CPSC 340: Machine Learning and Data Mining

Outlier Detection

Summer 2021

In This Lecture

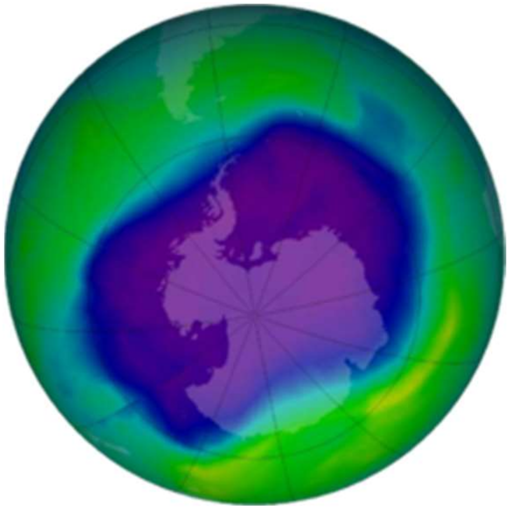
- Outlier Detection (30 minutes)
- Linear Regression Intro (20 minutes)

Coming Up Next

OUTLIER DETECTION

Motivating Example: Finding Holes in Ozone Layer

- The huge Antarctic ozone hole was "discovered" in 1985.



People before 1985

- It had been in satellite data since 1976:
  - But it was flagged and filtered out by a quality-control algorithm.

https://en.wikipedia.org/wiki/Ozone_depletion

What is an Outlier?

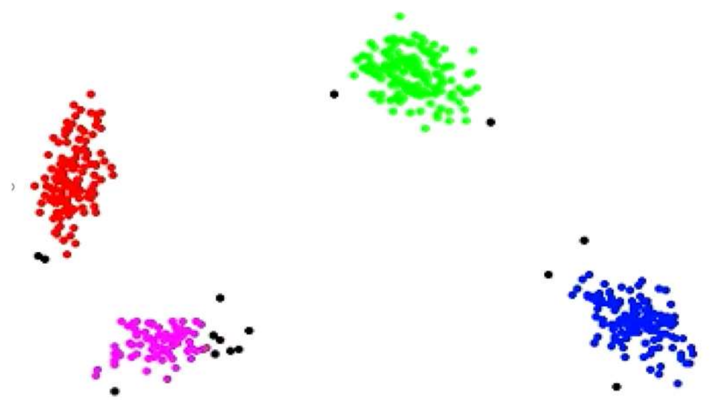
- Outlier := un-usually different observation
  - Usual difference: noise/variance in data, no worries
  - Unusual difference: even with noise/variance, this is weird

Outlier Detection in Learning

- Outlier detection is used in both supervised and unsupervised contexts

$	Hi	CPSC	340	Vicodin	Offer	...		Spam?
1	1	0	0	1	0	...	→	1
0	0	0	0	1	1	...	→	1
0	1	1	1	0	0	...	→	1
...	...	...	...	...	...	...	→	...

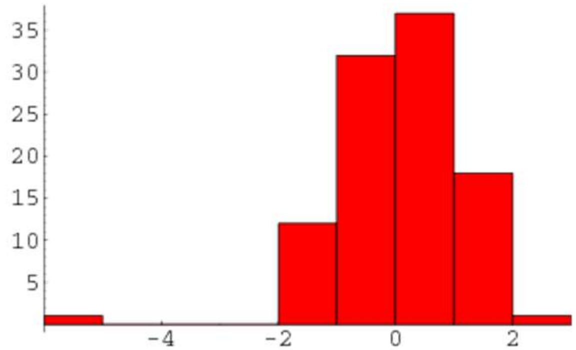
Supervised:
examples with weird labels

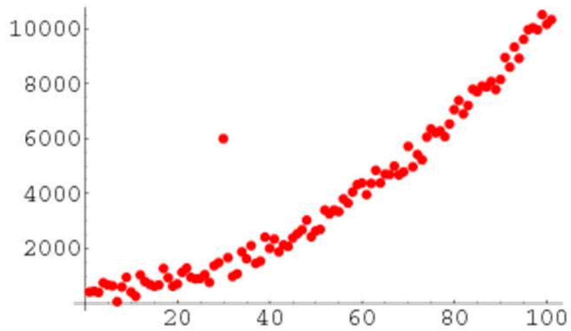


Unsupervised:
examples that look different from others

Outlier Detection

- Outlier detection:
  - Also known as “anomaly detection”.
  - May want to remove outliers, or be interested in the outliers themselves (security).



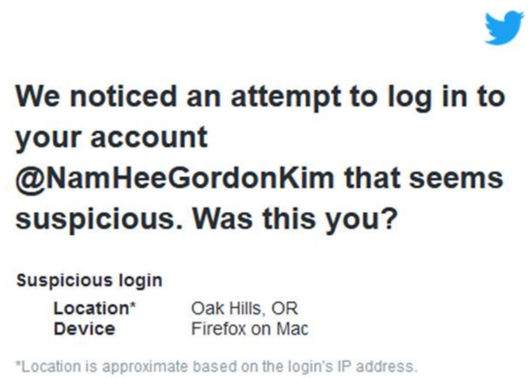


- Some sources of outliers:
  - Measurement errors.
  - Data entry errors.
  - Contamination of data from different sources.
  - Rare events.

http://mathworld.wolfram.com/Outlier.html

Applications of Outlier Detection

- Data cleaning: less outliers → better models
- Security and fault detection (network intrusion, DOS attacks).
- Fraud detection (credit cards, stocks, voting irregularities).



- Detecting natural disasters (underwater earthquakes).
- Astronomy (find new classes of stars/planets).
- Genetics (identifying individuals with new/ancient genes).

Classes of Methods for Outlier Detection

1. Model-based methods.
2. Graphical approaches.
3. Cluster-based methods.
4. Distance-based methods.
5. Supervised-learning methods.

- Warning: these solutions are highly ambiguous.
  - Human intuition is (usually) required for good results

But first…

- Usually it's good to do some basic sanity checking…

Egg	Milk	Fish	Wheat	Shellfish	Peanuts	Peanuts	Sick?
0	0.7	0	0.3	0	0	0	1
0.3	0.7	0	0.6	-1	3	3	1
0	0	0	"sick"	0	1	1	0
0.3	0.7	1.2	0	0.10	0	0	2
900	0	1.2	0.3	0.10	0	0	1

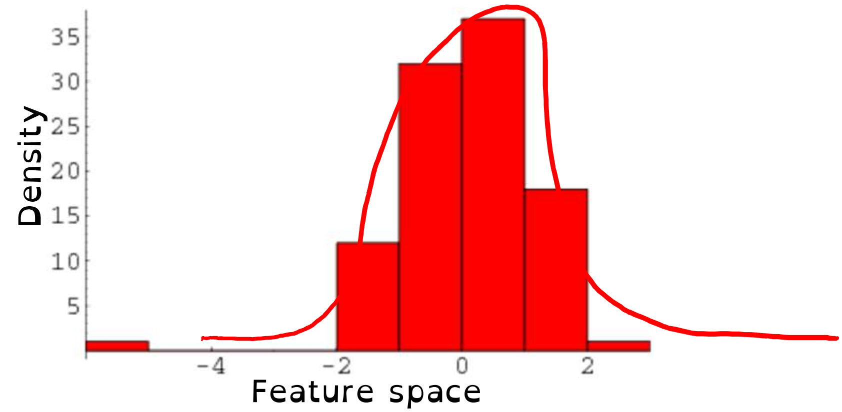
  - Would any values in the column cause a Python "type" error?
  - What is the range of numerical features?
  - What are the unique entries for a categorical feature?
  - Does it look like parts of the table are duplicated?
- These types of simple errors are VERY common in real data.

Coming Up Next

MODEL-BASED OUTLIER DETECTION

Model-Based Outlier Detection

- Model-based outlier detection:
  1. Fit a probability density function.
  2. Outliers are examples with _______________.



- Example:
  – Assume data follows normal distribution.
  – The z-score for 1D data is given by:

$$z_i = \frac{x_i - \mu}{\sigma} \quad \text{where } \mu = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{and } \sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2}$$

Q: When is z-score high? When is z-score low?

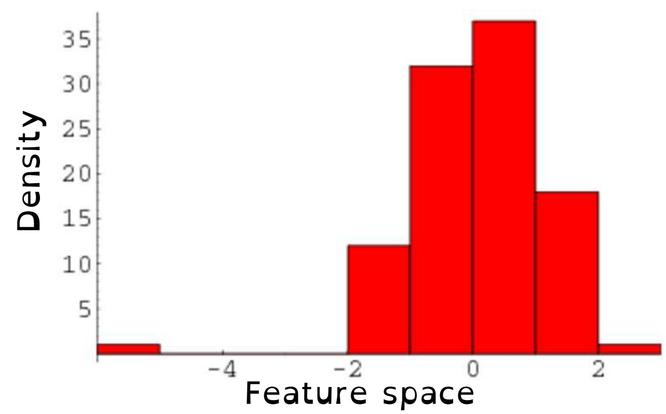
  – "Number of standard deviations away from the mean".
  – Say "outlier" if $|z| > 4$, or some other threshold.

Q: What's the problem with using mean and variance?

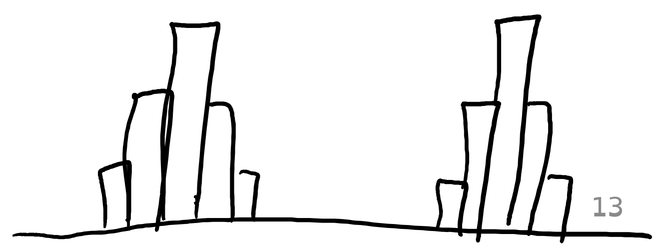
http://mathworld.wolfram.com/Outlier.html

Problems with Z-Score

- Unfortunately, the mean and variance are ___________ to outliers.



  - Possible fixes: use quantiles, or sequentially remove worse outlier.
- The z-score also assumes that data is "uni-modal".
  - Data is concentrated around the mean.
  - Bonus: why Mark Schmidt hates "curving" grades



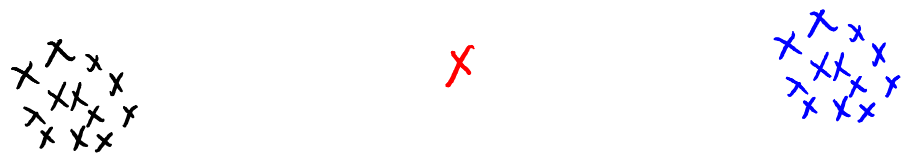
http://mathworld.wolfram.com/Outlier.html

Global vs. Local Outliers

- Is the red point an outlier?

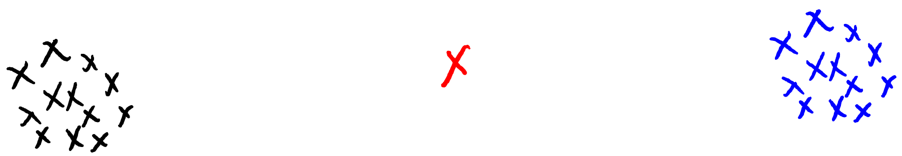
Global vs. Local Outliers

- Is the red point an outlier? What if we add the blue points?



Global vs. Local Outliers

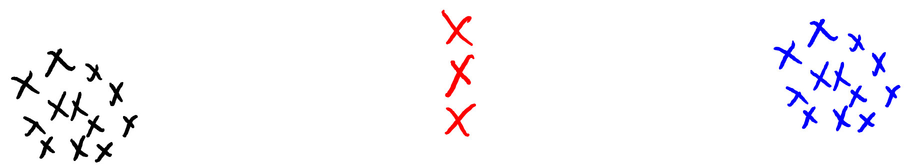
- Is the red point an outlier? What if we add the blue points?



- Red point has the lowest z-score.
  - In the first case it was a "global" outlier.
  - In this second case it's a "local" outlier:
    - Within normal data range, but far from other points.
- It's hard to precisely define "outliers".

Global vs. Local Outliers

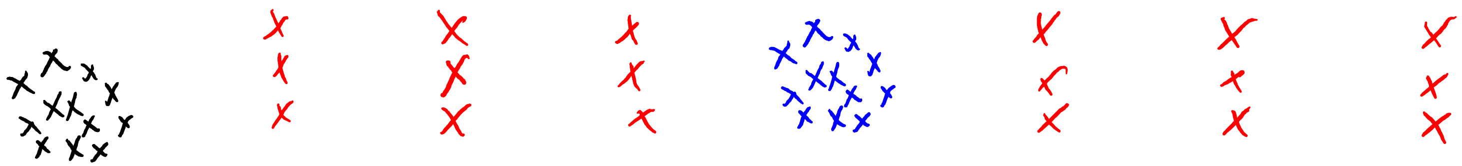
- Is the red point an outlier? What if we add the blue points?



- Red point has the lowest z-score.
  - In the first case it was a "global" outlier.
  - In this second case it's a "local" outlier:
    - Within normal data range, but far from other points.
- It's hard to precisely define "outliers".
  - Can we have outlier groups?

Global vs. Local Outliers

- Is the red point an outlier? What if we add the blue points?



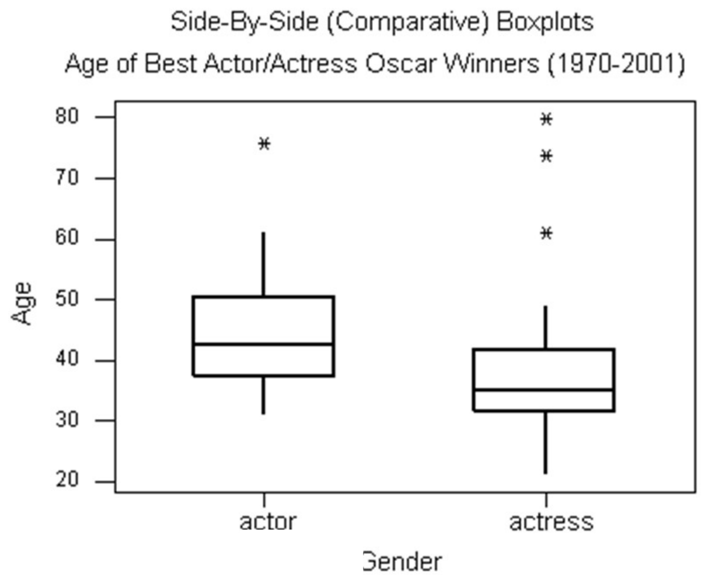
- Red point has the lowest z-score.
  - In the first case it was a "global" outlier.
  - In this second case it's a "local" outlier:
    - Within normal data range, but far from other points.
- It's hard to precisely define "outliers".
  - Can we have outlier groups? What about repeating patterns?

Coming Up Next

GRAPHICAL OUTLIER DETECTION

Graphical Outlier Detection

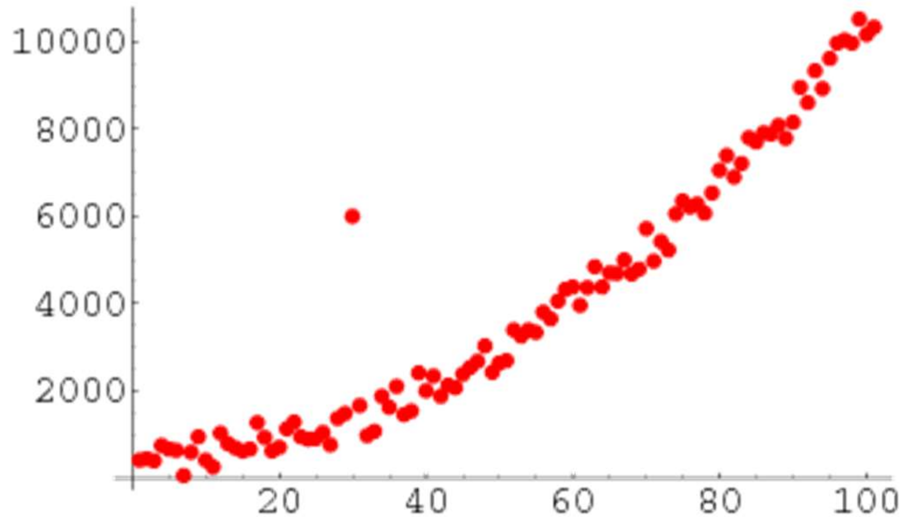
- Graphical approach to outlier detection:
  1. Look at a plot of the data.
  2. Human decides if data is an outlier.
- Examples:
  1. Box plot:
     - Visualization of quantiles/outliers.
     - Only 1 variable at a time.



http://bolt.mph.ufl.edu/6050-6052/unit-1/one-quantitative-variable-introduction/boxplot/

Graphical Outlier Detection

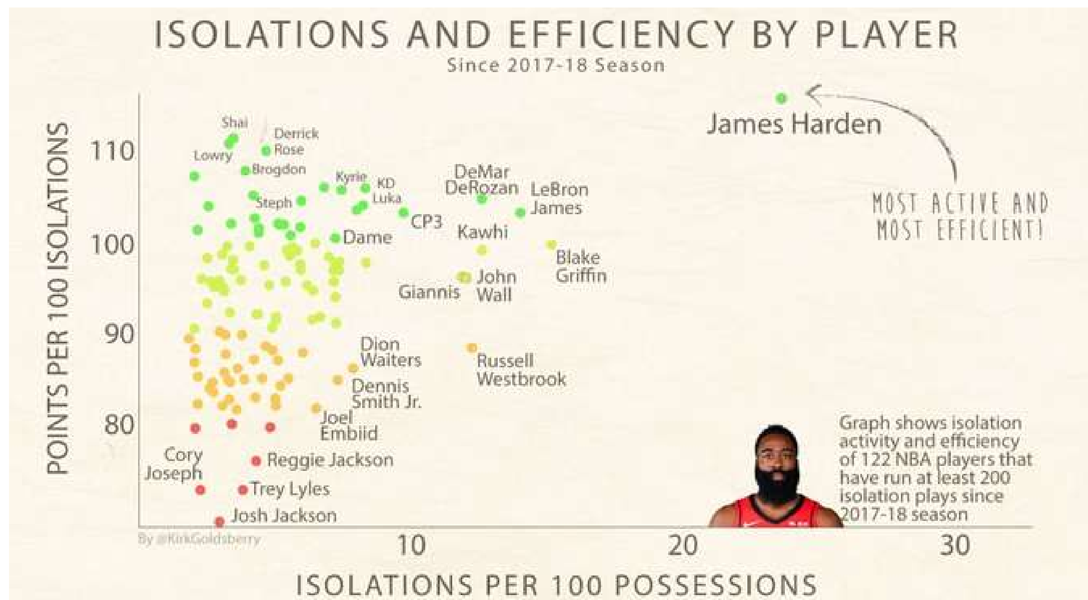
- Graphical approach to outlier detection:
  1. Look at a plot of the data.
  2. Human decides if data is an outlier.
- Examples:
  1. Box plot.
  2. Scatterplot:
     - Can detect complex patterns.



http://mathworld.wolfram.com/Outlier.html

Graphical Outlier Detection

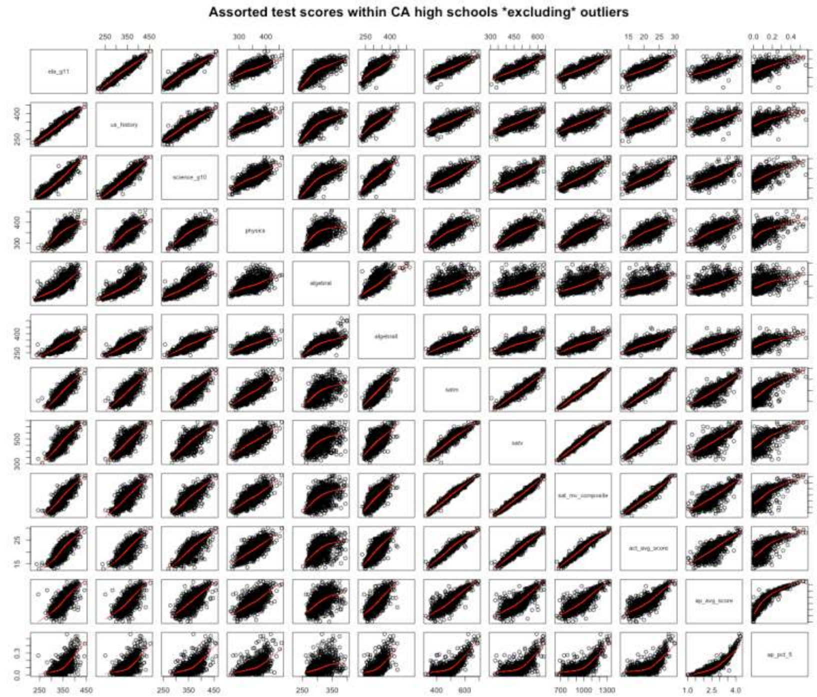
- Graphical approach to outlier detection:
  1. Look at a plot of the data.
  2. Human decides if data is an outlier.
- Examples:
  1. Box plot.
  2. Scatterplot:
     - Can detect complex patterns.
     - Only 2 variables at a time.



https://www.espn.com/nba/story/_/id/28214630/james-harden-greatest-scorer-nba-era

Graphical Outlier Detection

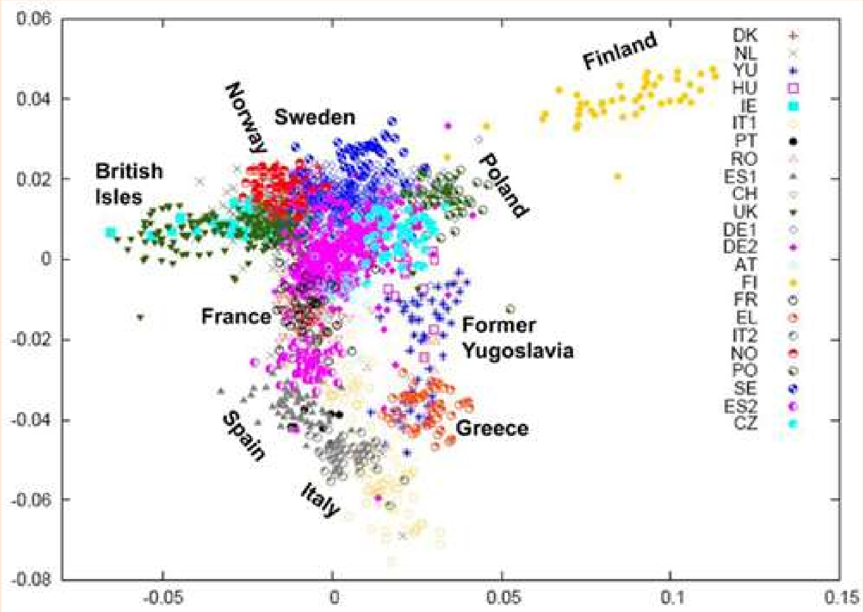
- Graphical approach to outlier detection:
  1. Look at a plot of the data.
  2. Human decides if data is an outlier.
- Examples:
  1. Box plot.
  2. Scatterplot.
  3. Scatterplot array:
     - Look at all combinations of variables.
     - But laborious in high-dimensions.
     - Still only 2 variables at a time.



https://randomcriticalanalysis.wordpress.com/2015/05/25/standardized-tests-correlations-within-and-between-california-public-schools/

Graphical Outlier Detection

- Graphical approach to outlier detection:
  1. Look at a plot of the data.
  2. Human decides if data is an outlier.
- Examples:
  1. Box plot.
  2. Scatterplot.
  3. Scatterplot array.
  4. Scatterplot of 2-dimensional PCA:
     - 'See' high-dimensional structure.
     - But loses information and sensitive to outliers.



We'll cover PCA later in course

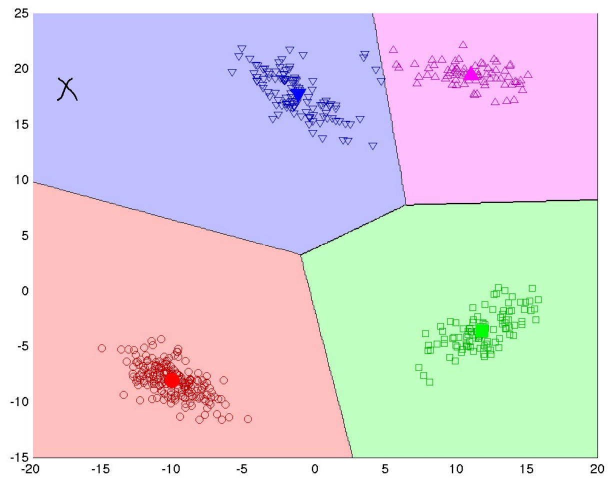
http://scienceblogs.com/gnxp/2008/08/14/the-genetic-map-of-europe/

Coming Up Next

CLUSTER-BASED OUTLIER DETECTION

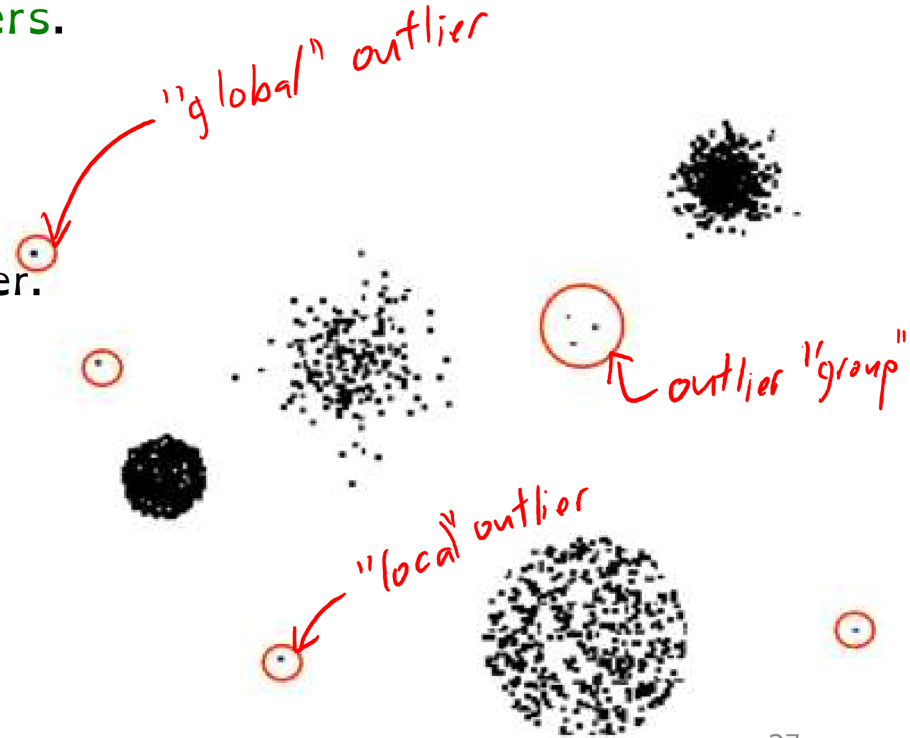
Cluster-Based Outlier Detection

- Detect outliers based on clustering:
  1. Cluster the data.
  2. Find points that don't belong to clusters.
- Examples:
  1. K-means:
     - Find points that are far away from any mean.
     - Find clusters with a small number of points.



Cluster-Based Outlier Detection

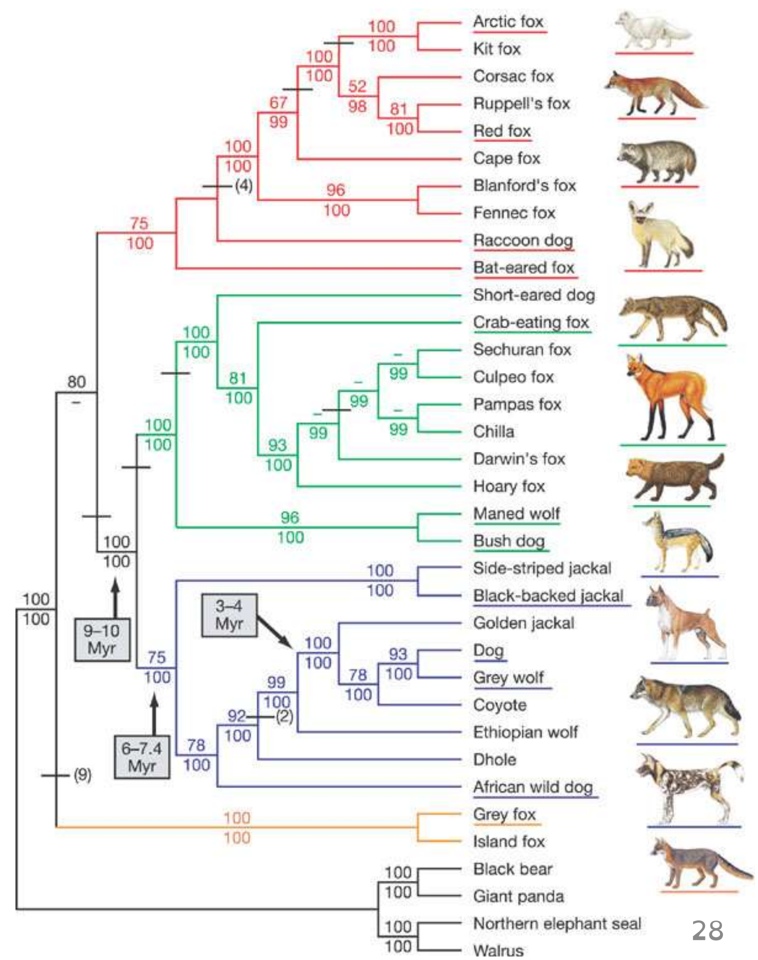
- Detect outliers based on clustering:
  1. Cluster the data.
  2. Find points that don't belong to clusters.
- Examples:
  1. K-means.
  2. Density-based clustering:
     - Outliers are points not assigned to cluster.



http://www-users.cs.umn.edu/~kumar/dmbook/dmslides/chap10_anomaly_detection.pdf

Cluster-Based Outlier Detection

- Detect outliers based on clustering:
  1. Cluster the data.
  2. Find points that don't belong to clusters.
- Examples:
  1. K-means.
  2. Density-based clustering.
  3. Hierarchical clustering:
     - Outliers take longer to join other groups.
     - Also good for outlier groups.



http://www.nature.com/nature/journal/v438/n7069/fig_tab/nature04338_F10.html

Coming Up Next

DISTANCE-BASED OUTLIER DETECTION

Distance-Based Outlier Detection

- Most outlier detection approaches are based on distances.
- Can we skip the model/plot/clustering and just measure distances?
  - How many points lie in a radius 'epsilon'?
  - What is distance to k^{th} nearest neighbour?

- First paper on this topic:

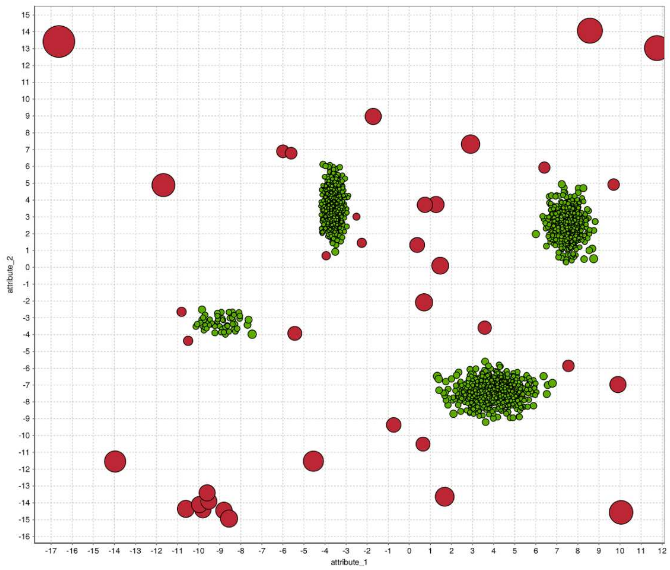
Algorithms for Mining Distance-Based Outliers in Large Datasets

Edwin M. Knorr and Raymond T. Ng
Department of Computer Science
University of British Columbia

Global Distance-Based Outlier Detection: KNN

- KNN outlier detection:
  - For each point, compute the average distance to its nearest neighbours.
  - Choose points with biggest values (or values above a threshold) as outliers.
    - "Outliers" are points that are far from their nearest neighbours.

- Goldstein and Uchida [2016]:
  - Compared 19 methods on 10 datasets.
  - KNN best for finding "global" outliers.
  - "Local" outliers best found with local distance-based methods…

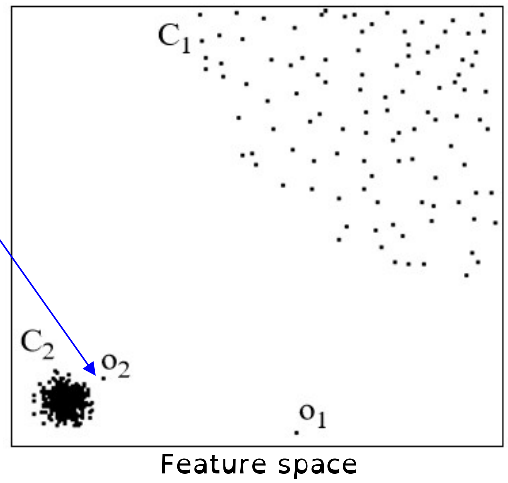


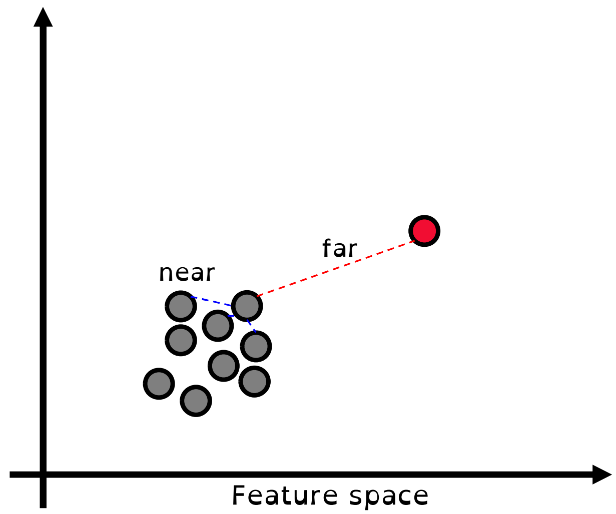
Feature space

Local Distance-Based Outlier Detection

- As with density-based clustering, problem with differing densities:

Q: Is o_2 an outlier? Why?





- Basic idea behind local distance-based methods:
  - Outlier o_2 is "relatively" far
    - compared to how close its neighbours are to one another

http://www.dbs.ifi.lmu.de/Publikationen/Papers/LOF.pdf

Slide 32

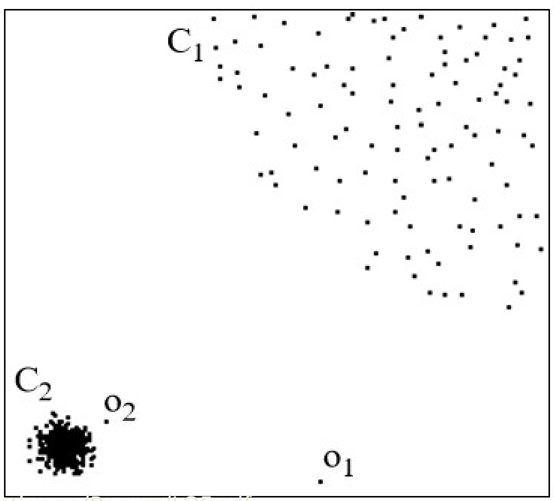
n1 nhgk@student.ubc.ca, 5/18/2021

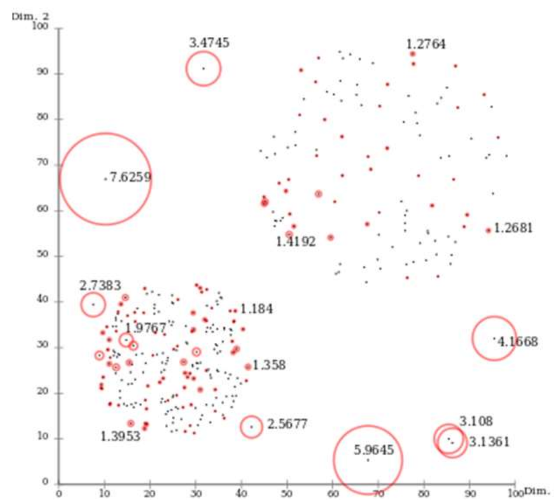
Local Distance-Based Outlier Detection

- "Outlier-ness" ratio of example 'i':

$$\frac{\text{average distance of 'i' to its KNNs}}{\text{average distance of neighbours of 'i' to their KNNs}}$$

- If outlier-ness > 1, x_i is further away from neighbours than expected.

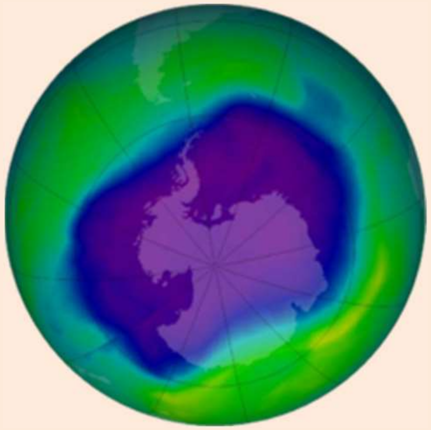




http://www.dbs.ifi.lmu.de/Publikationen/Papers/LOF.pdf
https://en.wikipedia.org/wiki/Local_outlier_factor

Problem with Unsupervised Outlier Detection

- Why wasn't the hole in the ozone layer discovered for 9 years?



- Can be hard to decide when to report an outler:
  - If you report too many non-outliers, users will turn you off.
  - Most antivirus programs do not use ML methods (see "base-rate fallacy")

https://en.wikipedia.org/wiki/Ozone_depletion

Supervised Outlier Detection

- Final approach to outlier detection is to use supervised learning:
    - $y_i = 1$ if x_i is an outlier.
    - $y_i = 0$ if x_i is a regular point.

- We can use our methods for supervised learning:
  - We can find very complicated outlier patterns.
  - Classic credit card fraud detection methods used decision trees.

- But it needs supervision:
  - We need to know what outliers look like.
  - We may not detect new “types” of outliers.

End of Part 2: Key Concepts

- We focused on 2 unsupervised learning tasks:
  - Clustering.
    - Partitioning (k-means) vs. density-based.
    - “Flat” vs. hierarachial (agglomerative).
    - Vector quantization.
    - Label switching.
  - Outlier Detection.
    - Surveyed common approaches (and said that problem is ill-defined).

- We will cover later in course:
  - Recommender systems and improving distance-based methods.
    - Amazon product recommendation.
    - Region-based pruning: fast “closest point” calculations.
    - Shingling: divides objects into parts, matches individual parts of measures part set distance.
    - Frequent itemsets: find items often bought together (a prior is an efficient method).

Part 3: Linear Models

Coming Up Next

LINEAR REGRESSION INTRO

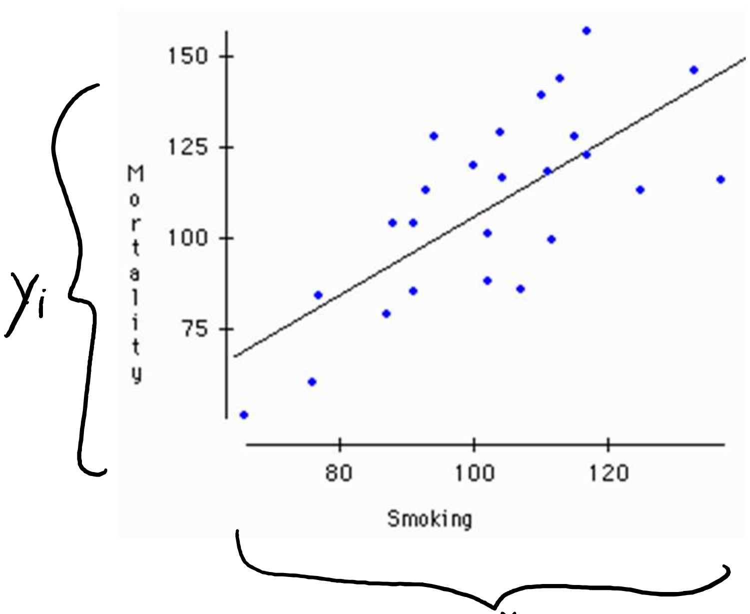
Supervised Learning Round 2: Regression

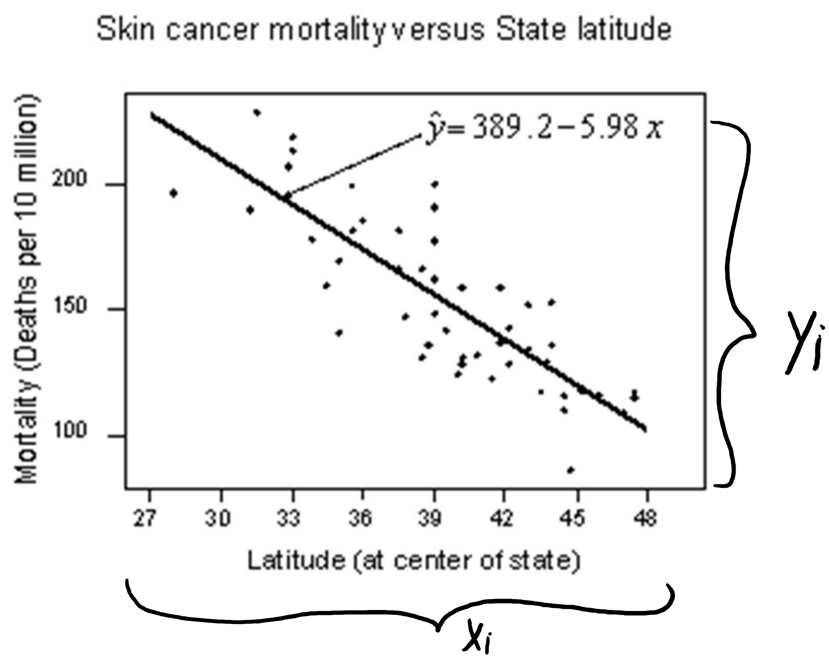
- We're going to revisit supervised learning:

- Previously, we considered classification:
  - We assumed y_i was discrete: y_i = 'spam' or y_i = 'not spam'.
- Now we're going to consider regression:
  - We allow y_i to be numerical: y_i = 10.34cm.

Example: Dependent vs. Explanatory Variables

- We want to discover relationship between numerical variables:
  - Does number of lung cancer deaths change with number of cigarettes?
  - Does number of skin cancer deaths change with latitude?

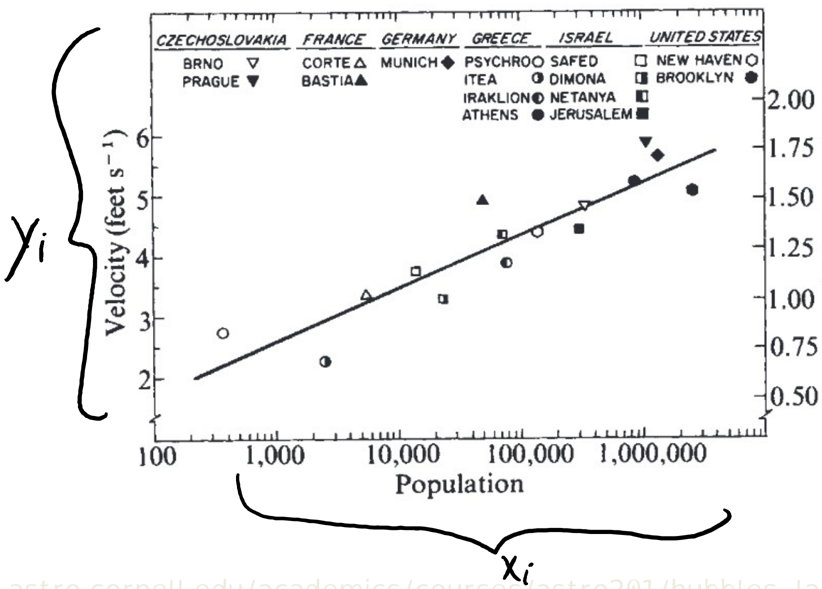


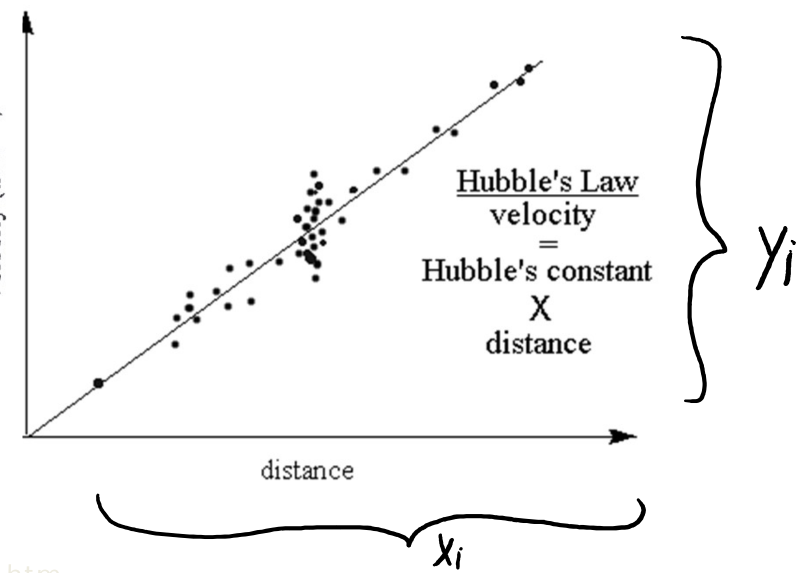


http://www.cvgs.k12.va.us:81/digstats/main/inferant/d_regrs.html
https://onlinecourses.science.psu.edu/stat501/node/11

Example: Dependent vs. Explanatory Variables

- We want to discover relationship between numerical variables:
  - Do people in big cities walk faster?
  - Is the universe expanding or shrinking or staying the same size?

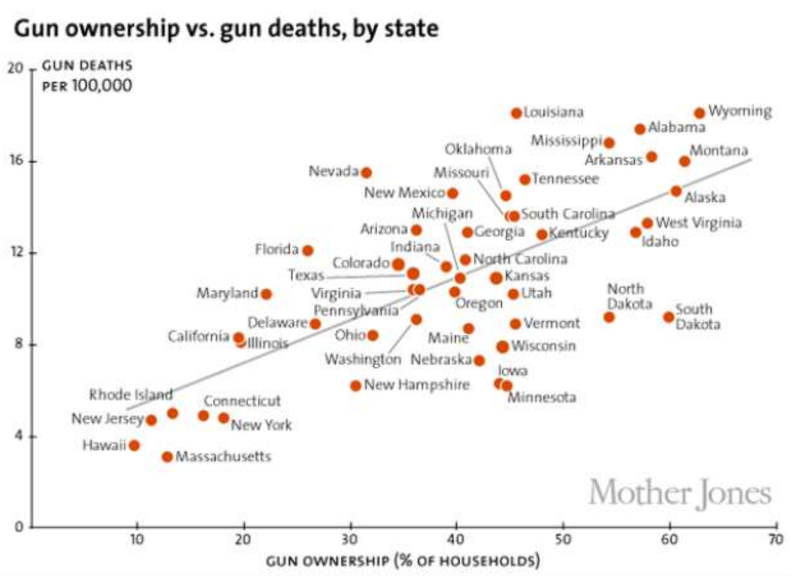


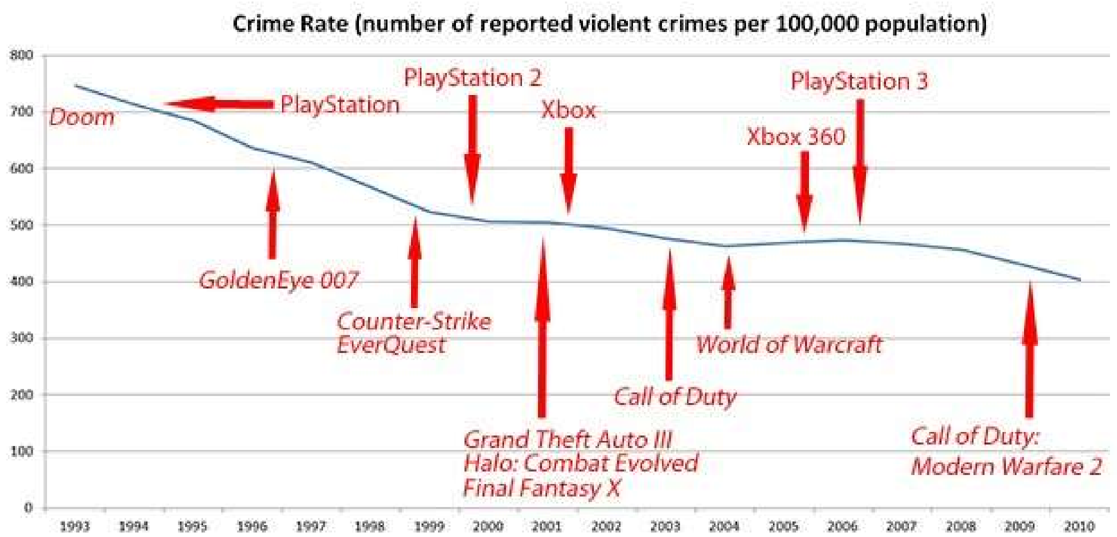


http://hosting.astro.cornell.edu/academics/courses/astro201/hubbles_law.htm
https://www.nature.com/articles/259557a0.pdf

Example: Dependent vs. Explanatory Variables

- We want to discover relationship between numerical variables:
  - Does number of gun deaths change with gun ownership?
  - Does number violent crimes change with violent video games?





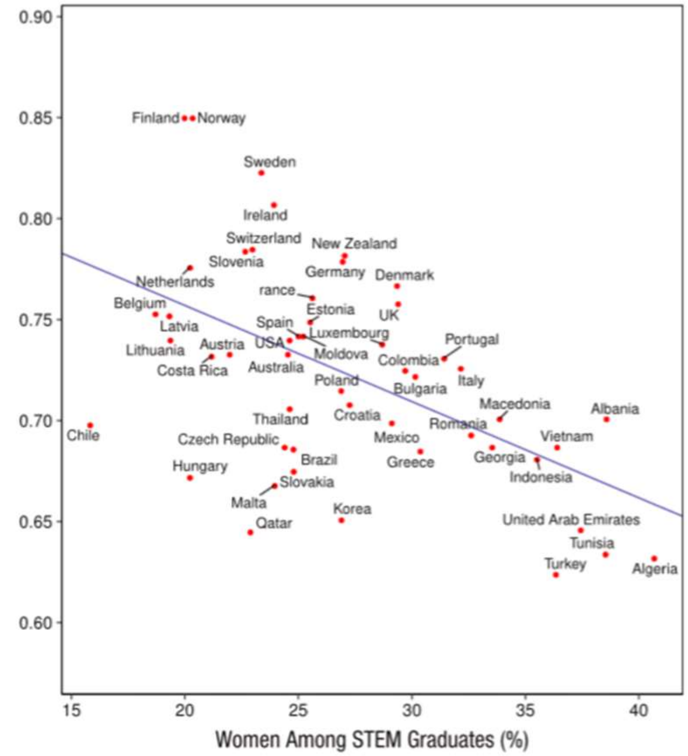
http://www.vox.com/2015/10/3/9444417/gun-violence-united-states-america
https://www.soundandvision.com/content/violence-and-video-games

Example: Dependent vs. Explanatory Variables

- We want to discover relationship between numerical variables:
  - Does higher gender equality index lead to more women STEM grads?

- Not that we're doing supervised learning:
  - Trying to predict value of 1 variable (the 'y_i' values). (instead of measuring correlation between 2).

- Supervised learning does not give causality:
  - OK: "Higher index is correlated with lower grad %".
  - OK: "Higher index helps predict lower grad %".
  - BAD: "Higher index leads to lower grads %".
    - People/media get these confused all the time, be careful!
    - There are lots of potential reasons for this correlation.



https://www.weforum.org/agenda/2018/02/does-gender-equality-result-in-fewer-female-stem-grads/

Handling Numerical Labels

- One way to handle numerical y_i: discretize.
  - E.g., for 'age' could we use {'age ≤ 20', '20 < age ≤ 30', 'age > 30'}.
  - Now we can apply methods for classification to do regression.
  - But coarse discretization loses resolution.
  - And fine discretization requires lots of data.
- There exist regression versions of classification methods:
  - Regression trees, probabilistic models, non-parametric models.
- Today: one of oldest, but still most popular/important methods:
  - Linear regression based on squared error.
  - Interpretable and the building block for more-complex methods.

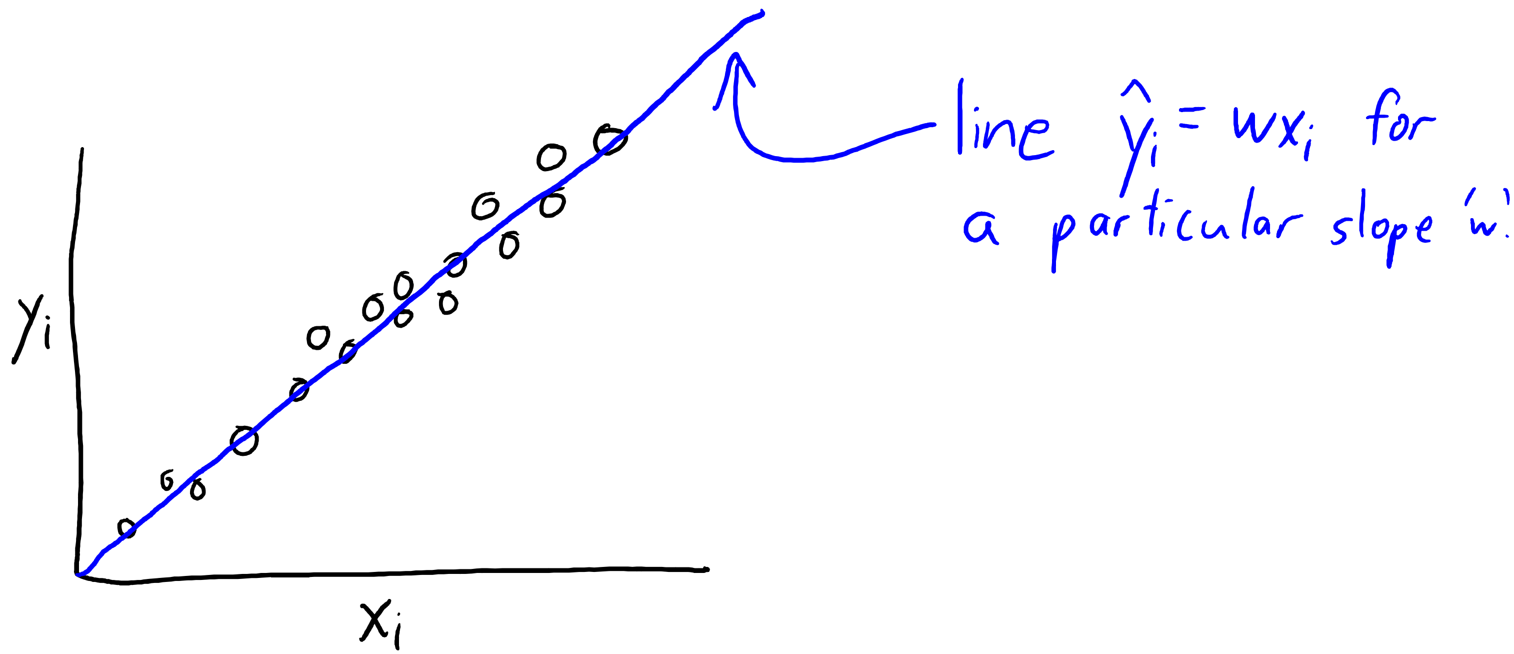
Linear Regression in 1 Dimension

- Assume we only have 1 feature (d = 1):
  - E.g., x_i is number of cigarettes and y_i is number of lung cancer deaths.
- Linear regression makes predictions $\hat{y}_i$ using a linear function of x_i:

$$\hat{y}_i = w x_i$$

- The parameter 'w' is the weight or regression coefficient of x_i.
  - We're temporarily ignoring the y-intercept.
- As x_i changes, slope 'w' affects the rate that $\hat{y}_i$ increases/decreases:
  - Positive 'w': $\hat{y}_i$ increase as x_i increases.
  - Negative 'w': $\hat{y}_i$ decreases as x_i increases.

Linear Regression in 1 Dimension



Aside: terminology woes

- Different fields use different terminology and symbols.
  - Data points = **objects** = **examples** = rows = observations.
  - **Inputs** = predictors = **features** = explanatory variables= regressors = independent variables = covariates = columns.
  - **Outputs** = outcomes = targets = response variables = dependent variables (also called a "label" if it's categorical).
  - Regression coefficients = **weights** = parameters = betas.
- With linear regression, the symbols are inconsistent too:
  - In ML, the data is X and y, and the weights are w.
  - In statistics, the data is X and y, and the weights are β.
  - In optimization, the data is A and b, and the weights are x.

Summary

- Biclustering: clustering of the examples *and* the features.
- Outlier detection is task of finding unusually different example.
  - A concept that is very difficult to define.
  - Model-based find unlikely examples given a model of the data.
  - Graphical methods plot data and use human to find outliers.
  - Cluster-based methods check whether examples belong to clusters.
  - Distance-based outlier detection: measure (relative) distance to neighbours.
  - Supervised-learning for outlier detection: turns task into supervised learning.
- Regression considers the case of a numerical y_i.

- Next time: using linear algebra to tackle linear regression

Review Questions

- Q1: What is the fundamental challenge in automated outlier detection?
- Q2: Why is using Z-score not optimal for outlier detection?
- Q3: How is distance-based outlier detection different from using density-based clustering?
- Q4: What is the problem with the usual reports of “linkage” between variables that we see in the news?

Issues with using z-scores for grades

I definitely sympathize with issues regarding baseline grades in different classes. The ideal solution is to encourage grades to have a standardized meaning across courses, and for courses to have a standardized difficulty, but obviously this is incredibly hard (and probably impossible).

The use of z-scores seems to be a nice solution, but I wanted to point out some potential issues:

1. Z-scores are quite sensitive to outliers. Basically, the mean will be pulled in the direction of outliers, and the variance will be made much larger by outliers. See Slide 8 here: https://www.cs.ubc.ca/~schmidtm/Courses/540-W20/L6.pdf

The major way this manifests is if you have a relatively-small class, and one person just catastrophically fails the course. This has weird effects on the z-score compared to if that person was not in the class: since the average moves lower, people who are slightly below average will actually appear slightly above average. This isn't a big deal, but the more serious issue is that since the variance is made larger the people who are a bit below average will appear very-far below average. (And students well above average get pushed way above average.)

The effect is much smaller in big classes, unless you have a cluster of catastrophic fails and in that case the effect is the same.

There are easy solution to this issue by using statistics based on more-robust measures that allow outliers (for examples, see Slide 9 in that lecture).

2. Z-scores assume the distribution is unimodal. See Slide 10 here: https://www.cs.ubc.ca/~schmidtm/Courses/540-W20/L6.pdf

If you have a group of "good" students and a group of "bad" students, it may reward the good group and punish the bad group more than their grade difference would justify. I think this is a less serious issue, and it's also harder to fix (you would probably need to use historic grade distribution data). In 340, I would expect the grade distribution to roughly look like this.

3. It doesn't address "skew" in the distribution. This could be the case if you have a lot of people at the very top and then the grades drop off slowly from there (another effect I've noticed in 340 grades). Similar to 2, I view this as a less-serious issue than 1 since the shifts probably aren't huge.

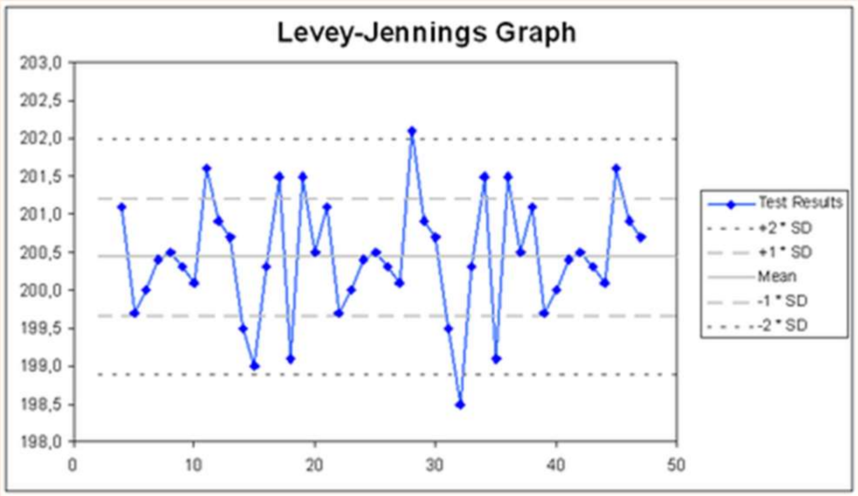
4. If you compare z-scores *across* classes, there is a confounding factor that the students may not come from the same distribution. E.g., one class may attract more strong students and one class may attract more weak students. In a simple setting where only top students take one class and only weak students take another class, the weaker "top" students will be hurt and the stronger "weak" students will be helped.

A simple approach that would address 1-3 is using quantiles. For example, just saying "student A ranked in the top 38% of grades" is simple and avoids some of the issues above. It's not perfect since it doesn't give the real spread (problematic if many students are really close, since it will push them apart). It also doesn't address issue 4, but I would be more comfortable making decisions with this than z-scores. Indeed, my criterion for whether I will write reference letters for students in class is based on ranking rather than absolute score. It's even-more informative to give the class size, like "student A ranked 14 out of 76", but that might be more-difficult to use in automated ways.

For addressing issue 4, you would really need data across classes and I would have to think about whether there is a simple/fair solution.

“Quality Control”: Outlier Detection in Time-Series

- A field primarily focusing on outlier detection is quality control.
- One of the main tools is plotting z-score thresholds over time:



- Usually don’t do tests like “$|z_i| > 3$”, since this happens normally.
- Instead, identify problems with tests like “$|z_i| > 2$ twice in a row”.

https://en.wikipedia.org/wiki/Laboratory_quality_control

Outlierness (Symbol Definition)

- Let $N_k(x_i)$ be the k-nearest neighbours of x_i.
- Let $D_k(x_i)$ be the average distance to k-nearest neighbours:

$$D_k(x_i) = \frac{1}{k} \sum_{j \in N_k(x_i)} \|x_i - x_j\|$$

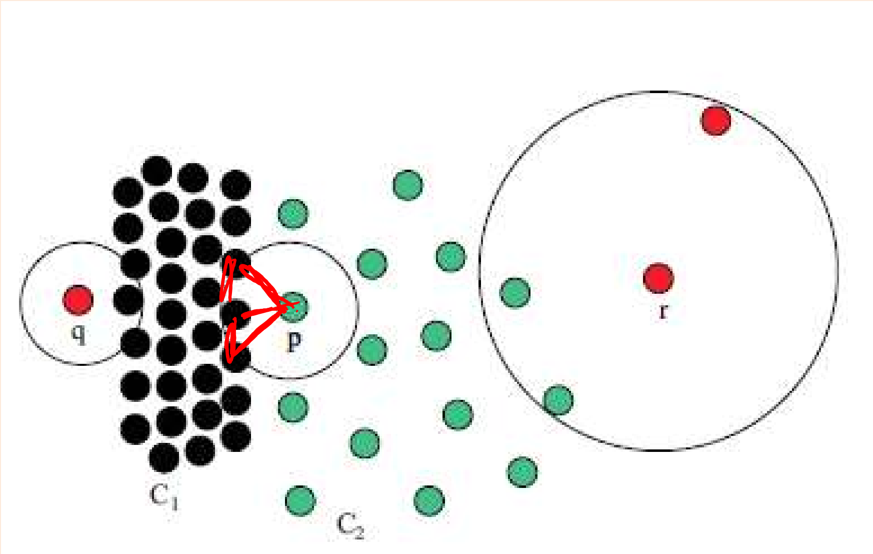
- Outlierness is ratio of $D_k(x_i)$ to average $D_k(x_j)$ for its neighbours 'j':

$$O_k(x_i) = \frac{D_k(x_i)}{\frac{1}{k} \sum_{j \in N_k(x_i)} D_k(x_j)}$$

- If outlierness > 1, x_i is further away from neighbours than expected.

Outlierness with Close Clusters

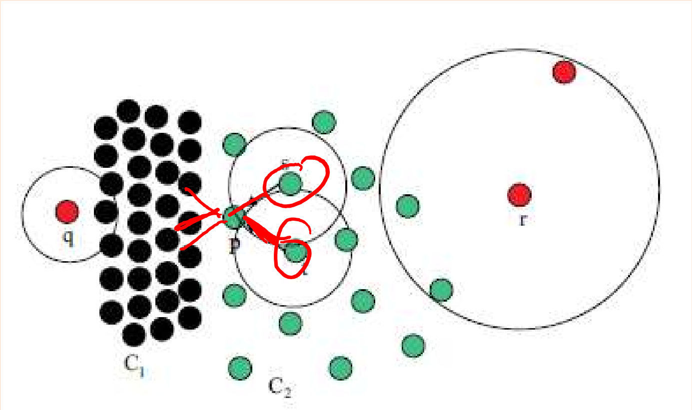
- If clusters are close, outlierness gives unintuitive results:



- In this example, 'p' has higher outlierness than 'q' and 'r':
  - The green points are not part of the KNN list of 'p' for small 'k'.

Outlierness with Close Clusters

- ‘Influenced outlierness’ (INFLO) ratio:
  - Include in denominator the ‘reverse’ k-nearest neighbours:
    - Points that have ‘p’ in KNN list.
  - Adds ‘s’ and ‘t’ from bigger cluster that includes ‘p’:



- But still has problems:
  - Dealing with hierarchical clusters.
  - Yields many false positives if you have “global” outliers.
  - Goldstein and Uchida [2016] recommend just using KNN.

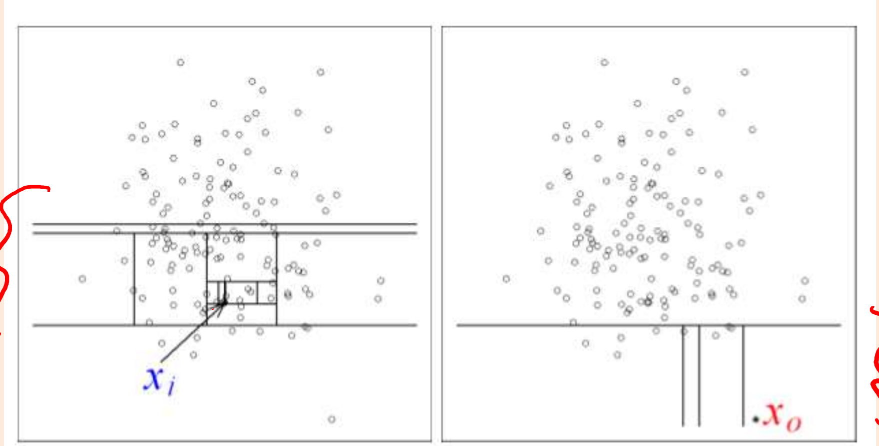
http://www.comp.nus.edu.sg/~atung/publication/pakdd06_outlier.pdf

Isolation Forests

- Recent method based on random trees is isolation forests.
  - Grow a tree where each stump uses a random feature and random split.
  - Stop when each example is “isolated” (each leaf has one example).
  - The “isolation score” is the depth before example gets isolated.
    - Outliers should be isolated quickly, inliers should need lots of rules to isolate.

Depth 12:
– needed 12 rules to isolate so may be inlier.

  - Repeat for different random trees, take average score.



(a) Isolating x_i (b) Isolating x_o

depth 4 so more likely to be outlier

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

Training/Validation/Testing (Supervised)

- A typical supervised learning setup:
  - Train parameters on dataset D_1.
  - Validate hyper-parameters on dataset D_2.
  - Test error evaluated on dataset D_3.

- What should we choose for D_1, D_2, and D_3?

- Usual answer: should all be IID samples from data distribution D_s.

Training/Validation/Testing (Outlier Detection)

- A typical outlier detection setup:
  - Train parameters on dataset D_1 (there may be no "training" to do).
    - For example, find z-scores.
  - Validate hyper-parameters on dataset D_2 (for outlier detection).
    - For example, see which z-score threshold separates D_1 and D_2.
  - Test error evaluated on dataset D_3 (for outlier detection).
    - For example, check whether z-score recognizes D_3 as outliers.

- D_1 will still be samples from D_s (data distribution).
- D_2 could use IID samples from another distribution D_m.
  - D_m represents the "none" or "outlier" class.
  - Tune parameters so that D_m samples are outliers and D_s samples aren't.
    - Could just fit a binary classifier here.

Training/Validation/Testing (Outlier Detection)

- A typical outlier detection setup:
  - Train parameters on dataset D_1 (there may be no "training" to do).
    - For example, find z-scores.
  - Validate hyper-parameters on dataset D_2 (for outlier detection).
    - For example, see which z-score threshold separates D_1 and D_2.
  - Test error evaluated on dataset D_3 (for outlier detection).
    - For example, check whether z-score recognizes D_3 as outliers.

- D_1 will still be samples from D_s (data distribution).
- D_2 could use IID samples from another distribution D_m.
- D_3 could use IID samples from D_m.
  - How well do you do at recognizing "data" samples from "none" samples?

Training/Validation/Testing (Outlier Detection)

- Seems like a reasonable setup:
  - D_1 will still be samples from D_s (data distribution).
  - D_2 could use IID samples from another distribution D_m.
  - D_3 could use IID samples from D_m.

- What can go wrong?

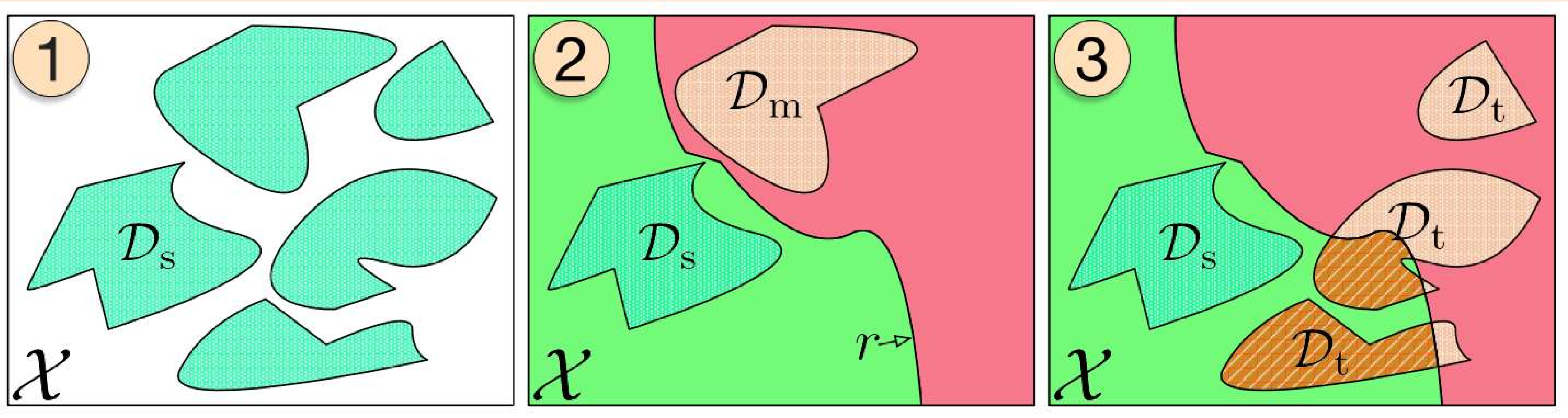
- You needed to pick a distribution D_m to represent "none".
  - But in the wild, your outliers might follow another "none" distribution.
  - This procedure can overfit to your D_m.
    - You can overestimate your ability to detect outliers.

OD-Test: a better way to evaluate outlier detections

- A reasonable setup:
  - D_1 will still be samples from D_s (data distribution).
  - D_2 could use IID samples from another distribution D_m.
  - ~~D_3 could use IID samples from D_m.~~
  - D_3 could use IID samples from yet-another distribution D_t.

- "How do you perform at detecting different types of outliers?"
  - Seems like a harder problem, but arguably closer to reality.

OD-Test: a better way to evaluate outlier detections



- “How do you perform at detecting different types of outliers?”

https://arxiv.org/pdf/1809.04729.pdf