# Agenda:

This week, we’ll continue with basic machine learning algorithms, this time focusing on clustering models. We’ll also cover couple of cross-cutting concepts, including distance norms and the important topic of model validation. As usual, you’ll have a hands-on homework assignment.

**Sub-courses**

1. **Validation**
2. **Clustering(K-means)**
3. **Supervised vs. Unsupervised**

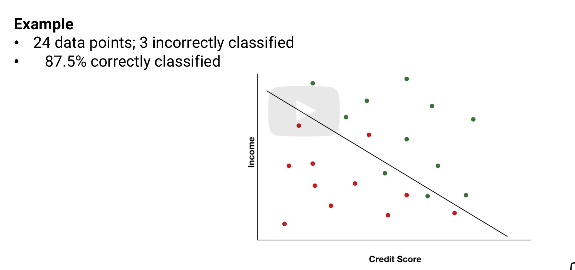
# 1.1.Validation

* check to see how good a model is

**Examples**

1. How good it predict loan applicant
2. Predict future sales
3. How often a Liver transplant will transmit disease

Model Accuracy is high in the training data set as the model is trained for that dataset



* Above accuracy of 87.% is on a high end(optimistic).
* - The test results or real time data will have lesser accuracy

This is also due to the **Patterns**

Two kinds pf pattern

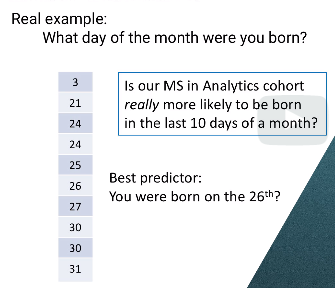
1. Real effects: real relationship b/w attributes and results
2. Random : random , but looks real

**Challenge:** cannot distinguish real and random

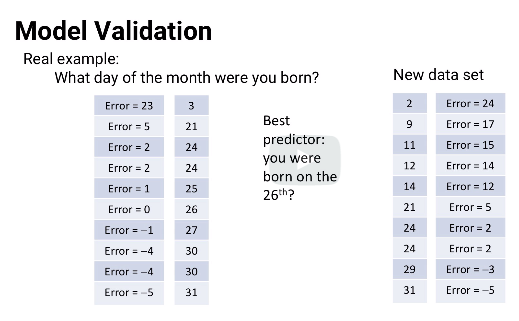
* model is good on validation data as the random effects will be applicable only to validation set and will fade away with real data

**Example**

Day of birth of students in analytics data



* Prediction is wrong.
* Just random
* Model’ effectiveness should be measured on different set of data



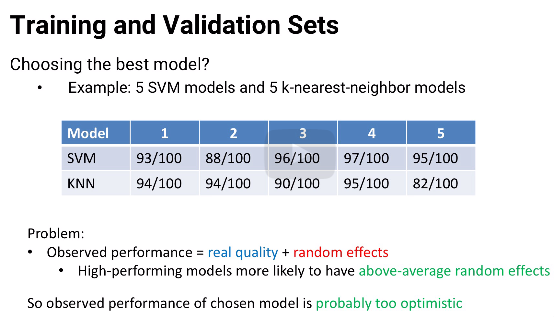
In sample above, random is measured by taking mid point(26) and calculation error as subtraction of mid point

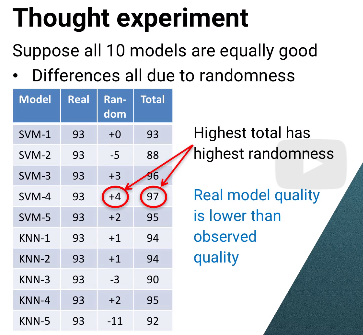
# Validation and Test data sets

1. Use two different data
2. larger set to fit model (Training set); smaller set to measure model’ effectiveness(validation)

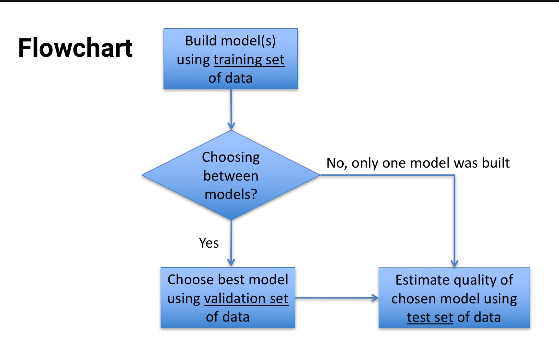
**Scenario#1:** More than one model:

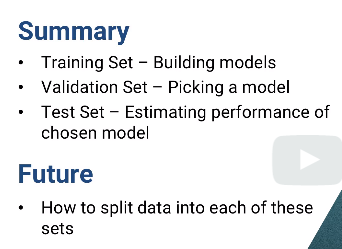
Best performing is #4 of SVM: But is it the best?. Most random effects model are high performing.

****

in diagram, the best performing model is inflated by randomness, not realness

**How to overcome?** Use third set, called “Test data”





# 1.3 Splitting Data

## How much data to use for each?

1. 1 model : training and test data

**70-90% for training; 30-30% test**

1. More than 1 model : Training, validation and test
2. Training(50%), validation(25%) ;test(25%)
3. Training(60%), validation(20%) ;test(20%)
4. Rule of thumb : 50-70% for training; remaining split equally for test and validation

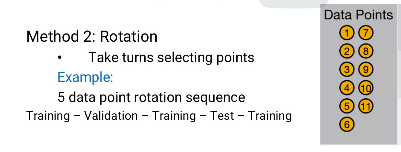
## How to split data?

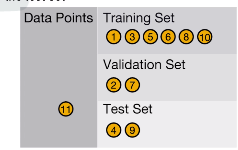
Example: we have 1000 data points

1. **Random**

* choose random 600 for training
* remaining 400 ( 200 each for validation and test)

1. **Rotation**

****

* Meaning out of 5 data points, training is after each different data set
* Results:
* ****

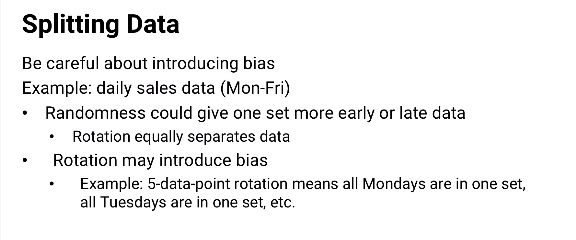
**Advantage of rotation:**

1. Each part of equally separated
   1. Especially with 10 years worth of data
   2. all data is equally separated in rotation
   3. vs. random, there might be a change that training will get year 1 and 2 and miss year 10th, where a condition might have changed

**Disavd of rotation:**

Unexpected bias

-assuming 5 point rotation in scenario of 5 day retailer shop, training data will not include day 2 and day 4 for training



**Last Option :**Combined Approach :60% of Monday data in training, remaining 40% in test and validation; repeat for all the days

# Cross validation

**Why is it required?**

=>“new” patterns are present only in validation and test data

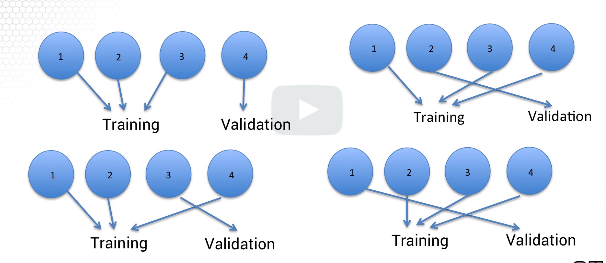
=> model doesn’t know about it and doesn’t get trained on it

**Solution:** cross- validation

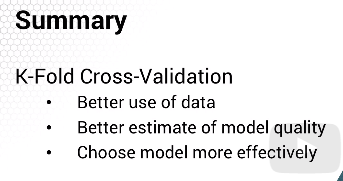
## K Fold Cross validation

If k=4

* Training & Validation will be split into 4 parts
* While training model, we use 3 split dataset and validate on 4th so each data point occurs 3 times in a set
* To choose best of the models, take average
* K=10 is fairly common



* For SVM, using k fold , how to choose model? We cannot average co-efficients; So with SVM, after k-fold, train the model again with all four parts together to get the final model.



# 2.1.CLUSTERING

-categorize the data points to groups that are close to each other

Examples:

1)target marketing , each group gets customized messaegs

2) if selling suv, for some people we highlight price, other the size; each set is cluster

3) personalized medicine ; how they react to a treatment

4) location civic facility, cluster group of people and establish fire station, coffee shops

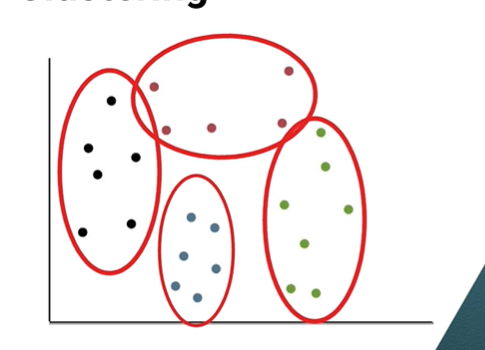
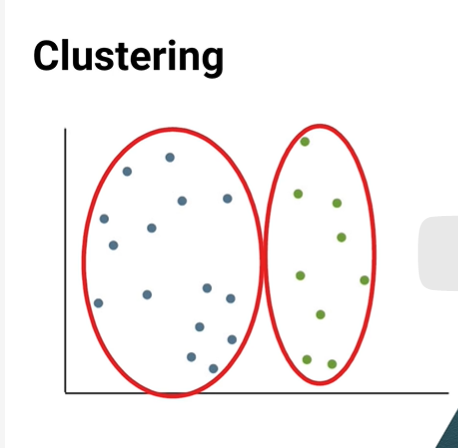
5) image analysis ; online CAptha..the picture might get more ambigious as other people as getting better

6) prediction/data investigation

From cluster, we might find inituitive groups

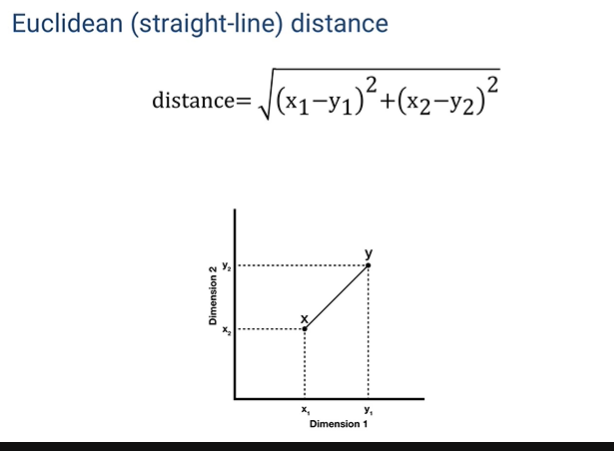
Graph: shows suv buyers

x-axis :age; y-axis : no of children



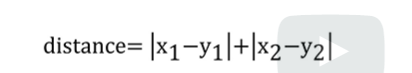
# 2.2 Distance Norms

Distance between two points: squareroot of pythagorus theorem

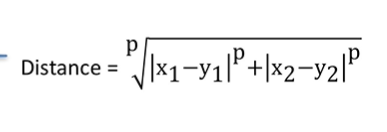


Recti-linear distance: used for find in cities that fit a grid model.

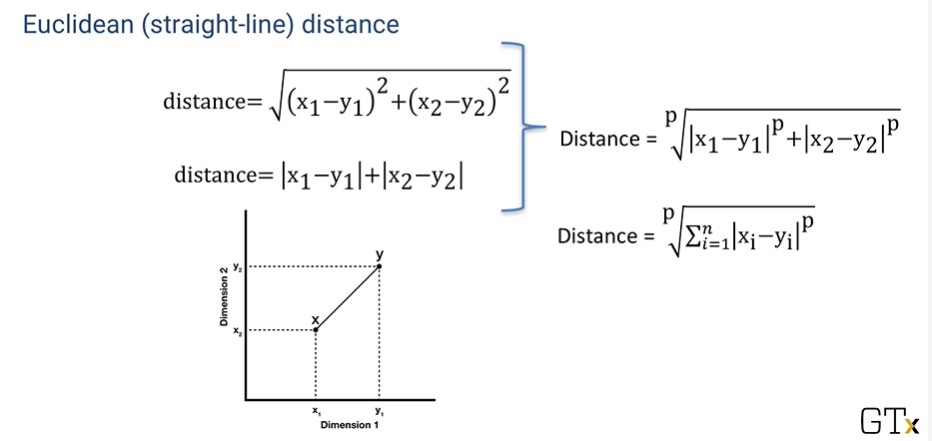
-also called “Manhattan distance”



Consolidated



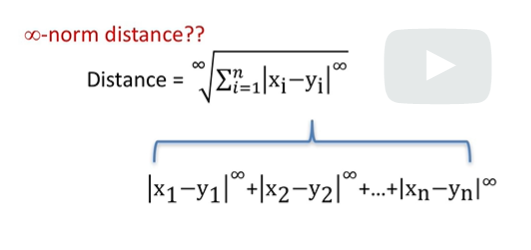
P=1 for rectilinear distance and p=2 for straight line distance (Euclidean)



This is called “P-norm” distance or “monkowski” distance.

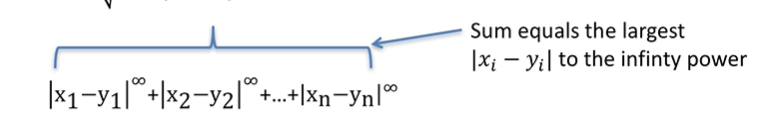
P can be any value

-very popular is “Infinity norm” distance.



Assumng there are n points, the formula looks like above when expanding.

However, “2” to the power infinity is not close to 2 to the power inifinity. So formula comes to the highest difference between x and y co-ordinates. As the square root cancels power, it is simply the highest distance



**Example:**



If the machine can have up and down movement and also horizontal movement to reach the package.

1. In “1” norm, the machine moves to the correct location and then stretches its hand
2. In infinity norm, machine moves and stretches simultaneously, which means the time taken is based on the longest one , whichever is the highest number

# 2.3. K Means clustering

- clustering algorithm

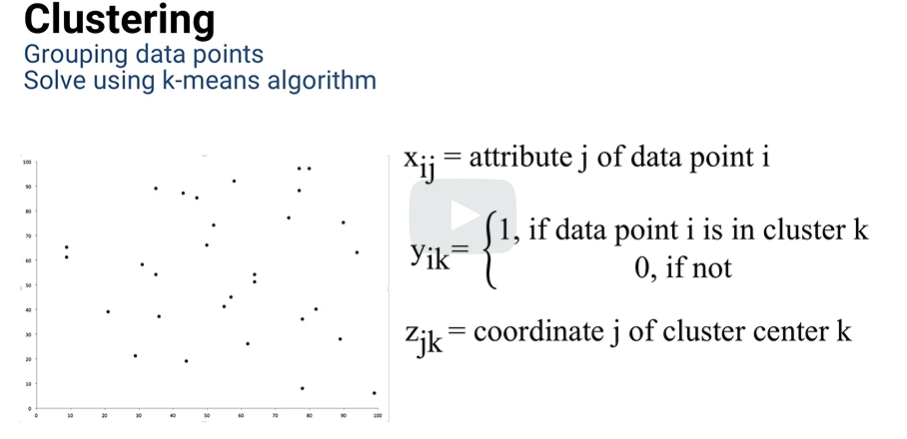
-more thatn 60 years

Example: x-axis:age; y-axis average daily temperature

**Use case: daily sales of SUV**

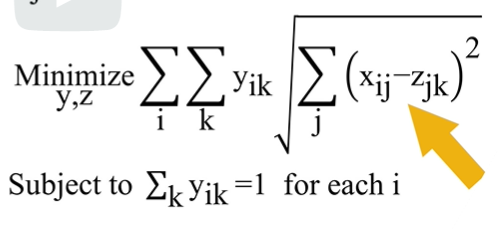
Data point: n

Number of attributes : m

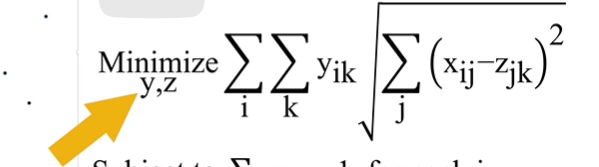


Purpose: find “k” cluster centers

Then assign each data points to the cluster center that minimize the distance from the data point to the cluster center.

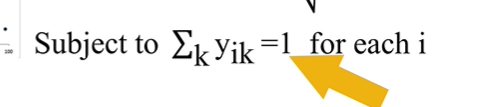


The piece shown in arrow shows the distance of all the data points from its cluster center.



But the arrow highlighted part ensures that the data point is part of the cluster center.

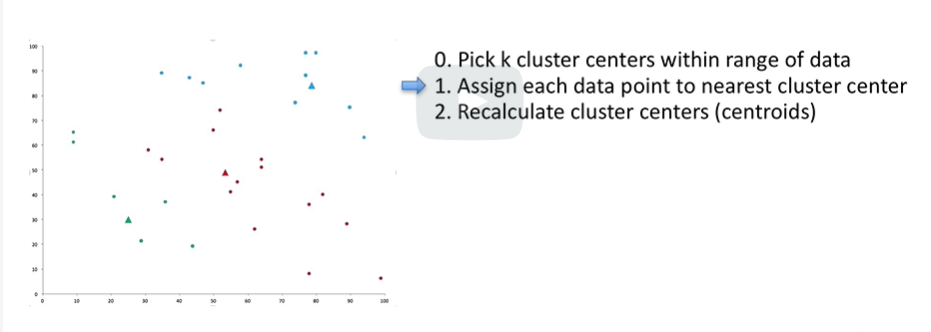
Finally, below arrow noted piece ensure the data point is assigned to only one cluster



This is a hard optimization problem to solve. So we use K means algorithm

**Approach:**

1. Pick number of cluster, which is “k”.this are cluster center
2. Temporarily assign cluster to each data point
3. But, the centers we picked are not really the cluster, so we recalculate.
4. For each cluster, we find centroid of the cluster and that is our new cluster center
5. Then, we reassign data points back to new center



So we go in loop again and again until we know no data point change cluster.

Doing in hand in 2- dimension; for more dimension ,k-means is best

**Jargons:**

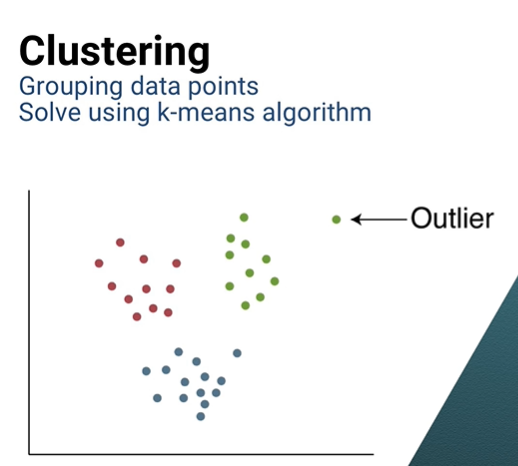
1) this is heuristic : not guarantee for best solution. But get close to best solution; it is fast

2) it is also “expectation maximization”(EM)

- we alternate expectation step, but finding center and cluster repeatedly

-maximise ; assign data to center; meaning we are maximization the negative of distance to the cluster center

# 2.4 practical details of K means



How to handle outlier

1. Remove outlier and run the algorithm so it doesn’t drag the cluster center biased to one side
2. Best :investigate what it means (combination of data, alogorithm and real scenario to find the right answer)

**Disadv**: K means is heuristic

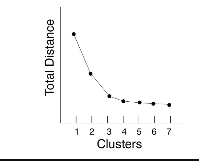
**Advantage** : speed

So to handle heuristic, run it different times with different starting initial cluster center

-use algorithm’ speed to pick the right number of clusters

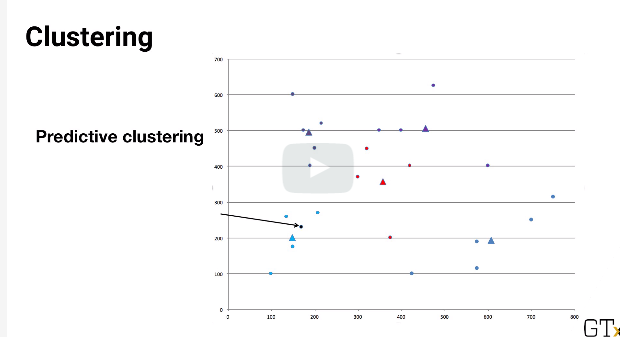
**Is more clusters better than fewer?** (in extreme case, it might one cluister for each data point)

1. Qualitative : If city needs 4 new fire stations, then that gives the size of the cluster
2. Quantitative: find k means for different k

 Elbow chart : the curve shows that from that point, the addition of clusters will not [[1]](#footnote-1)be helpful

# 2.5 Clustering for prediction

- use K means in predictive way (meaning where will be new data point fit?)



-if new point is inside cluster, the answer is straightforward

- If new point is not in any cluster, we choose the cluster which has the center closer to the new point

**Question:** What range of possible data point can be assigned to a cluster?

**Ansner:** each cluster can include data point that is close to itself that other clusters

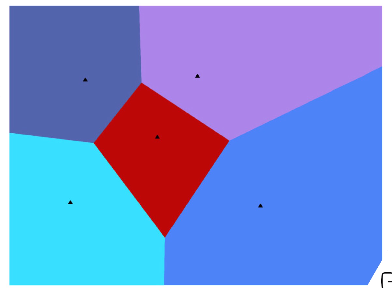
 It called voronoi diagram named after mathematician (100 years ago).

Diagram like this was used to analyze the outbreak of cholera in London

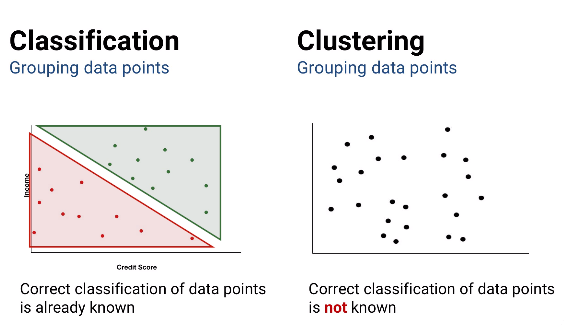
Not all ideas are new and cutting edge.the best of them have been around a long time and just needs to be dusted off

# 2.6 Supervised vs. unsupervised

Classification and Clustering are same and does the same thing

What is the difference?

1. In classification, we have input and output - Supervised
2. In clustering, we know only input, no output - Unsupervised



\*Most algorithms are Supervised

1. [↑](#footnote-ref-1)