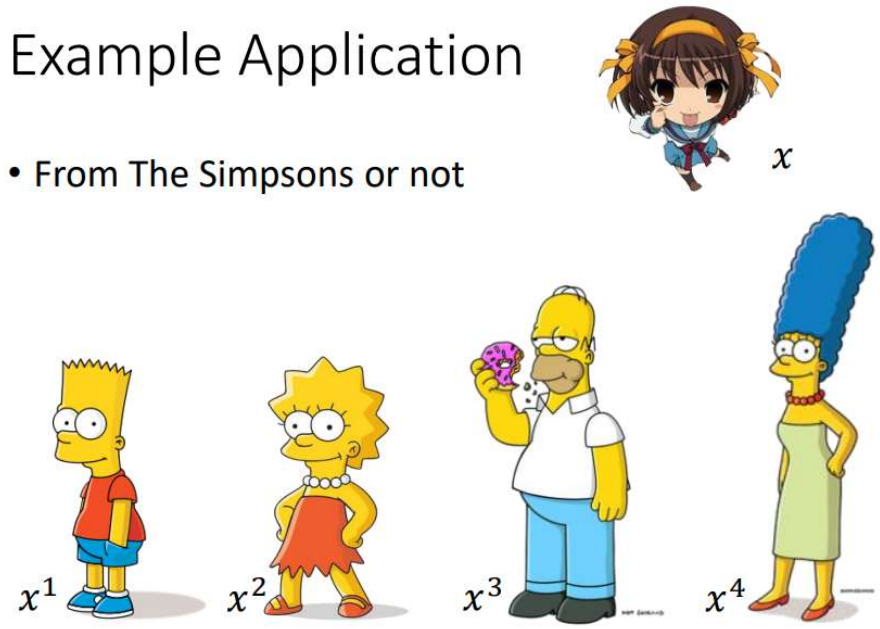
Open-set Recognition

Clean: All the training data is normal.

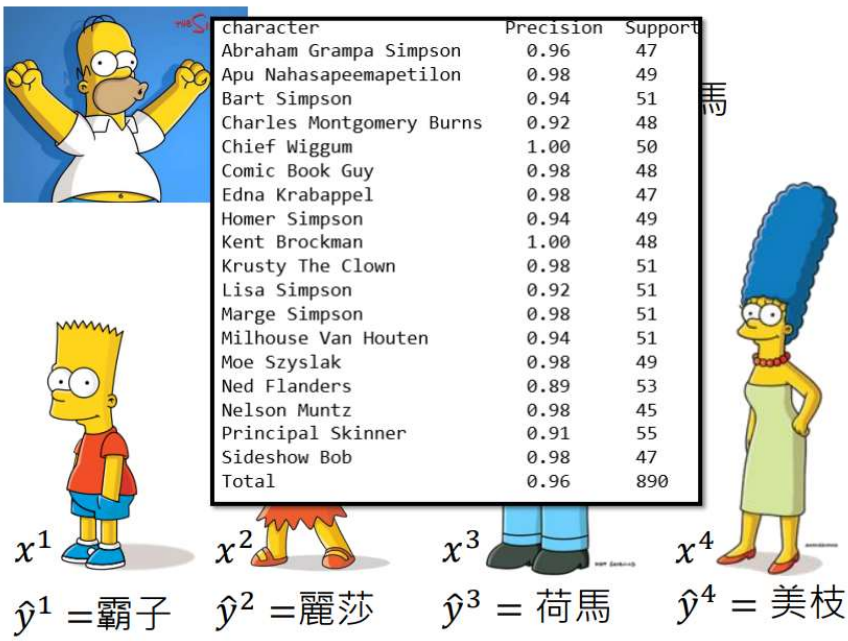
Polluted: A little bit of training data is anomaly.

- 一般的 classifier 可以用
  - generative model
  - logistic regression
  - bitnami
- 對異常問題簡單的做兩個分類
  - 不只有訓練資料，並且是有某種類型的 label
    - 期待訓練好的 classifier，可以知道沒看過的東西是 unknown (異常偵測 - Open-set Recognition)，可以辦識沒看過的東西
  - 訓練資料沒有任何 label，只能去判斷新的資料跟這些訓練資料，“像” 還是 “不像”
    - 分成兩個 case
      - Clean : 所有資料都是正常的資料
      - Polluted : 有一些資料是異常的資料

Case 1 : With Classifier



Source of model: https://www.kaggle.com/alexattia/the-simpsons-characters-dataset/



character	Precision	Support
Abraham Grampa Simpson	0.96	47
Apu Nahasapeemapetilon	0.98	49
Bart Simpson	0.94	51
Charles Montgomery Burns	0.92	48
Chief Wiggum	1.00	50
Comic Book Guy	0.98	48
Edna Krabappel	0.98	47
Homer Simpson	0.94	49
Kent Brockman	1.00	48
Krusty The Clown	0.98	51
Lisa Simpson	0.92	51
Marge Simpson	0.98	51
Milhouse Van Houten	0.94	51
Moe Szyslak	0.98	49
Ned Flanders	0.89	53
Nelson Muntz	0.98	45
Principal Skinner	0.91	55
Sideshow Bob	0.98	47
Total	0.96	890

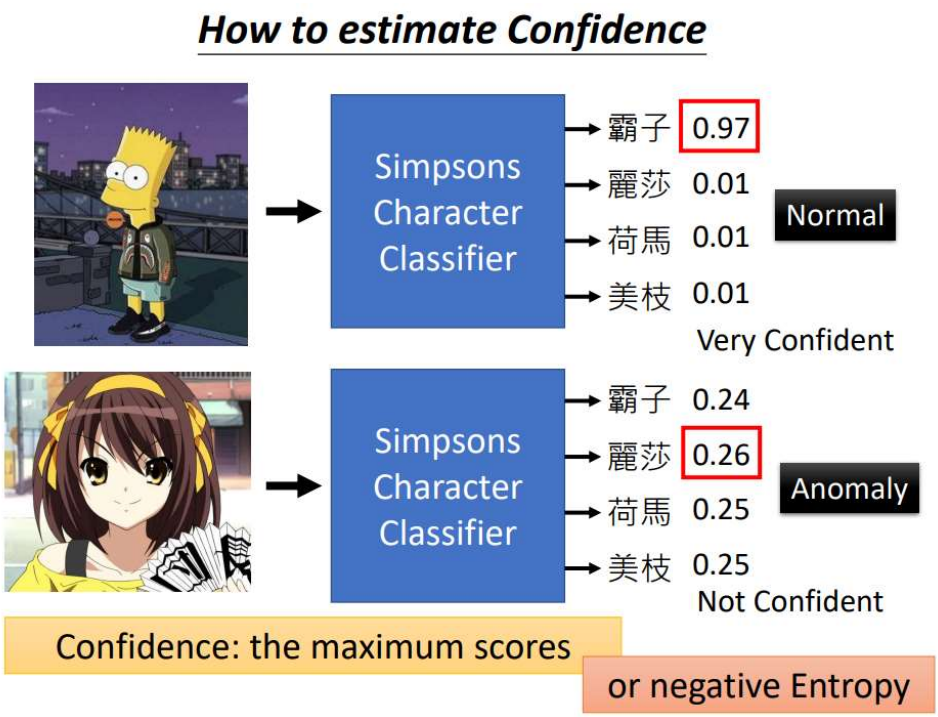
- 每一個辛普森家族的人物都有 label，每個圖片的人物是誰

How to use the Classifier

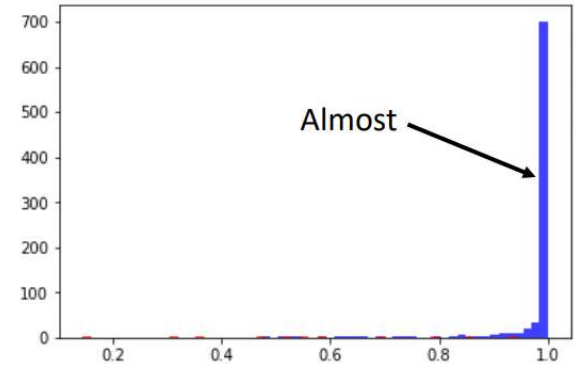
Anomaly Detection:

$$f(x) = \begin{cases} normal, & c(x) > \lambda \\ anomaly, & c(x) \leq \lambda \end{cases}$$

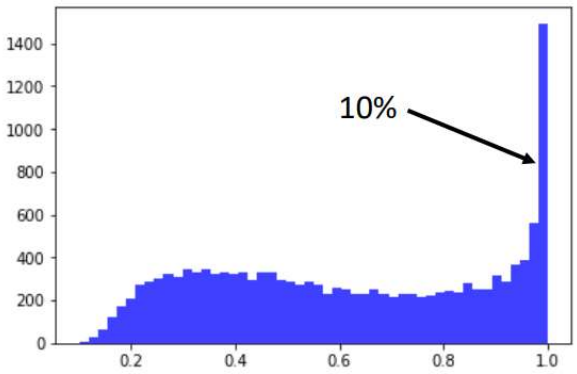
- 根據這個 classifier 來做異常偵測
- 現在我們希望 classifier 不只是做分類這件事
  - 給它一張圖片
    - 不只會告訴我們是哪個人物
    - 還會輸出一個數值代表它的信心分數
- 根據信心分數可以拿來做異常偵測這件事情
  - 如果大於 λ，代表是正常
  - 如果小於等於 λ，代表是異常



- output 是一個 distribution
- 最高的分數就是信心分數
- 給一個 distribution 就可以算 entropy
  - entropy 越大，就是 output 的 distribution 越平均，代表新的機器的分數是比較低的



Confidence score distribution for ***characters from Simpsons***

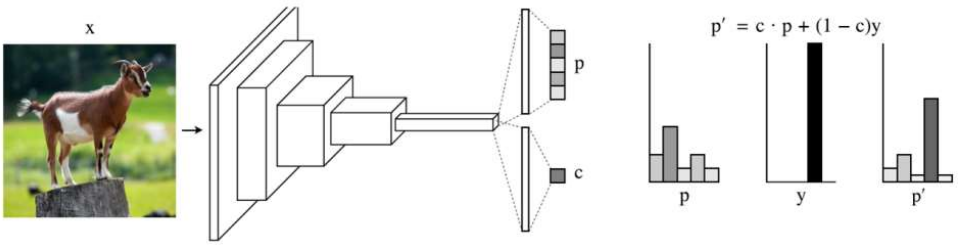


Confidence score distribution for ***anime characters***

- 辛普森家族的人物跟不是辛普森家族的人物，在 classifier 的 confidence score 上的分佈有很大的差異
- 異常偵測問題這是一個應該要做的 baseline，雖然很簡單但是不見得 performance 差

Outlook: Network for Confidence Estimation

- Learning a network that can directly output confidence



Terrance DeVries, Graham W. Taylor, Learning Confidence for Out-of-Distribution Detection in Neural Networks, arXiv, 2018

(not today)

- 之前是訓練一個 classifier 之後，在不知道怎麼回事從 classifier 的輸出得到一個信心分數

- 而這個技術是訓練一個 classifier，不只是可以做分類這件事，還會直接 output 一個分類的信心分數

Example Framework

Training Set: Images x of characters from Simpsons.

Each image x is labelled by its characters $\hat{y}$.

Train a classifier, and we can obtain confidence score $c(x)$ from the classifier.

$$f(x) = \begin{cases} normal, & c(x) > \lambda \\ anomaly, & c(x) \leq \lambda \end{cases}$$

Dev Set: Images x

Label each image x is from Simpsons or not.

We can compute the ***performance*** of $f(x)$

Using dev set to determine λ and other hyperparameters.

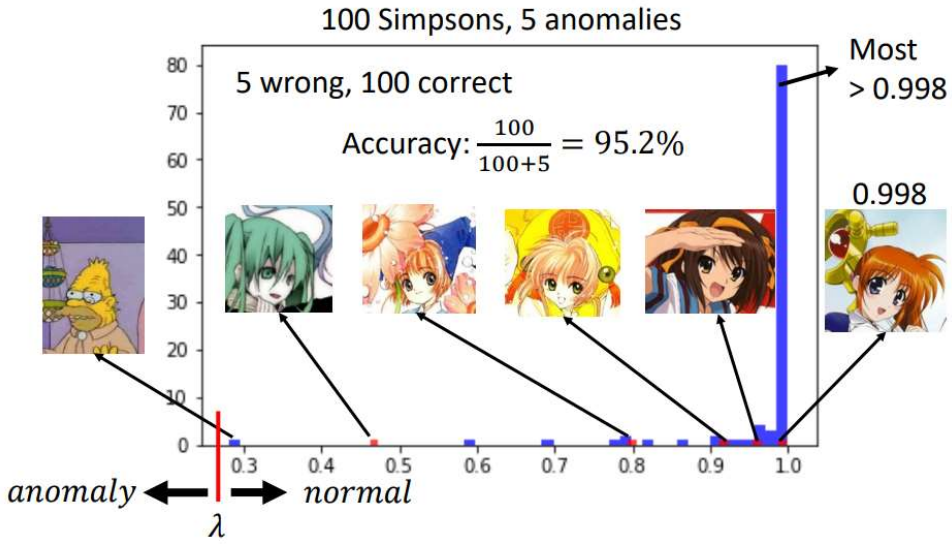
Testing Set: Images x ➡ from Simpsons or not

- 關鍵的 Development set，才能調整一些 hyperparameter，才不會 overfitting 在 testing data 上
- 在異常偵測的系統裡面，需要怎麼樣的 development set ?
  - 需要大量的 image，並且要被 label 是來自於辛普森家族人物，還是不是來自於辛普森家族的人物
  - 在 training 的時候，所有的資料都是辛普森家族的人物，而我們需要的 label 是哪一個辛普森家族的人物
  - 但是在做 develop 的時候， development set 要模仿 testing set
  - development set 要的不是一張圖片，它是辛普森家族的哪一個人物，而是 development set 裡面應該要有辛普森家族的人物跟不是辛普森家族的人物
- 接著就可以把異常偵測系統用在 development set 上面，然後計算系統在這個 development set 上面的 performance 有多好
- 然後可以用 development set 去調整 threshold，找出 threshold

Evaluation

Accuracy is not a good measurement!

A system can have high accuracy, but do nothing.

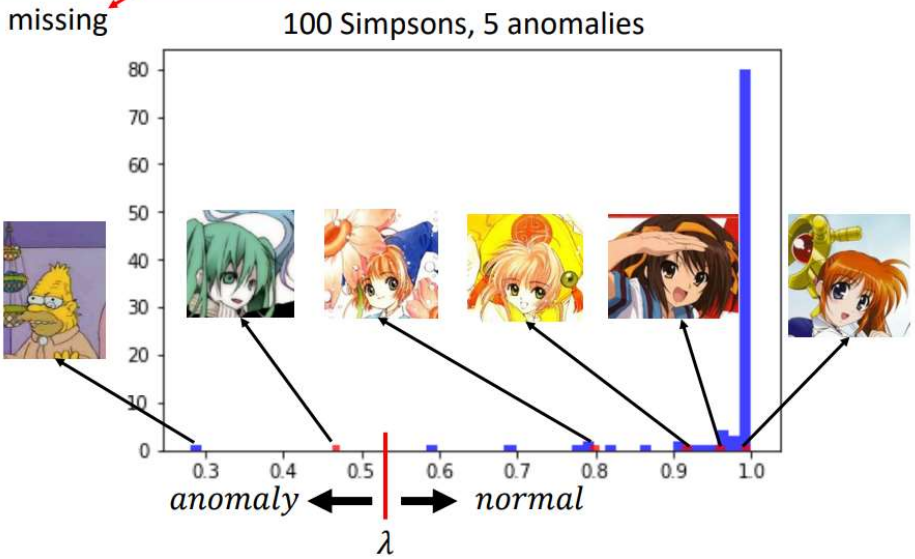


- 如何評估一個異常偵測系統的好壞？
  - 在異常評估系統裡，正確率不是一個好的評估系統的方式
  - 因為異常的東西很少，所以算出來的正確率仍然是很高的

	Anomaly	Normal
Detected	1	1
Not Det	4	99

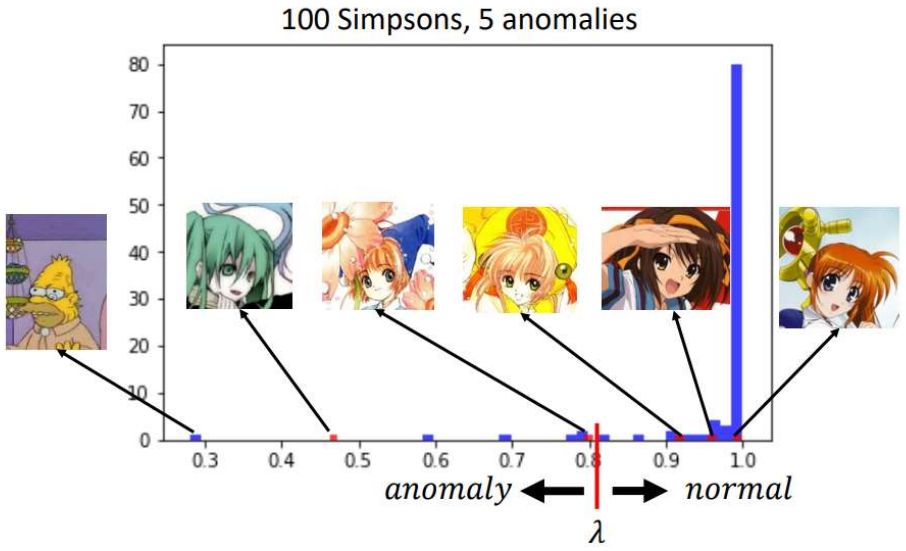
False alarm

missing



	Anomaly	Normal
Detected	1	1
Not Det	4	99

	Anomaly	Normal
Detected	2	6
Not Det	3	94



- 在異常偵測問題裡面，有兩種錯誤
  1. 異常的資料被判斷成正常的資料
  2. 正常的資料被判斷成異常的資料

	Anomaly	Normal
Detected	1	1
Not Det	4	99

Cost = 104

Cost = 401

	Anomaly	Normal
Detected	2	6
Not Det	3	94

Cost = 603

Cost = 306

Cost	Anomaly	Normal
Detected	0	100
Not Det	1	0

Cost Table A

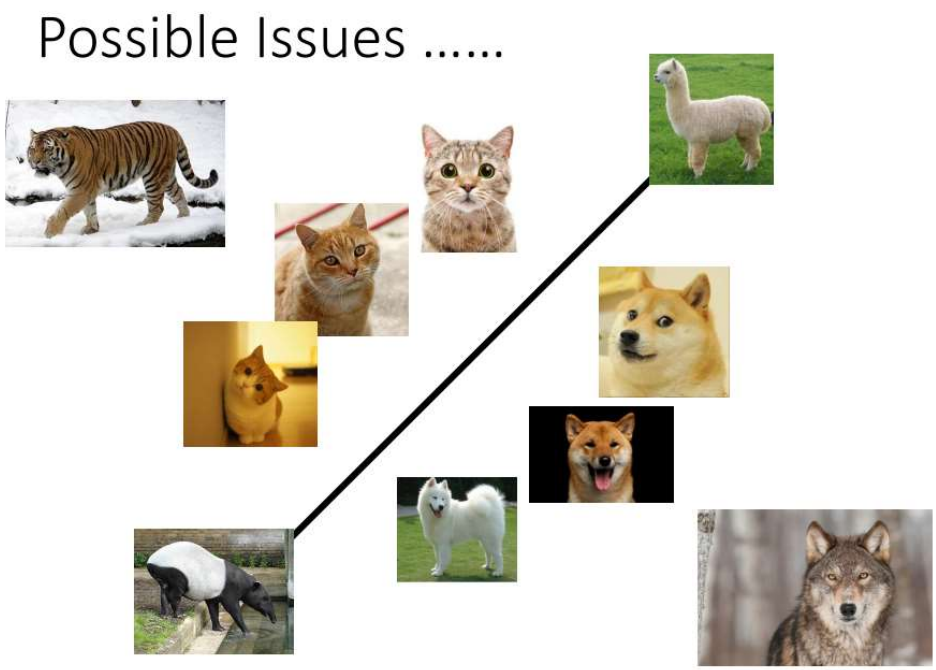
Cost	Anomaly	Normal
Detected	0	1
Not Det	100	0

Cost Table B

Some evaluation metrics consider the ranking

For example, Area under ROC curve

- 系統的好壞取決於我們覺得 missing 比較嚴重還是 False alarm 比較嚴重
- 在不同任務下會有不同的 cost table
- 常用的 measure 叫做 AUC (Area Under ROC Curve)
  - 不需要決定 threshold
  - 是看你把 testing data 裡面所有的結果，根據分數做一個排序，由高分排到低分，由最可能是異常的排到最不可能是異常的，根據這個排序來決定這個結果的好壞

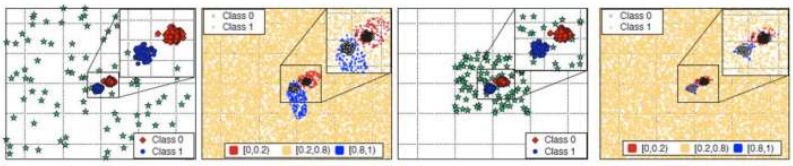


- 對機器來說，有一些東西雖然在訓練的時候沒有看過，是異常的，但是它有一些非常強的特徵會給你的分類器很大的信心說覺得看到的是某一種類別

- 臉是不是黃色，是機器判斷是不是辛普森家族的一個非常重要的特徵

To Learn More .....

- Learn a classifier giving low confidence score to anomaly



Kimin Lee, Honglak Lee, Kibok Lee, Jinwoo Shin, Training Confidence-calibrated Classifiers for Detecting Out-of-Distribution Samples, ICLR 2018

- How can you obtain anomaly?

Generating by Generative Models?

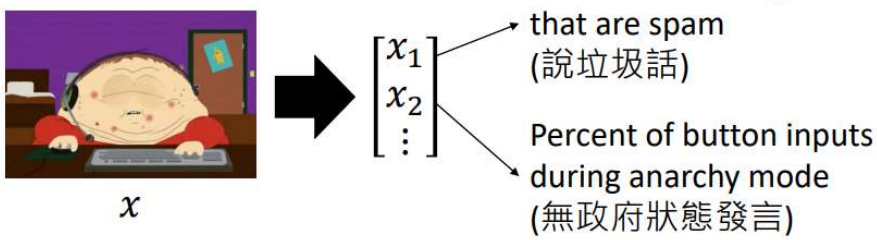
Mark Kliger, Shachar Fleishman, Novelty Detection with GAN, arXiv, 2018

- 解決的方法：
  - 假設我們可以收集到一些異常的資料
  - 可以教機器，今天看到正常的資料的時候不要只學做分類這件事
  - 要學會一邊做分類，一邊學看到正常的資料 confidence 就高
  - 看到異常的資料，confidence 就低
  - 既然收集不到異常的 data
  - 我們認一個 generative model 來生成異常的 data

Case 2 : Without Labels

Problem Formulation

- Given a set of training data $\{x^1, x^2, \cdots, x^N\}$
- We want to find a function detecting input x is *similar* to training data or not.



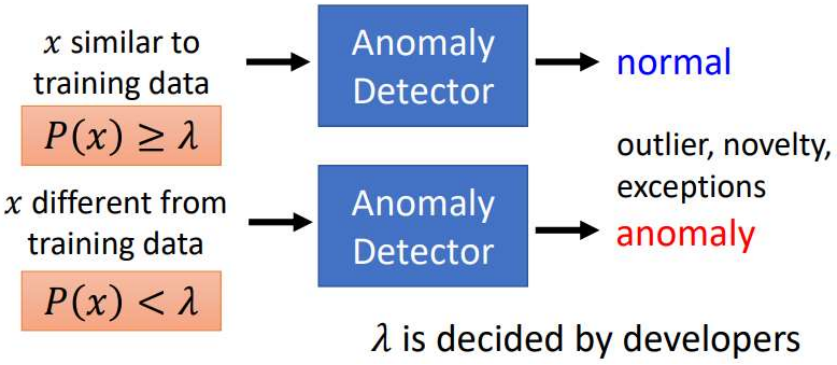
https://github.com/ahaque/twitch-troll-detection (Albert Haque)

- 需要一些訓練的資料，每一個 x 就是一個玩家
- 每個玩家必須被表示為一個 vector

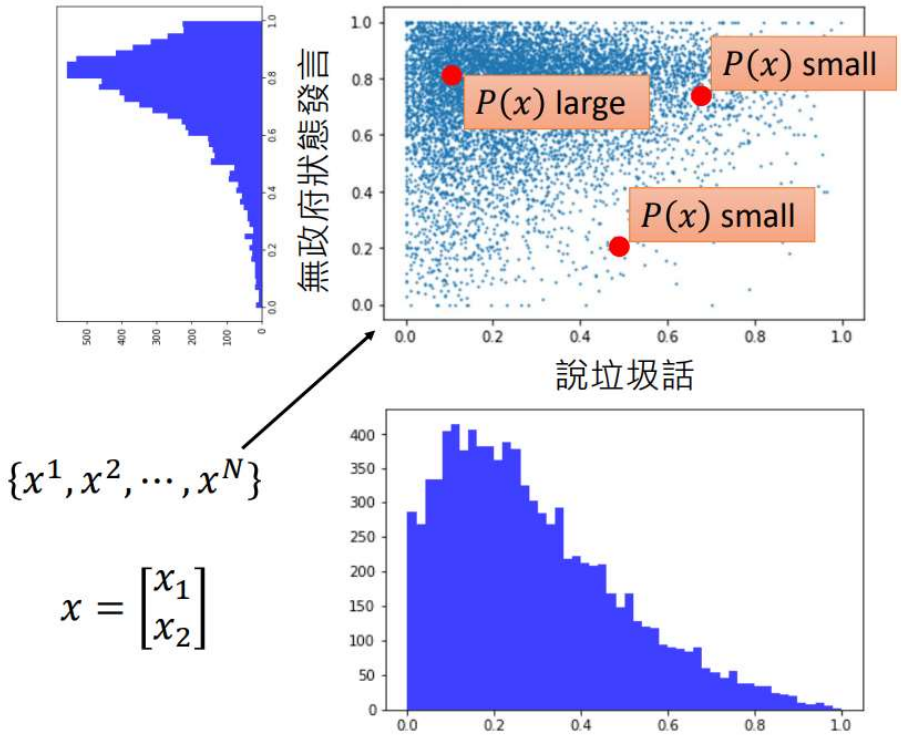
Problem Formulation

Generated from $P(x)$

- Given a set of training data $\{x^1, x^2, \cdots, x^N\}$
- We want to find a function detecting input x is *similar* to training data or not.



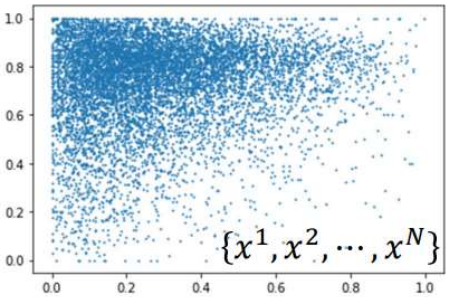
- 現在的問題
  - 只有大量的 x，但是沒有 y
- 在沒有 classifier 的情況下，我們可以建立一個模型
- 這個模型可以拿來告訴我們 $P(x)$ 的機率有多少
- 根據我們的訓練資料，可以找出一個機率模型，這個機率模型告訴我們某一種使用者這樣的行為發生的機率有多大
- 把 confidence 換成機率模型，用 threshold 去判斷



- 每一個玩家用一個二維的向量去描述他
- 希望有數值化的方法比較玩家之間誰比較異常

Maximum Likelihood

- Assuming the data points is sampled from a probability density function $f_\theta(x)$
  - θ determines the shape of $f_\theta(x)$
  - θ is unknown, to be found from data



$$L(\theta) = f_\theta(x^1) f_\theta(x^2) \cdots f_\theta(x^N)$$

Likelihood

$$\theta^* = arg \max_\theta L(\theta)$$

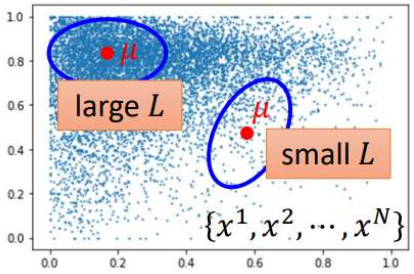
- 需要用到一個 likelihood 的概念
- 收集 N 筆資料
- 假設有一個 probability density function $f_\theta(x)$，θ 是 function 的參數決定它的形狀
- θ 必須要從 training data 裡面，根據 training data 去找出 θ
- likelihood 意思是，根據現在手上的 probability density function 左下圖的資料被產生出來的機率有多大
- $L(\theta) = f_\theta(x^1)f_\theta(x^2)\dots f_\theta(x^N)$
- 這個 likelihood 的可能性是由 θ 所控制的
- 選擇不同的 θ 就會有不同的 probability density function，就會算出不同的 likelihood
- 可以讓 likelihood 最大的 θ，是我們要找出來的

Gaussian Distribution D is the dimension of x

$$f_{\mu,\Sigma}(x) = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\Sigma|^{1/2}}exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\}$$

Input: vector x, output: probability density of sampling x

θ which determines the shape of the function are **mean** μ and **covariance matrix** Σ



$$L(\theta) = f_\theta(x^1)f_\theta(x^2)\cdots f_\theta(x^N)$$

$$L(\mu,\Sigma) = f_{\mu,\Sigma}(x^1)f_{\mu,\Sigma}(x^2)\cdots f_{\mu,\Sigma}(x^N)$$

$$\theta^* = arg\max_\theta L(\theta)$$

$$\mu^*,\Sigma^* = arg\max_{\mu,\Sigma} L(\mu,\Sigma)$$

How about $f_\theta(x)$ is from a network, and θ is network parameters? (out of the scope)

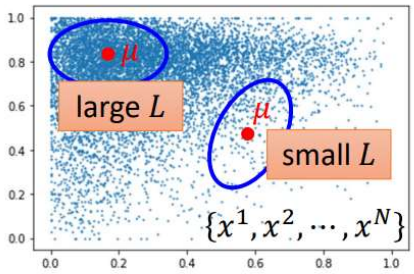
- 常用的 probability density function 就是 Gaussian Distribution
- 輸入左下圖空間裡面的一個向量 x，輸出就是這個 x 被 sample 到的機率

Gaussian Distribution

$$f_{\mu,\Sigma}(x) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} exp\left\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right\}$$

Input: vector x, output: probability of sampling x

θ which determines the shape of the function are **mean μ** and **covariance matrix Σ**



$$L(\theta) = f_\theta(x^1) f_\theta(x^2) \cdots f_\theta(x^N)$$

$$L(\mu, \Sigma) = f_{\mu,\Sigma}(x^1) f_{\mu,\Sigma}(x^2) \cdots f_{\mu,\Sigma}(x^N)$$

$$\theta^* = arg \max_\theta L(\theta)$$

$$\mu^*, \Sigma^* = arg \max_{\mu,\Sigma} L(\mu, \Sigma)$$

$$\mu^* = \frac{1}{N}\sum_{n=1}^{N} x^n = \begin{bmatrix} 0.29 \\ 0.73 \end{bmatrix}$$

$$\Sigma^* = \frac{1}{N}\sum_{n=1}^{N} (x-\mu^*)(x-\mu^*)^T = \begin{bmatrix} 0.04 & 0 \\ 0 & 0.03 \end{bmatrix}$$

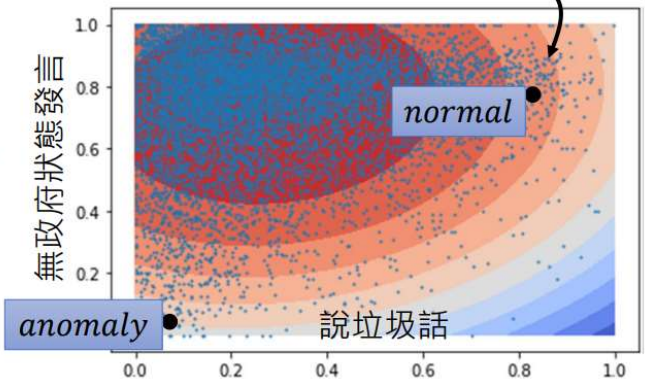
$$f_{\mu^*,\Sigma^*}(x) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma^*|^{1/2}} exp\left\{-\frac{1}{2}(x-\mu^*)^T \Sigma^{*-1}(x-\mu^*)\right\}$$

$$\mu^* = \begin{bmatrix} 0.29 \\ 0.73 \end{bmatrix} \quad \Sigma^* = \begin{bmatrix} 0.04 & 0 \\ 0 & 0.03 \end{bmatrix}$$

$$f(x) = \begin{cases} normal, & f_{\mu^*,\Sigma^*}(x) > \lambda \\ anomaly, & f_{\mu^*,\Sigma^*}(x) \leq \lambda \end{cases}$$

λ is a contour line

The colors represents the value of $f_{\mu^*,\Sigma^*}(x)$



$$f(x) = \begin{cases} normal, f_{\mu^*,\Sigma^*}(x) > \lambda \\ anomaly, f_{\mu^*,\Sigma^*}(x) \leq \lambda \end{cases}$$

More Features

x_1: Percent of messages that are spam (說垃圾話)
x_2: Percent of button inputs during anarchy mode (無政府狀態發言)
x_3: Percent of button inputs that are START (按 START 鍵)
x_4: Percent of button inputs that are in the top 1 group (跟大家一樣)
x_5: Percent of button inputs that are in the bottom 1 group (唱反調)

$\begin{bmatrix} 0.1 \\ 0.9 \\ 0.1 \\ 1.0 \\ 0.0 \end{bmatrix} \rightarrow log f_{\mu^*,\Sigma^*}(x) \rightarrow -16$

$\begin{bmatrix} 0.1 \\ 0.9 \\ 0.1 \\ 0.0 \\ 0.3 \end{bmatrix} \rightarrow log f_{\mu^*,\Sigma^*}(x) \rightarrow -22$

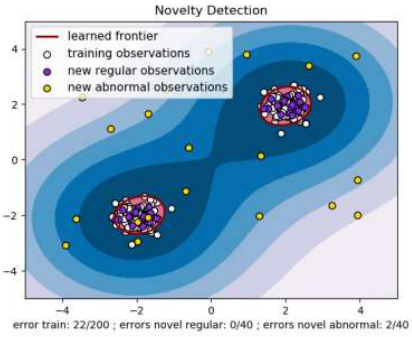
$\begin{bmatrix} 0.1 \\ 0.9 \\ 0.1 \\ 0.7 \\ 0.0 \end{bmatrix} \rightarrow log f_{\mu^*,\Sigma^*}(x) \rightarrow -2$

- 假設有一個玩家，他的決策永遠都跟別人一樣，他反而是一個奇妙的異常玩家
- 如果是有時候不一樣反而看起來比較真實，根據模型算出來的分數反而是比較高的

More ...

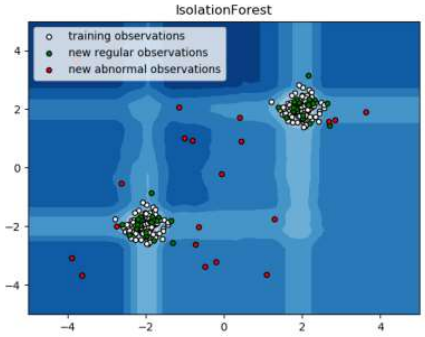
Source of images: https://scikit-learn.org/stable/modules/outlier_detection.html#outlier-detection

One-class SVM



Ref: https://papers.nips.cc/paper/1723-support-vector-method-for-novelty-detection.pdf

Isolated Forest



Ref: https://cs.nju.edu.cn/zhouzh/zhouzh.files/publication/icdm08b.pdf

- SVM，有一個叫做 one-class SVM
  - 只需要一個正常的 data 就可以訓練一個 SVM
  - 讓你分類正常的 data 跟異常的 data
- Isolated Forest
  - 在 decision tree based 方法裡面，有一個叫做 isolated forest
  - 跟 one-class SVM 做的事很像
  - 給他正常的資料，可以訓練出一個模型告訴你異常的資料長怎樣

tags: 2022 李宏毅_機器學習