Họ và Tên sinh viên: Nguyễn Thế Ngọc

Môn học: Khai Thác Dữ Liệu & Ứng Dụng

Giảng Viên: *Thầy* Phạm Trọng Nghĩa

Tìm Hiểu Outlier Detection

Mục lục

I. Outlier là gì, các loại Outlier

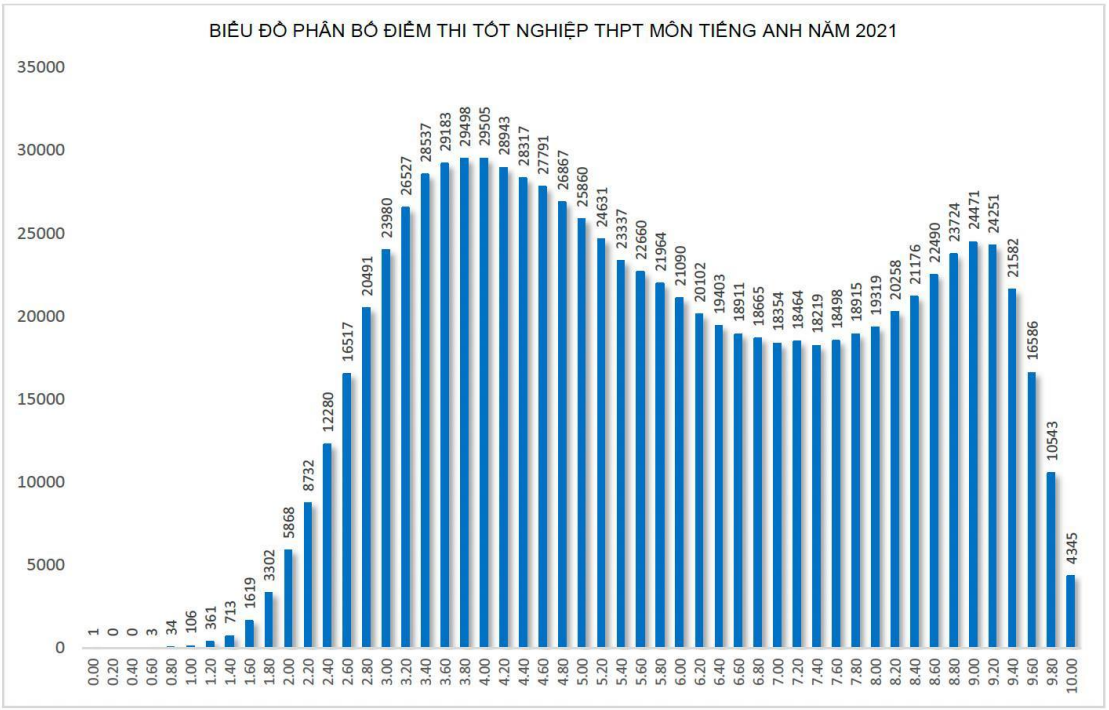
Outliers: Có tên gọi khác là anomalies, là một phần tử hay một nhóm nhỏ có sự khác biệt cách xa so với phần lớn dữ liệu khác trong tập dữ liệu, sự khác biệt này có thể là giá trị quá xa (quá lớn, quá bé…), hoặc có thể là thuộc tính quá khác (có thuộc tính hoặc giá trị thuộc tính đặt biệt, đôi khi là duy nhất).

Một số ví dụ:

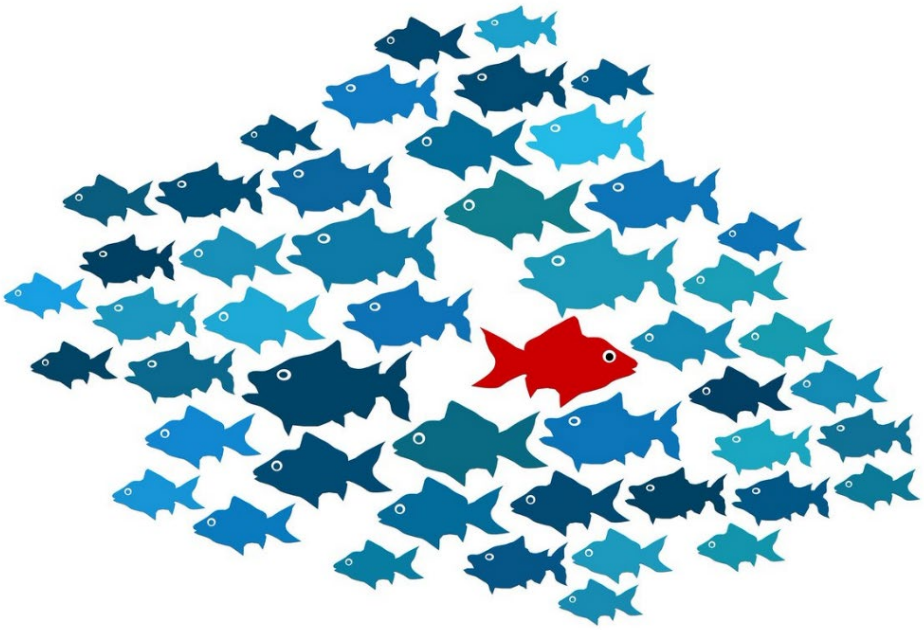
- *Giá trị cách xa phần lớn dữ liệu:* Điểm thi THPT năm 2021 môn Tiếng anh có điểm 0.0 rất cách xa các điểm khác, chỉ có duy nhất 1 bạn. (nguồn: https://vietnamnet.vn/vn/giao-duc/pho-diem-cac-mon-thi-tot-nghiep-thpt-2021-759287.html)

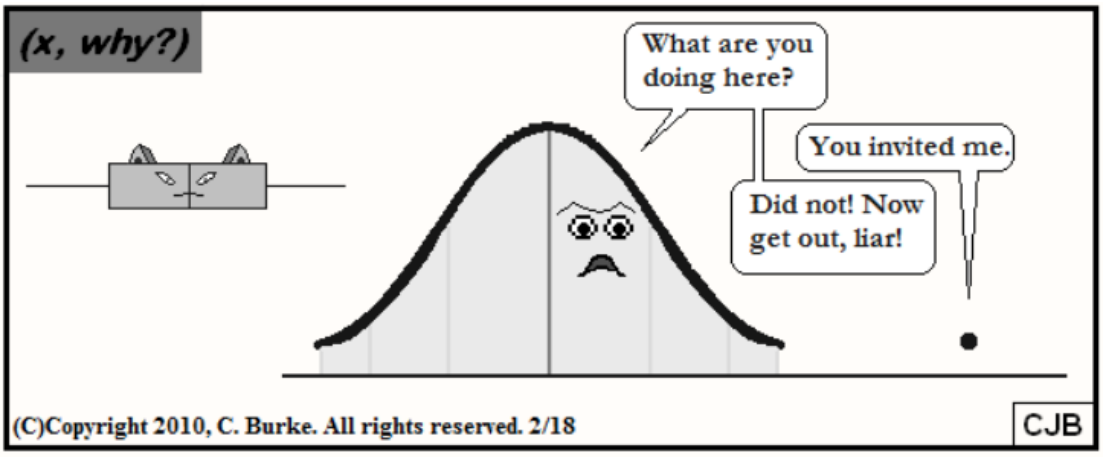
9. MÔN TIẾNG ANH

a. Phổ điểm



- *Thuộc tính đặt biệt hoặc độc nhất so với dữ liệu còn lại:* Danh sách nhân viên telesale của một công ty có duy nhất 1 người nước ngoài trong số tổng 100 nhân viên, nhân viên này công ty thuê để tư vấn cho số ít lượng khách ngoại quốc. Vì có thuộc tính khác hẳn 99 nhân viên kia về quốc tịch, mức độ lương có thể khác cũng sẽ khác, khả năng làm việc cũng có thể khác.

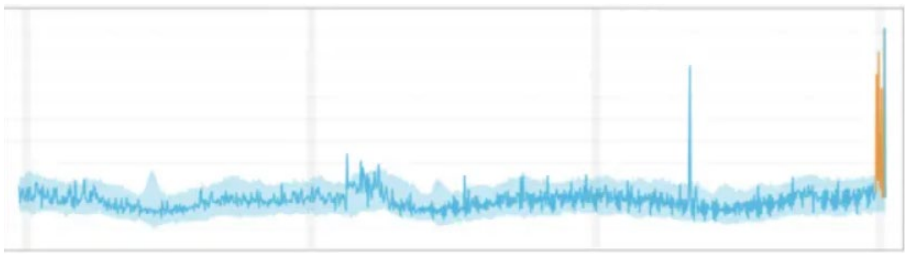




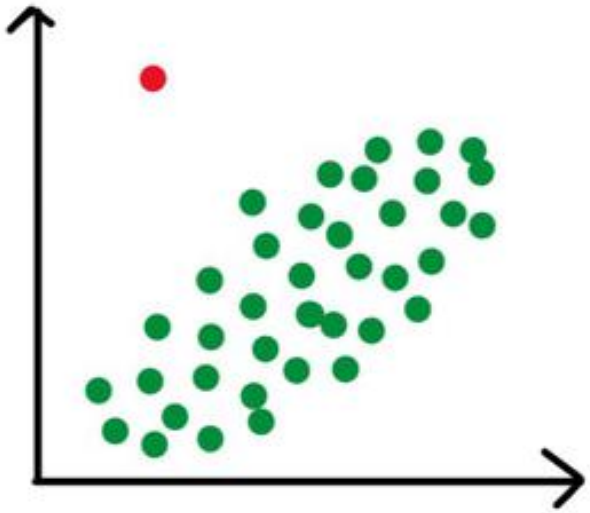
Các loại Outlier

1. Global Outlier:

Một data được coi là global outlier nếu nó nằm quá xa, nằm ngoài toàn bộ tập dữ liệu, Hầu hết các phương pháp phát hiện outlier đều nhằm mục đích tìm Global outlier.. Hình bên dưới, các dữ liệu màu vàng được coi là outlier. Ví dụ:



Ví dụ khác, khi visualize lên thì điểm đỏ nó nằm quá xa nhóm còn lại:

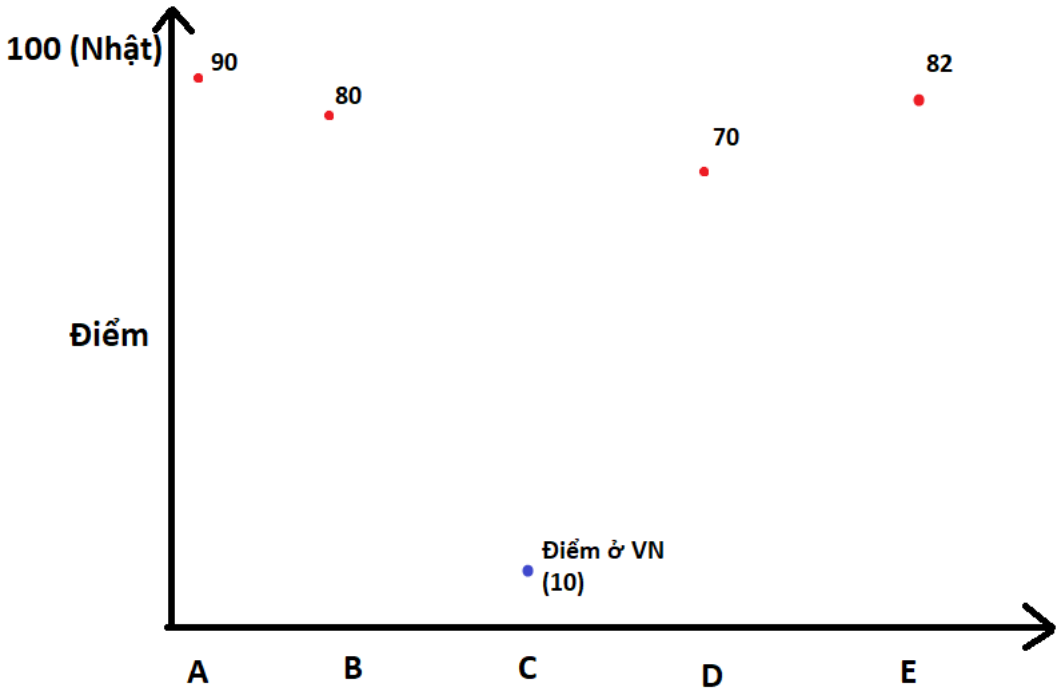


2. Contextual (Conditional) Outliers:

Còn được gọi là ngoại lệ có điều kiện, một data được coi là Contexual outlier nếu như giá trị của nó lệch đáng kể so với phần còn lại của các điểm dữ liệu, tại cùng một ngữ cảnh, điều kiện. Đôi khi ở điều kiện khác, ngữ cảnh khác nó không xảy ra hoặc giá trị ngoại lệ đó không được coi là ngoại lệ ở điều kiện khác. Vì vậy cần phân tích kỹ ngoại lệ, điều kiện của dữ liệu.

Ví dụ : Nhiệt độ 40 độ sẽ được coi là bình thường vào mùa hè, nhưng tại mùa đông như hình dưới nó là điểm outlier.

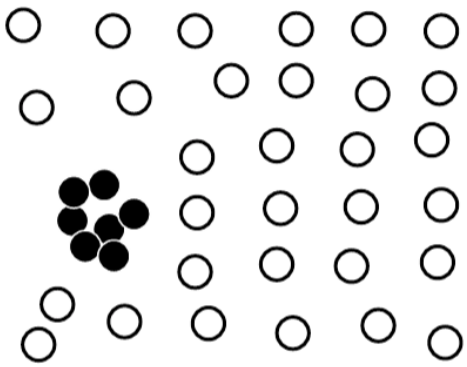
Một ví dụ khác như hình dưới, điểm số ở Việt nam thường than điểm kiểm tra phổ biến là 0-10, ở Nhật là 0-100, nếu đem điểm số 10 của 1 bạn ở Việt Nam qua bảng điểm 1 lớp ở Nhật sẽ thấy nó cực thấp, và ngược lại nếu đem điểm số 1 bạn ở Nhật sang 1 lớp học ở Việt Nam nó sẽ vượt ngưỡng tối đa.



3. Collective Outliers:

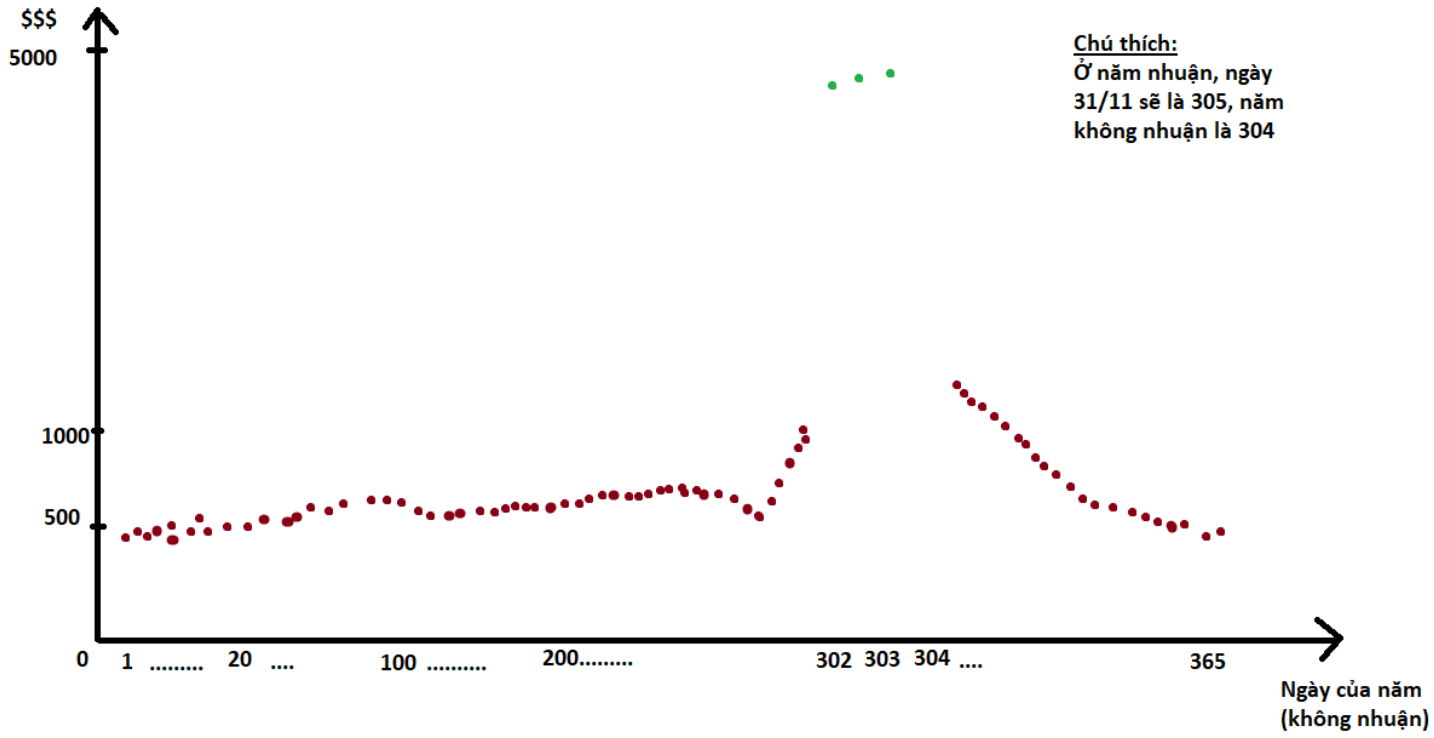
Một số điểm trong dữ liệu có sai lệchđánh kể với phần còn lại của tập dữ liệu, có thể coi các điểm này là Collective Outlier, khi nhìn riêng lẻ 1 vài điểm dữ liệu có thể khôn phải outlier nhưng khi nhìn tổng quát, các điểm này hoạt động như outlier. Để phát hiện Outlier này, cần phải có thông tin về quan hệ, thông tin nền giữa các đối tượng dữ liệu, hành vi của chúng.

Ví dụ: Hình dưới, khi ta xét riêng lẻ các điểm đen so với toàn bộ thì thấy có vẻ bình thườn, nhưng khi xét theo nhóm hành vi các điểm đứng quá sát nhau lại, ta thấy được các điểm đen có gì bất thường trong hành vi so với các điểm trắng, chúng đúng sát nhau quá mức và co cụm, các điểm đen này là collective outlier



II. Các khó khăn trong việc phát hiện Outlier

Một vài ví dụ trước khi liệt kê: Trong điểm số thi cử, nếu cả lớp, kể cả các bạn giỏi có điểm trải dài trong phổ điểm từ 4-6 điểm, có duy nhất 1 bạn học lực kém được điểm 10 thì có thể xem bạn này outlier, khá đơn giản. Tại một trường hợp khác, công ty bán đồ halloween, quanh năm họ thu nhập mỗi ngày trung bình 500$, nhưng tới cụm ngày 29-30-31 halloween và cận halloweeen, thu nhập của cửa hàng sẽ cao bất thường, khoảng 5000$ hơn, nếu thống kê ngày của năm, thì 3 ngày trên sẽ outlier nhưng đây là trường hợp đặt biệt, ta xem xét kỹ thông tin nền và nhiều thông tin khác, thay vì loại bỏ outlier này, ta tận dụng nó sẽ có ích hơn cho halloween năm sau. Nếu ta không xem xét kỹ có khi đã bỏ outlier ngày 29-30-31.

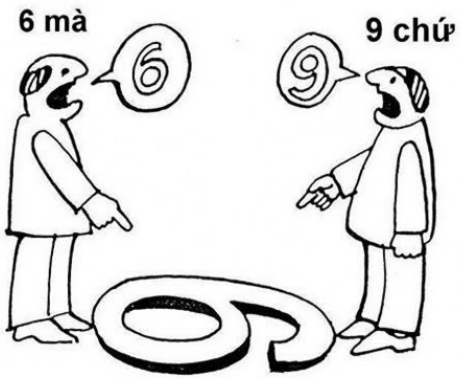


Một vài khó khăn trong việc phát hiện Outlier:

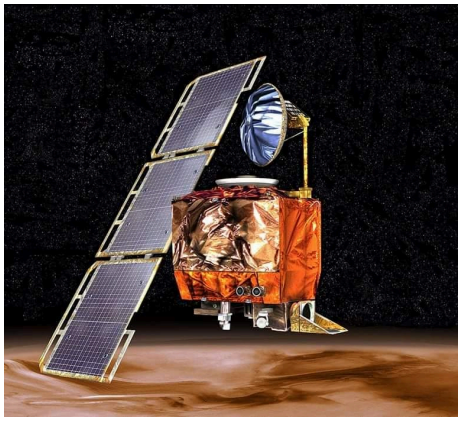
- Như đã ví dụ ở trên, với mỗi tập dữ liệu có những điều kiện riêng biệt, chủng dữ liệu cũng riêng biệt, nên việc lựa chọn mô hình thuật toán để phát hiện outlier cũng là một trong số khó khăn. Nếu tạo một mô hình riêng biệt cũng khó đảm bảo rằng tương lai có thể xử lý những trường dữ liệu lạ xuất hiện.
- Xếp loại bất thường, chọn mốc outlier. Có thể một số điểm cao bất thường nhưng lại được xếp vào nhóm bình thường thay vì nhóm outlier (False positive), các điểm xếp thấp hơn có thể là các outlier cần giữ nhưng lại xếp vào nhóm outlier (False negative, thiếu kết quả đúng). Việc chọn mốc (hay gọi là threshold) cho các điểm outlier rất quan trọng và vô cùng khó, nếu chọn quá cao, sẽ có rất nhiều điểm False negative, nếu chọn quá thấp, sẽ có nhiều

điểm false positive. Mỗi góc nhìn chủ quan của mỗi người cũng khác, có thể là xác định điểm outlier theo góc nhìn khác nhau.

- Một số tập dữ liệu quá lớn và phân bố nhiễu loạn quá nhiều cũng là trở ngại khi xác định outlier.
- Xác thực các ngoại lệ có liên quan dữ liệu, thường sẽ do góc nhìn chủ quan, góc suy nghĩ chủ quan 1 chiều hoặc mô tả thiếu từ các người cung cấp dữ liệu dẫn đến việc xác định outlier kém chính xác hơn.

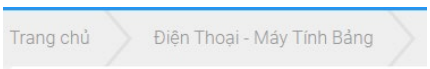


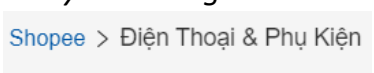
- Dữ liệu bị rỗng, thiếu quá nhiều trường.
- Không đồng nhất các đơn vị đo lường trong tập dữ liệu đó dẫn đến tính toán sai, xác định sai.



Ví dụ: có 1000 dữ liệu mà 100 dữ liệu xài đo lường feet, 900 dữ liệu xài chuẩn mét.
Trong thực tế đã xảy ra 1 tai nạn đáng tiếc vì không sử dụng đồng nhất dữ liệu, trong ngành hàng không vũ trụ, Năm 1999, NASA đã mất tàu vũ trụ Mars Climate Orbiter, thiệt hại hơn 125 triệu USD, trong phần mềm tính toán lực đẩy lên tên lửa, phần mềm dùng Pound nhưng đoạn mã thì tính theo hệ mét (N/m^2). Trong giai đoạn thiết kế, các kỹ sư động cơ đẩy tại Lockheed Martin đã dùng pound để tính lực và nhà thầu Lockheed Martin Astronautics sử dụng đơn vị đo của Anh nhưng các sứ mệnh không gian đều phải đổi sang đơn vị hệ SI để tính toán. Sự cố này đã đẩy Mars Climate Orbiter đến quá gần khí quyển Sao Hỏa. Kết quả là tàu thăm dò này bị thiêu rụi.

- Tệp dữ liệu có thể đã được trộn pha từ nhiều nguồn khác nhau dẫn đến khó đồng nhất.
Ví dụ: Dữ liệu của điện thoại samsung trên shopee, sẽ thuộc về danh mục *Điện thoại & Phụ Kiện*, còn trên tiki sẽ thuộc về danh mục *Điện thoại-Máy tính bảng*

Tiki: 

Shopee: 

III. Các phương pháp phát hiện outlier, các nhóm phương pháp

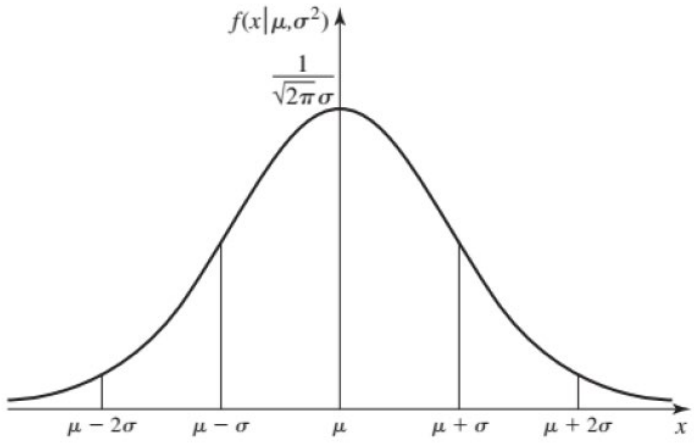
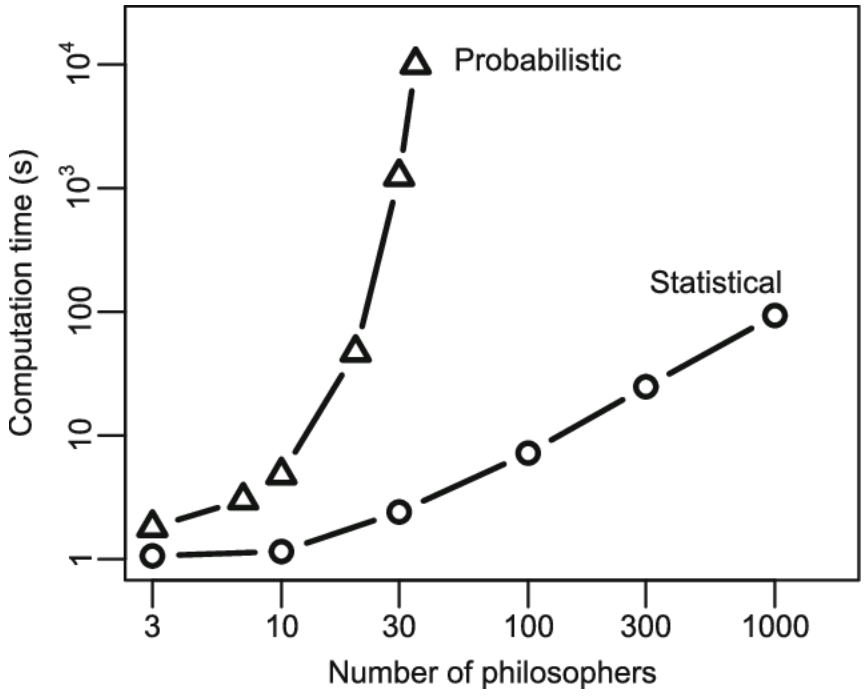
1. Các nhóm phương pháp

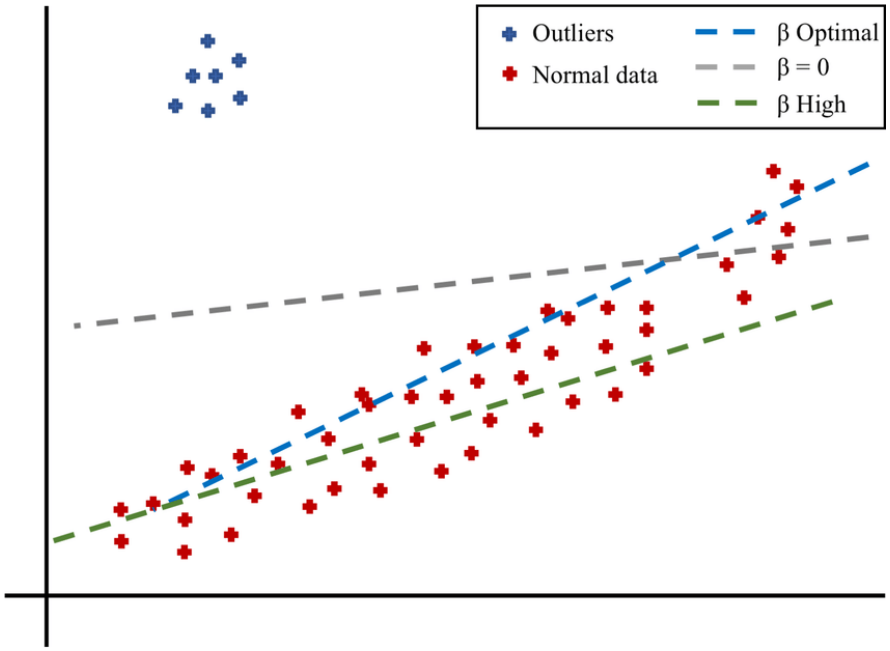
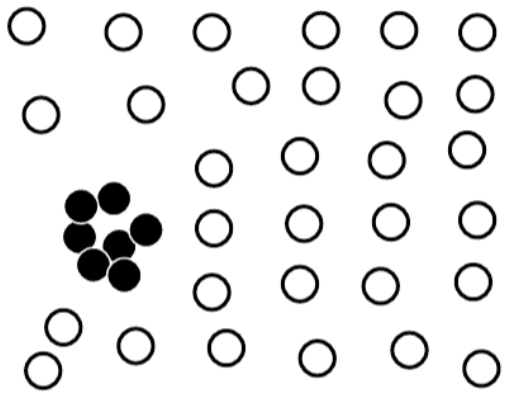
Dựa vào cuốn sách Outlier Analysis của Charu Aggarwal (https://www.amazon.com/gp/product/1461463955/). Tác giả đã phân loại các mô hình phát hiện các outlier thành 6 nhóm sau:

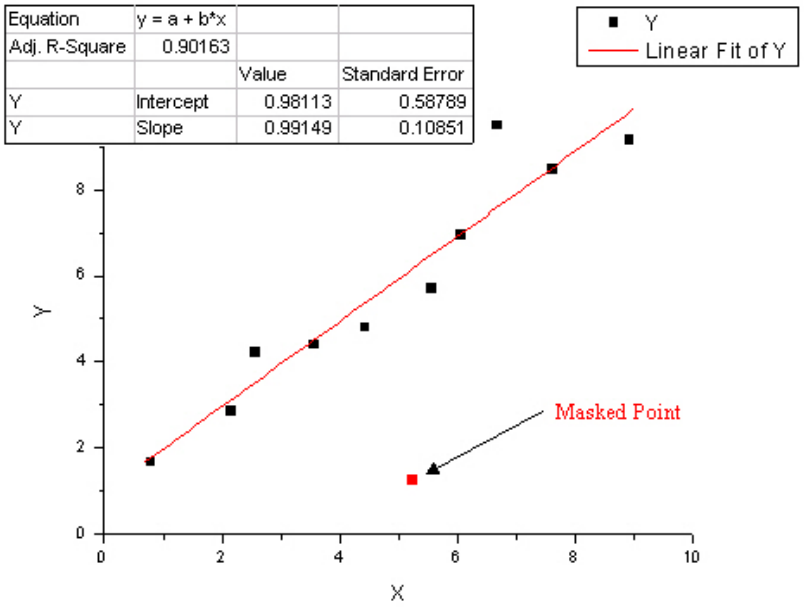
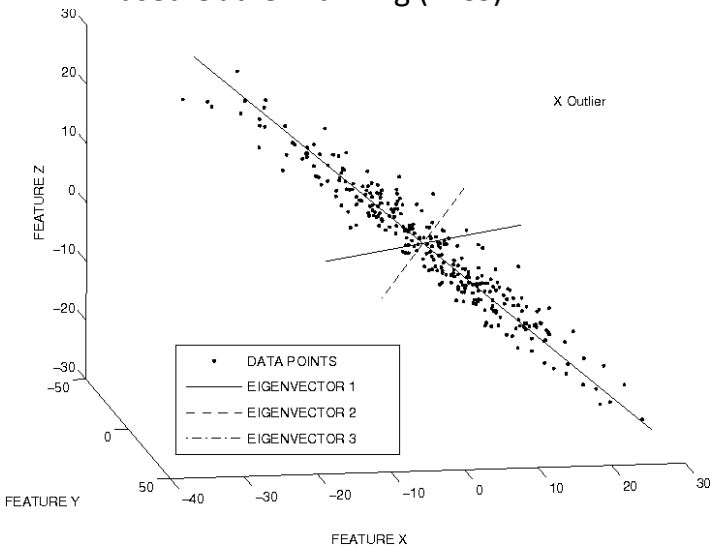
- Extreme Value Analysis
- Probabilistic and Statistical Models
- Linear Models
- Proximity-based Models
- Information Theoretic Models
- High-Dimensional Outlier Detection

2. Đặc trưng và sơ lược từng nhóm phương pháp

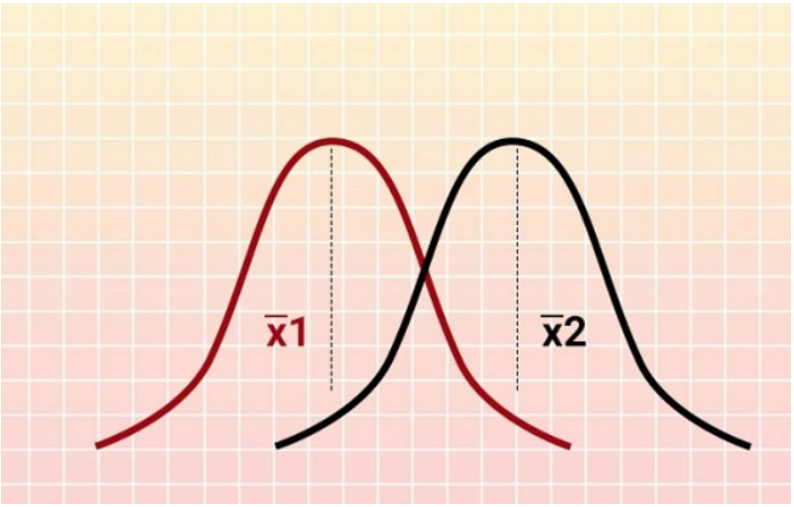
Tên nhóm phương pháp	Sơ lược, đặc trưng
Extreme Value Analysis (EVA)	• Một trong những dạng cơ bản để phát hiện outlier, tốt cho dữ liệu 1 chiều, trong mô hình này, các điểm nào có giá trị quá lớn hoặc quá bé so với nhóm còn lại bị coi là outlier. • Điểm yếu là khi gặp các dữ liệu nhiều chiều, khi đó phát hiện các outlier sẽ bị kém chính xác đi. • Các lĩnh vực hay được sử dụng như kỹ thuật, khí tượng, thủy văn, hải dương học. • Thích hợp với các dữ liệu sử dụng có các mốc thời gian khác nhau. • Mô hình được sử dụng phổ biến cho các bước cuối của quá trình nghiên cứu và phân tích. • Mô hình này rất hữu ích trong việc tìm ra các global outliers. • Một số phương pháp nổi tiếng như Z-test, Studens's t test

	
Probabilistic and Statistical Models	• Áp đặt một phân bố cụ thể cho tập dữ liệu (normal distribution, Bernoulli distribution, poisson distribution, ..). Sau đó, ta sử dụng phương pháp expectation-maximization(EM) để ước lượng tham số cho các mô hình thống kê này. Cuối cùng, ta tính xác suất cho các phần tử thuộc tập dữ liệu ban đầu. Các phần tử nào có xác suất thấp sẽ được cho là điểm ngoại lai. 

Linear Models	• phương pháp này chuyển đổi tập dữ liệu ban đầu sang không gian ít chiều hơn (sub-space) bằng cách sử dụng tương quan tuyến tính (linear correlation). Sau đó, khoảng cách của từng điểm dữ liệu đến mặt phẳng ở không gian mới sẽ định tính toán. Khoảng cách tính được này được dùng để tìm ra các điểm ngoại lai. • Một vài phương pháp: PCA (Principal Component Analysis). 
Proximity-based Models	• ý tưởng của phương pháp này là mô hình hóa các điểm ngoại lai sao cho chúng hoàn toàn tách biệt (isolated) khỏi toàn bộ các điểm dữ liệu còn lại. • Một vài phương pháp: Cluster analysis, density based analysis và nearest neighborhood. 

Information Theoretic Models	• ý tưởng của phương pháp này là dựa trên nguyên lý các điểm ngoại lai sẽ làm tăng giá trị minimum code length khi mô tả tập dữ liệu. • Một vài phương pháp: The Conventional, Information-Theoretic 
High-Dimensional Outlier Detection	• phương pháp đặc biệt để xử lý các tập dữ liệu nhiều chiều và rời rạc (high dimensional sparse data). • Một vài phương pháp: High Contrast Subspaces for Density-Based Outlier Ranking (HiCS) 

Mô tả sơ lược một vài phương pháp thuộc nhóm **Extreme Value Anaysis.**

Tên phương pháp	Sơ lược
Z-test	• Sử dụng phân phối chuẩn. • Sử dụng giá trị trung bình của một phân phối. • Xác định xem hai giá trị trung bình tổng thể có khác nhau hay không khi biết các phương sai và kích thước mẫu lớn. • Để kiểm định này được thực hiện một cách chính xác, giả định đưa ra là có phân phối chuẩn và những sai số như độ lệch chuẩn là đo đạc được. • Ngoài ra, các kiểm định T giả định độ lệch chuẩn là không xác định, trong khi các kiểm định Z gỉa định biết trước độ lệch chuẩn.
t-test	• Được sử dụng trong kiểm định sự khác biệt về giá trị trung bình của tổng thể với một giá trị cho trước, hoặc kiểm định sự khác biệt về giá trị trung bình giữa hai tổng thể. • Khác với z-test, t-test không biết trước độ lệch chuẩn, giá trị trung bình. • Được áp dụng phổ biến khi thống kê thử nghiệm tuân theo phân phối chuẩn. • T-test có thể được sử dụng để kiểm tra xem 2 bộ dữ liệu có khác nhau đáng kể hay không. 

3. Phương pháp chọn để trình bày chi tiết

Nhóm phương pháp chọn: Extreme Value Analysis (EVA)

Phương pháp chọn: Z-test

- Z-test (Kiểm định Z) là một kiểm định thống kê để xác định xem hai trung bình tổng thể có khác nhau hay không khi biết phương sai và kích thước mẫu lớn.
- Sử dụng phân phối chuẩn.
- Sử dụng giá trị trung bình của một phân phối.
- Xác định xem hai giá trị trung bình tổng thể có khác nhau hay không khi biết các phương sai và kích thước mẫu lớn.
- Để kiểm định này được thực hiện một cách chính xác, giả định đưa ra là có phân phối chuẩn và những sai số như độ lệch chuẩn là đo đạc được.
- Dùng cho cỡ mẫu lớn
- Khi tiến hành Z-test, cần nêu rõ giả thuyết không (giả thuyết ban đầu), giả thuyết nghịch, α và z-score.
- Giá trị thống kê z, hay z-score, là một giá trị biểu thị kết quả từ phép thử z.
- Các kiểm định Z có liên quan chặt chẽ với các kiểm định T, nhưng các kiểm định T được thực hiện tốt nhất khi kiểm định có cỡ mẫu nhỏ.
- Ngoài ra, các kiểm định T giả định độ lệch chuẩn là không xác định, trong khi các kiểm định Z giả định biết trước độ lệch chuẩn.

Ví dụ: **(tham khảo từ wikipedia – z test)**

Giả sử ta có điểm trung bình của học sinh là 100, độ lệch chuẩn là 12. Một dữ liệu khác thống kê từ 55 học sinh cho ra điểm trung bình là 96. Nhận xét về dữ liệu 55 học sinh?

Tính sai số SE, với 12 là độ lệch chuẩn có từ trước

$$\mathrm{SE} = \frac{\sigma}{\sqrt{n}} = \frac{12}{\sqrt{55}} = \frac{12}{7.42} = 1.62$$

Tính Z-score với giá trị trung bình của mẫu và trung bình của dữ liệu:

$$z = \frac{M - \mu}{\mathrm{SE}} = \frac{96 - 100}{1.62} = -2.47$$

Tra cứu bảng phân phối chuẩn, -2.47 = -2.0 + 0.47 = 1 – (0.9932)

= 17/2500 = 0.0068

Nhân đôi lên cho 2 phía = 0.0068 x2 = 0.014

Kết luận độ tin cậy là 1- 0.014 = 98.6%

Ví dụ ta tăng độ lệch chuẩn lên thành 40, thì SE sẽ là 5.39, Z sẽ là -0.74

tra cứu bảng thì -0.74 = 1 – (0.7704) = 0.2296, nhân đôi lên khoảng 0.46

Độ tin cậy là 1 - 0.46=54%

Kết quả cho thấy rằng nếu ta tăng độ lệch chuẩn lên hay giảm xuống sẽ ảnh hưởng tới kết quả.

4. Kết luận và đánh giá

Nhóm thuật toán	Đánh giá
Extreme Value Analysis	• Thích hợp dùng cho người mới tìm hiểu phân tích outlier. • Phù hợp với dữ liệu 1 chiều. • Thích hợp để xác định global outlier • Phương pháp t-test sẽ thích hợp cho dữ liệu nhỏ. • Phương pháp z-test sẽ phù hợp cho dữ liệu lớn.
Probabilistic and Statistical Models	• Sử dụng phương pháp EM. • Sử dụng cho dữ liệu được tạo nên từ các dữ liệu nhiều nguồn khác nhau.
Linear Models	• Xử lý trên không gian nhiều chiều từ dữ liệu ban đầu. • Sử dụng linear correlation. • Sử dụng 1 đường tuyến tính làm chuẩn. • Tính khoảng cách từ điểm đến đường tuyến tính để tìm điểm outlier. • Thích hợp cho việc sử dụng tập dữ liệu có quá nhiều trường, bị nhiễu.

Proximity-based Models	• Có 3 loại phương pháp:
	  ○ Dựa trên cụm (Cluster based methods)
	  ○ Dựa trên khoảng cách (Distance based methods)
	  ○ Dựa trên mật độ (Density based method)
	• Các phương pháp dựa trên cụm phân loại dữ liệu thành các cụm khác nhau và đếm các điểm không phải là thành viên của bất kỳ cụm nào trong số các cụm đã biết dưới dạng ngoại lệ.
	• Mặt khác, các phương pháp dựa trên khoảng cách chi tiết hơn và sử dụng khoảng cách giữa các điểm riêng lẻ để tìm ra các điểm bất thường. Phương pháp Hệ số ngoại lệ cục bộ được thảo luận ở đây bằng cách sử dụng phương pháp dựa trên mật độ.

	• Các phương pháp tiếp cận dựa trên khoảng cách sẽ gặp vấn đề trong việc tìm kiếm một điểm ngoại lai.
Information Theoretic Models	• Dựa vào giá trị tối thiểu để xác định outliers. • Phù hợp cho dữ liệu có nhiều số, bị nhiễu.
High-Dimensional Outlier Detection	• Phù hợp cho xử lý đặt biệt và người có kinh nghiệm xử lý outlier • Các tập dữ liệu bị nhiễu, nhiều chiều, rời rạc sẽ dùng phương pháp này.

Ý nghĩa của phát hiện Outlier:

- Phát hiện outlier có ý nghĩa quan trọng, để áp dụng rất nhiều trong cuộc sống.
- Nếu dữ liệu ta không xác định và loại bỏ được outlier thật thì sẽ rất ảnh hưởng đến độ chính xác của thuật toán dự đoán của chúng ta.
- Một số phương pháp phát hiện outlier còn được áp dụng để phát hiện bất thường trong dữ liệu, từ đó phân tích được chuỗi dữ liệu bất thường đó để tìm ra các phương pháp tức thì, ví dụ như lỗi thanh toán một con số quá cao.

IV. Video demo, source code một thuật toán phát hiện outlier

- Link google drive chứa toàn bộ tệp, source code liên quan bài tìm hiểu: https://drive.google.com/drive/folders/1C9xo16kuFxutuc-wGpyEMC7fFz834TrC?usp=sharing

- Link Youtube xem trực tuyến video demo thuật toán Z-test dựa vào ví dụ halloween: https://youtu.be/HcFPhuFMXEQ

V. Các nguồn tham khảo

- https://data-fun.com/outliers-loai-bo-du-lieu-ngoai-lai-mysql/
- https://ongxuanhong.wordpress.com/2016/01/31/lay-va-lam-sach-du-lieu-xu-ly-du-lieu-ngoai-lai-outliers/
- https://www.geeksforgeeks.org/types-of-outliers-in-data-mining/
- https://www.anodot.com/blog/quick-guide-different-types-outliers/
- https://pub.towardsai.net/why-outlier-detection-is-hard-94386578be6c
- https://towardsdatascience.com/anomaly-detection-with-extreme-value-analysis-b11ad19b601f

- https://www.sciencedirect.com/topics/computer-science/outlier-detection?fbclid=IwAR3GOuJ42fslxOCFgw9kpUXCGXtgzqblBMX-5yZZHeEzTW3oi5U1fR0c1kw
- https://towardsdatascience.com/anomaly-detection-with-extreme-value-analysis-b11ad19b601f
- https://en.wikipedia.org/wiki/Z-test
- https://en.wikipedia.org/wiki/Student%27s_t-test
- https://cs.nju.edu.cn/zlj/Course/DM_16_Lecture/Lecture_8.pdf
- https://www.datasciencecentral.com/m/blogpost?id=6448529%3ABlogPost%3A375126
- https://courses.lumenlearning.com/odessa-introstats1-1/chapter/types-of-outliers-in-linear-regression/
- https://stats.stackexchange.com/questions/30999/differences-between-a-statistical-model-and-a-probability-model
- https://www.investopedia.com/terms/z/z-test.asp
- https://www.geeksforgeeks.org/z-score-for-outlier-detection-python/
- https://towardsdatascience.com/z-score-for-anomaly-detection-d98b0006f510
- https://medium.com/analytics-vidhya/outlier-detection-with-python-79a82514223b