Tools & Models for Data Science

Outliers

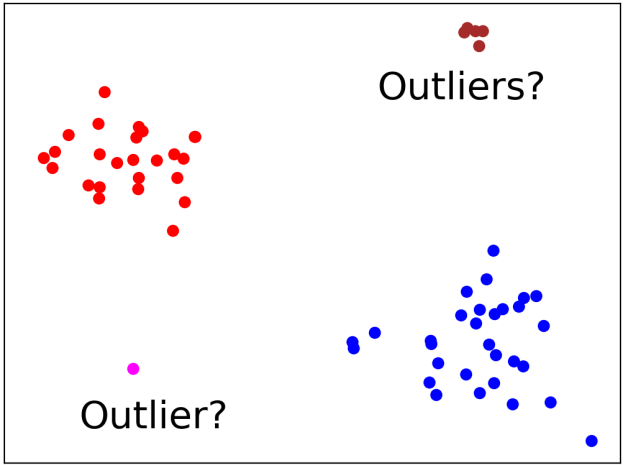
Chris Jermaine & Risa Myers

Rice University

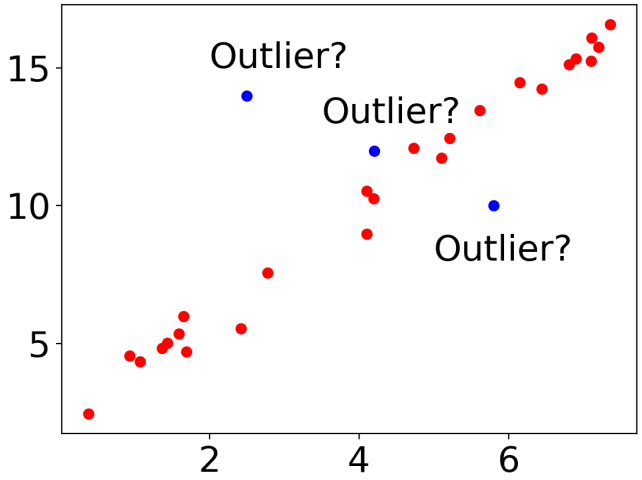
What Are Outliers In Data Science?

- Data points that are unlike the other points, unexpected, or unusual in some way
  - Weather data set: low of -12°C degrees in Houston
  - Blood pressure > 300 mmHg
  - Drug price change from $13.50/pill to $750/pill

Outlier Pic: 1



Outlier Pic: 2



Outliers Not Same As Unusual Data

- Low of 12°C degrees in Houston unusual, probably not an outlier
- Low of -12°C in Chicago (a bit) unusual, probably not an outlier
- It's the combination of Houston, -12°C that makes this difficult to believe

Why Look For Outliers?

- Classically, two reasons:
  1. Classically, to remove them from data...
     - ...because outliers can hurt learning process

Why Throw Out Data?

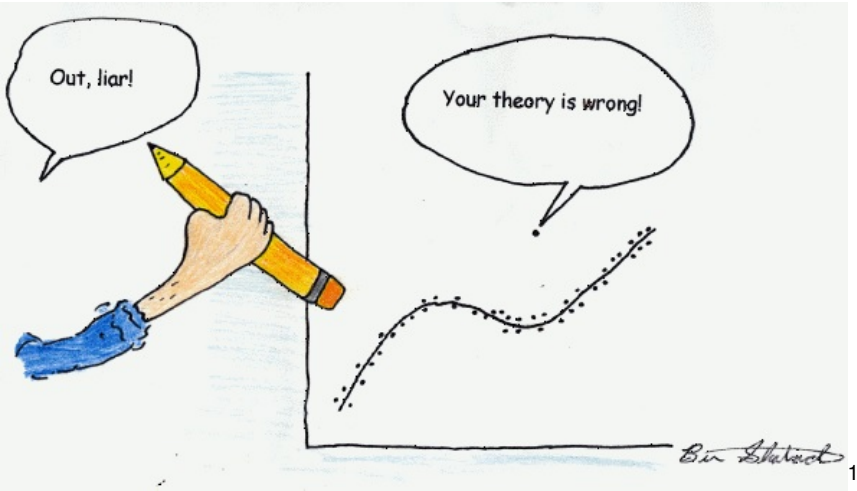
- It messes up the model
- Garbage in, garbage out
- Huge impact on least squares
  - Most common loss / error model
  - Outliers have a magnified effect

distance	distance^2
1	1
2	4
-3	9
-4	16
5	25

Is it Really Garbage Data?

- Values that defy natural laws
  - Negative blood pressure values
  - Speed faster than light
- Sometimes it's meaningful
  - Age 0 in an adult ER
  - Value 99 to indicate unknown
- It's a slippery slope

But This Can Be Dangerous



[1]

Change your model to include the outlier!

[1] Reprinted with permission `http://davidmlane.com/ben/cartoons.html`

Why Look For Outliers?

- Classically, two reasons:
  1. Classically, to remove them from data...
     - ...because outliers can hurt learning process
  2. To find them for further examination...
     - ...because outliers might enhance understanding of data

? What are some examples of applications where we want to find outliers?

The Data Scientist's Goal

- Find the outliers!
  - Computer security
  - Fraud detection
  - Medical crisis alerts

? What if you find 1 MD prescribing lots of opioids. Is it fraud?

The Data Scientist's Goal

- Find the outliers!
  - Computer security
  - Fraud detection
  - Medical crisis alerts
- What if you find 1 MD prescribing lots of opioids. Is it fraud?
  - Or is it a large practice and all the Rx's go under 1 name?

Supervised vs. Unsupervised Learning

- Supervised learning – labels are provided
- Unsupervised learning – no labels provided

Outlier Detection is an Unsupervised Task

? Why?

Outlier Detection is an Unsupervised Task

- Why?
  - We don't always know what the outliers look like
  - By definition, we don't expect them
  - They are rare
- Supervised learning requires labels

How Do We Define Outliers?

- Two standard definitions:
  - (1) Distance-based
  - (2) Model-based

Distance Based Outlier Detection

- Classical approach
- Most frequently used
- Feasible - if we have a high dimensional space, outliers are very far from the other points

Model Based Outlier Detection

- Requires a proper statistical model
- If a data point has a very low probability of occurring, flag that
  - Coin flips:
  - 10 flips: 9 heads
  - ? Is our coin biased?

Model Based Outlier Detection

- Requires a proper statistical model
- If a data point has a very low probability of occurring, flag that
  - Coin flips:
  - 10 flips: 9 heads
  - $10 * \frac{1}{2}^9 * (1 - \frac{1}{2}) = 10 * \frac{1}{2}^{10} = 0.00977$
  - 10 * P(heads) * P(Tails)
  - Your coin is likely biased

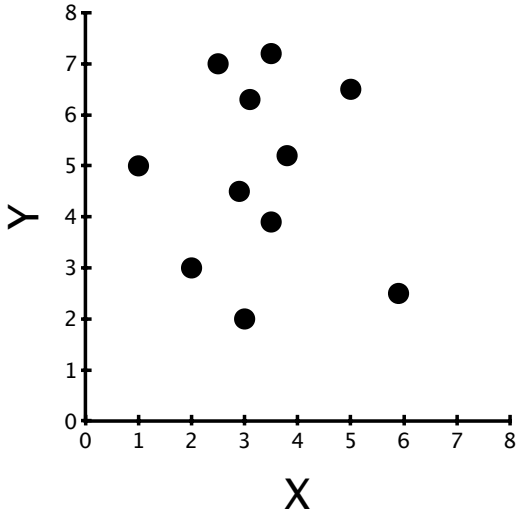
kNN

- k Nearest Neighbors
- Very popular machine learning algorithm
- Relatively simple
  - Look at a data point
  - Consider the k closest points to it
  - Make a decision based on that information

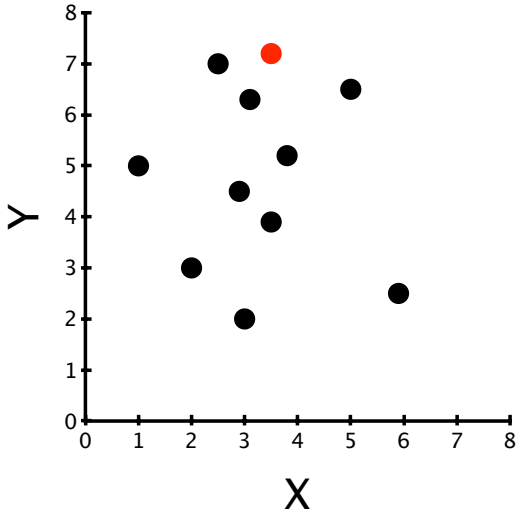
Distance-Based Outliers

- Definition: A point is an outlier if it is far from all other points
- Outlier search often defined in terms of kNN:
  - Let $d(x_i)$ be the distance to point x_i's $k$thNN in the data set
  - Then, given data set $\langle x_1, x_2, ..., \rangle$, we want to compute the set O of outliers such that...
  - $|O| = m$ the number of points in the set is m
  - and $\forall (x_o \in O, x_i \in \mathbf{X} - \mathbf{O}), d(x_o) \geq d(x_i)$
  - That is, for every point x_o in our outlier set
  - and **every** point x_i **NOT** in our outlier set
  - the outlier distance is further than the non-outlier distance

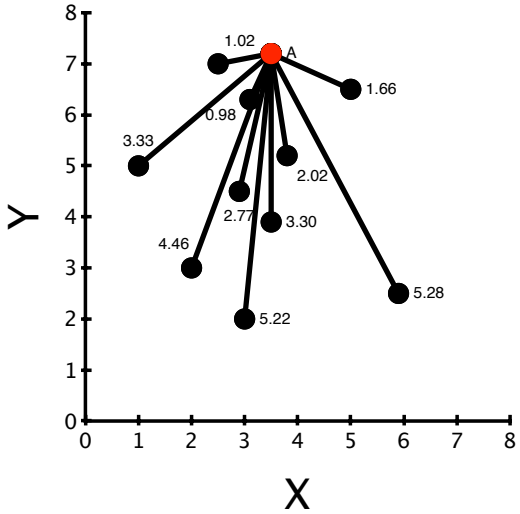
How to Find Distance-Based Outliers, Conceptually (2NN)?



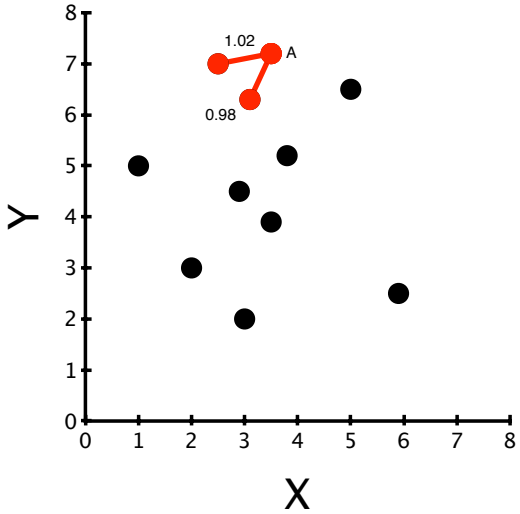
Pick a Point (2NN)



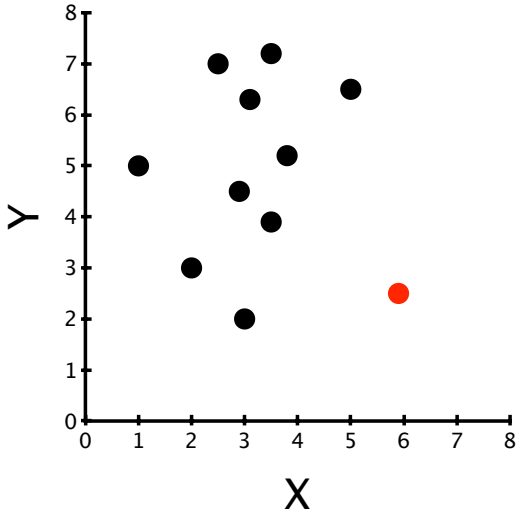
Compute Distance to All Other Points (2NN)



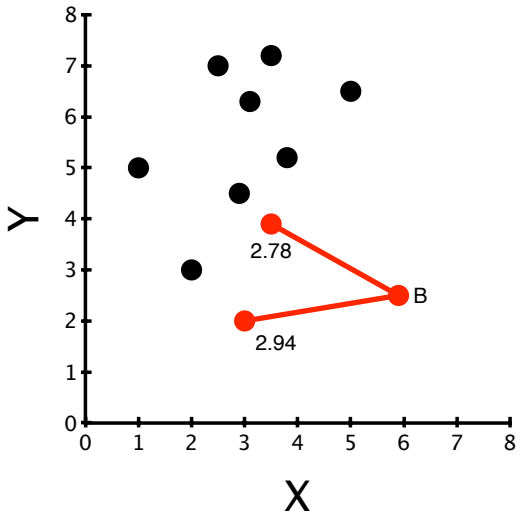
Find the Distance to the 2nd NN (2NN)



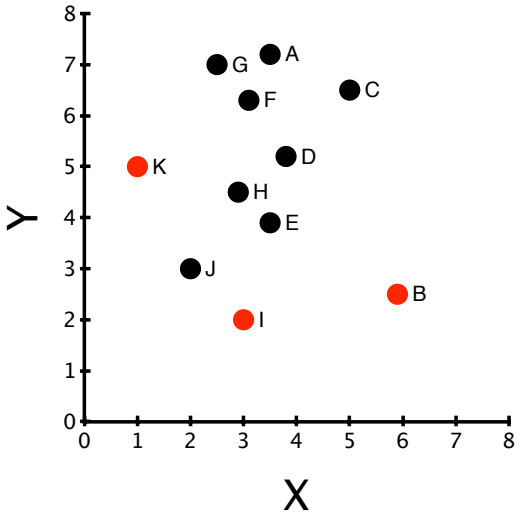
Repeat for All Points



Finding the 2nd NN Distance



Select the Top m "Outlierly" Points



Point	2NN distance
A	1.02
B	2.94
C	1.77
D	1.30
E	1.33
F	0.98
G	1.02
H	1.14
I	1.96
J	1.75
K	2.24

How To Find Distance-Based Outliers?

- Simple algorithm:
  - O is the set of current Outliers
  - Q is the closest distances to the current point

```
init min-priority queue O
for x1 ∈ X:
  init max-priority queue Q
  for x2 ≠ x1 ∈ X:
    insert dist(x1,x2) into Q
    if |Q| > k
      remove max from Q
  insert x1 into O with key max(Q)
  if |O| > m
    remove point with min key from O

return O
```

- m is the number of outliers we are looking for
- k is the nearest neighbor we consider
- Let's walk through an example in detail

How To Find Distance-Based Outliers?

- So, this works!
- ? Are there any problems here?

How To Find Distance-Based Outliers?

- What are the problems here?
  - m & k are magic/arbitrary numbers
  - Nested loop thru entire database O(N^2)
  - Too slow for big data
  - How to address (the computational problem)?

How To Find Distance-Based Outliers?

- Better algorithm:
  - O is the set of current Outliers
  - Q is the closest distances to the current point

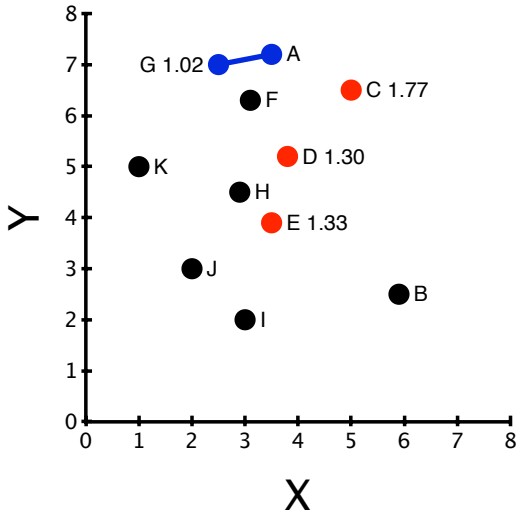
```
init min-priority queue O
for x1 ∈ X:
  init max-priority queue Q
  for x2 ≠ x1 ∈ X:
    insert dist(x1,x2) into Q
    if |Q| > k
      remove max from Q
    if |Q| == k and |O| == m and max(Q) < min(O)
      discard x1; not an outlier
  insert x1 into O with key max(Q)
  if |O| > m
    remove point with min key from O

return O
```

- Q & O must be full
- If I find a neighbor who is closer than the current outlier distance, I **can't** be a top outlier
- So move to the next candidate point

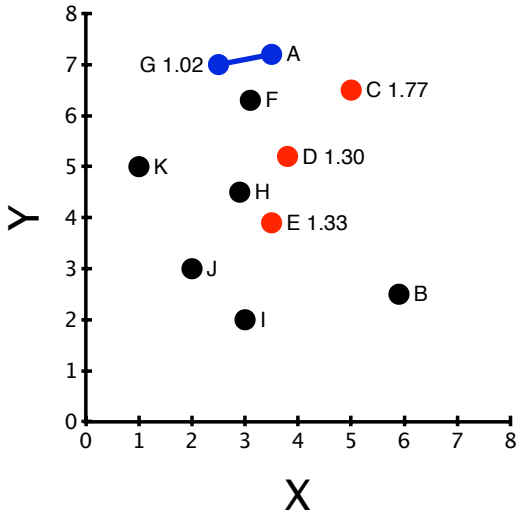
How To Find Distance-Based Outliers?

- Why does this help?
  - $\texttt{max}(Q)$ is an upper bound on distance to kth NN
  - So distance to kth NN can't ever be greater
  - If this is not good enough to get point into top m in O
  - Then can discard it early



How To Find Distance-Based Outliers?

- Example: O: {(1.33, E), (1.77, C) (1.30, D) }
  - Process F
  - Find 2NN
    $distance(F, A) = 0.98$
  - $0.98 < 1.30$ so discard F
- Can get a 100x speed up
- But, it's still $O(N^2)$



How To Find Distance-Based Outliers?

- Even better:
  - Store X in randomized order
  - That way, won't get unlucky and find all far points first

How to Choose m and k?

- What goes into m?
  - Size of the data set
  - Ability of humans to process / deal with the identified outliers
  - Start small and make m bigger. Stop when nothing “interesting” is added
  - Very application specific
- What goes into k?
  - Size of the data set
  - How far apart are your data? k “smooths" the data.
  - Try $\sqrt{N}$
  - Determine empirically using your validation set

Model-Based Outliers

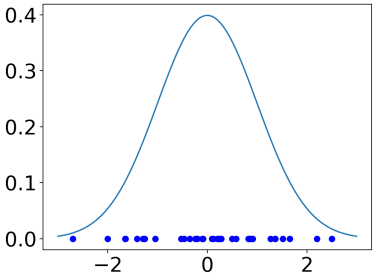
- Basic idea:
  - Learn a model for what is “typical”
  - Then outlier is data point with low score according to the model

Example of Model-Based Detection

- Learn parameters of a Normal distribution by choosing μ, σ^2 to maximize

$$P(x_1, x_2, ..., x_n) = \prod_i \left(\text{Normal}(x_i|\mu, \sigma^2)\right)$$

- Then choose m points with highest $\text{Normal}(x_i|\mu, \sigma^2)$ PDF
- Whichever points are not described well by the model are considered outliers



- Most likely data occurs in the middle, points in the tails are less likely

Use Case Real-time Detection

- Write software to sound an alarm
- Use existing dataset to identify a new outlier
- Alarm fatigue (high number of False Positives) can be a huge problem

Use Case Retrospective Analysis

- Find outliers in a big dataset
- Learn a model
- Considering all the points, choose the most “outlierly”
  - Anesthesiologist No. 7 [2]
  - Study to evaluate pre-op condition’s (reduced blood flow to the heart) impact on post-op heart attack
  - Local hospital (Texas Heart Institute)
  - One anesthesiologist was afraid of bradycardia
  - Also was the least experienced anesthesiologist
  - Factor of 10 difference

[2] Slogoff S, Keats AS. Does perioperative myocardial ischemia lead to postoperative myocardial infarction? Anesthesiology. 1985;62(2):107–14.

Wrap up

1. What is an outlier?
2. Why does it matter?
3. Review of detection algorithm using kNN

? How can we use what we learned today?

? What do we know now that we didn't know before?