# Analytical Modeling (ISYE6501)

Investigators:

Principal Investigator: Yakut Gazi, Ph.D.

Co-Principal Investigator: Stephen Harmon, Ph.D.

Co-Principal Investigator: Shawn Staudaher, Ph.D

TA:

**Alex Arnon, Ozan Aygun:** Piazza gurus, the people the rest of us ask if we don't know how to answer a question.  They work together to make sure that none of your posts slip through the cracks and all get answered in a timely way.  (At least, if you follow the [Communication Guidelines](/class/koqwy2g8qw55n1?cid=6) and keep your posts visible by all TAs.)

**Daniel Rodgers:** head proctor, reviews exams that our proctors flag for potential misconduct and fields technical difficulties.  (If your small child runs screaming into the room during your exam, he is the one that makes sure it doesn't get held against you.  If your computer catches fire while you're taking a test, he's the one that figures out how to keep it from affecting your grade.)

**Chase Nicholson:** head grader, handles the dirty work, makes sure everyone gets grades on their homework assignments and projects, makes sure complaints about uneven or inaccurate grading are investigated.  (At least, if you follow the Homework Guidelines and include the appropriate explanation + file your challenge on time.)

**Tim Pollard:** head of integrity, handles the *really* dirty work, investigates complaints/concerns about submissions that appear to be copied from another source.  (It's an ugly fact of large online courses that someone, somewhere, will attempt to cheat.  Thankfully, Tim reviews the evidence to make sure the honest students win out in the end.)

**Pin Hsu:** head TA, gets all the glory, handles logistics and important course announcements (like this one), decides what to do in weird emergencies and one-off cases.  (If a TA replies to your post with "I've notified the head TA and he'll update you when he's made a ruling on your case", congratulations!  You have won the 'weird situation' lottery.  Past winners include folks who got dropped from the course due to Bursar error, that one guy who was drafted to fight in a war mid-semester, and most everyone who was traveling somewhere when COVID-19 hit.)

**Patrick Kriengsiri:** head of office hours, makes sure no one is left totally in the dark about the homework assignments.  (At least, if you attend the office hours or [watch the videos later](/class/koqwy2g8qw55n1?cid=8).)

# \\ Agenda:

Couple of basic machine learning algorithms.

cross-cutting concepts about data and scaling.

**Instructor:**

Joel Sokol

**Shortcuts**

“M” (modeling), “C” (cross-cutting concept), or “X” (experiential),

**Analytics:**

**Use case:**

across business, sports team pilots,

To compile, build and forecast

**why is it used?**

To answer

1. descriptive questions (what happened? why this happened, what the impact of a change? (is the type of analysis of data that helps describe, show or summarize data points in a constructive way such that patterns might emerge that fulfill every condition of the data.)
2. Predictive question: what is going to happen?
3. Prescriptive questions: what action(s) would be best?

Prescriptive Analytics is **a form of advanced analytics which examines data or content to answer** the question “What should be done?” or “What can we do to make \_\_\_\_\_\_\_ happen?”

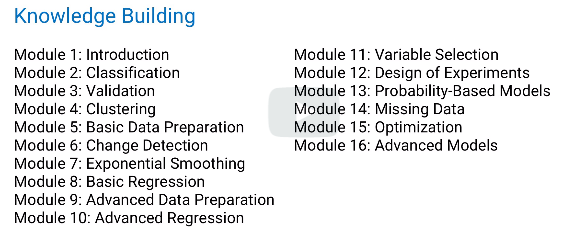
1. General question: new technology’ impact

**Modeling**

* Take a real life situation and expressing this situation in math
* Analyze math
* Turn math answer back to real life solution

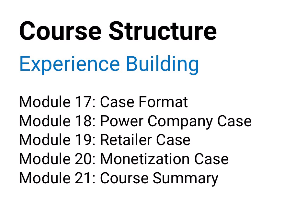
# Structure of the course

**Part-1**

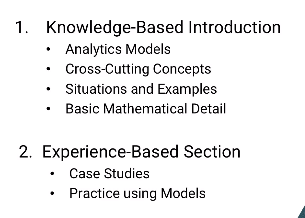
****

* Models
  + ML
  + Regression
  + Optimization, etc
* Cross-cutting
  + Data preparation
  + Output quality
  + Missing data

**Part-2 experience building**

****

Summary:



# [\\WEEK1](file:///\\WEEK1)

# Module 1: Model

**3 different ways**

* regression to predict delivery date of packages company ships out ;This case regression is a model
* use regression to predict delivery date based on size, weight and distance; This case regression+all the detail=model
* it can be a linear equation with 81\*(sum of linear dimension)+76\* weight+4\*distance; then that is also a model

# Module2: Intro to classification

## Intro to classification

* putting things to category

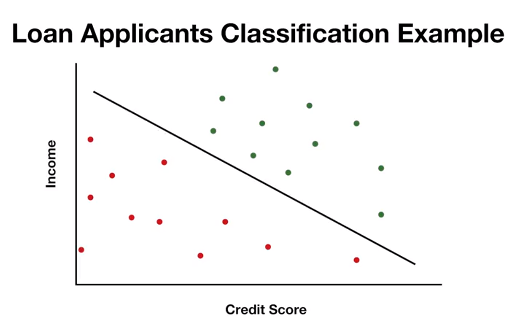
**use cases**

1. answer yes or no question
   1. bank : customer who will fully repay the loan or not
   2. security agency : differentiate regular vs. terrorist
   3. email filter l :spam or not
   4. legal doc : relevant or not
   5. cdc : is this organ safe to transmit : if donor is recent infected ; test is not positive
2. having more than 2 category is possible
   1. politics : support, opssition voters and undecided
   2. paleantologist : newly found dinosaur across existing ones

**example** : bank loan application

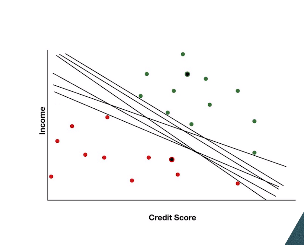
collect data = income, credit score, more attributes => length of time of customer in bank,number of dependants, total assets and liabilities

green :repaid ; red: defaulted



**Challenge:**

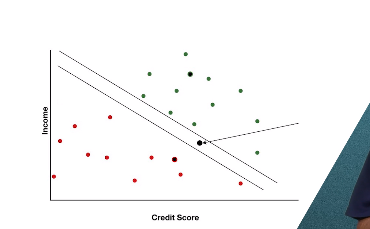
1. more than one line possible (like below): how to decide?
2. What happens to the new entry point that is in middle



## Choosing a classifer:

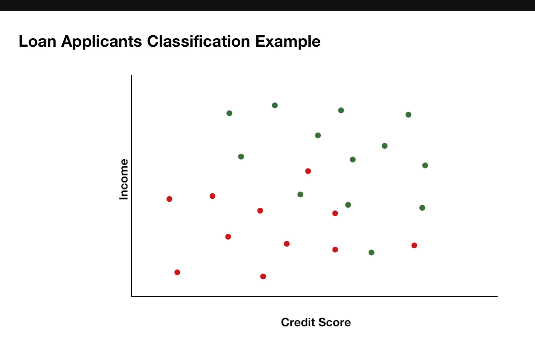
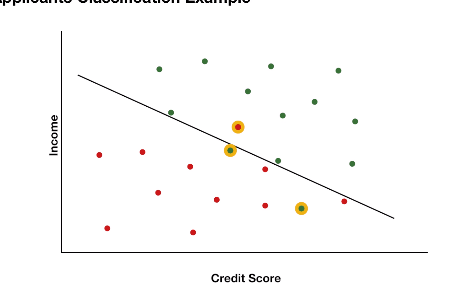
**Agenda**: tradefoffs in classification

\*consider two lines as classifers

 Lower classifier (low line) will suggest we give the loan and higher classifier (top line) will suggest , we deny the loan

**Suggestion:**

1. Outlier shouldn’t cause classicization error
2. Use soft classification for below example

1. **Minimize error:** We can look to reduce total mistakes ; trade off actual mistakes and near mistakes
2. Priority of mistakes (cost of making mistake in green side > cost of making mistake in red side)

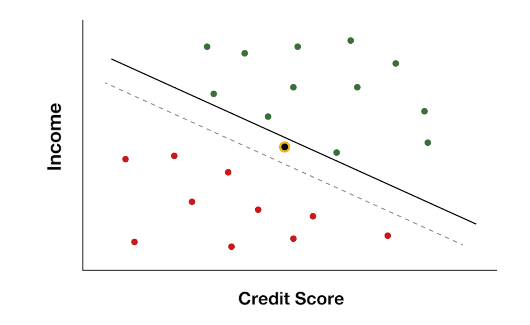
Meaning,

cost of giving loan to wrong candidate is more costlier than cost of not granting loan to right candidate

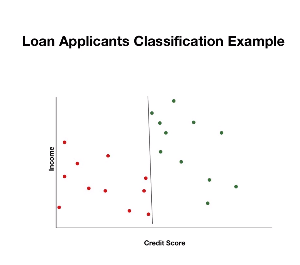
cost of eating a poisonous food is more costlier/worse than cost of not eating a edible food

**suggestion:** The more costly one type of bad decision is,the more we want to move the line away from it.

Like graph below

**** If new yellow point appear, though its closer to green, it will deny the loan/suggest not to eat the food

If data looks like this

this means, only horizontal axis impacts the results, not the vertical axis

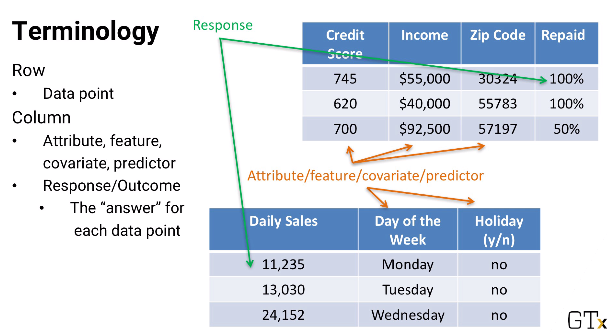
## Data types

1. Mental mode: table ; every row is a data point; i.e single point of observation

Each applicant is data point (every row)

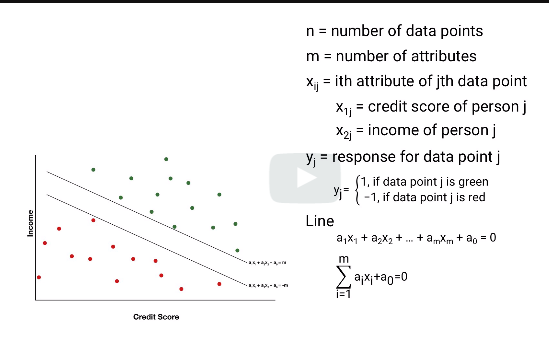
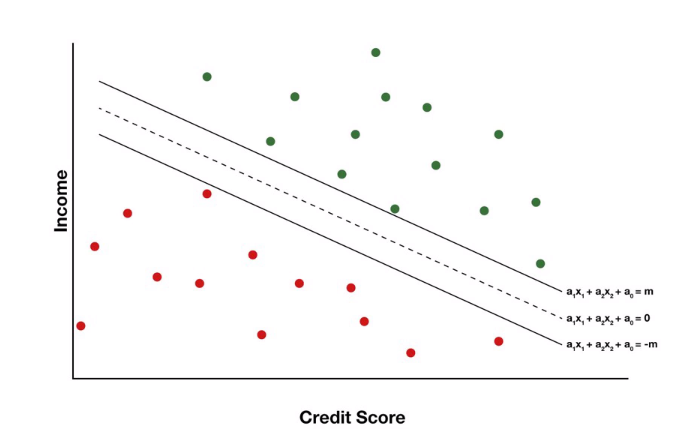
Column is attribute; feature, covariate, predictor

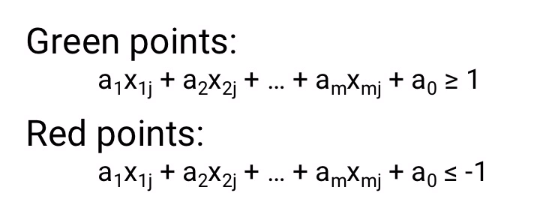
Response: answer to our question



1. Structured vs. unstructured (use cases: police officer incident report; twitter)
2. Structured
   1. Numbers (quantitative)
   2. Numbers are categorical as well . E.g zipcodes,year
   3. Binary (categorical) – M/F; day can be holiday or not; (sometimes treated as quantitative)
   4. Unrelated data vs. time series data(same data recorded over time; usually equal interval; e.g ‘ daily stocks logged every 2 min, height of a child recorded every year; also possible for time series is not recorded with same interval)

## Support vector machines (SVM)

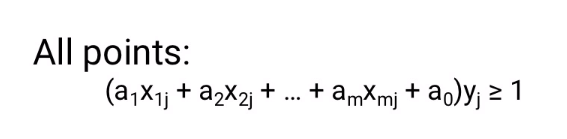


1 : is the “y” results of data point green;

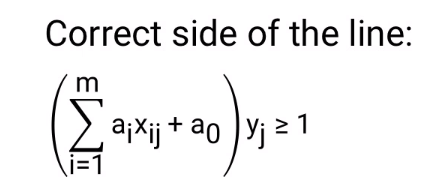
-1 : is the “y” result of data point red:

We don’t use 0 for red and 1 for green. Instead we use +1 and -1 for results to obtain generic results for all points

**Generic equation of all points**

****

Below is the simplified version of this expression



-we need to find value of a0 throguh a(m) so the two parallel lines are far from each other

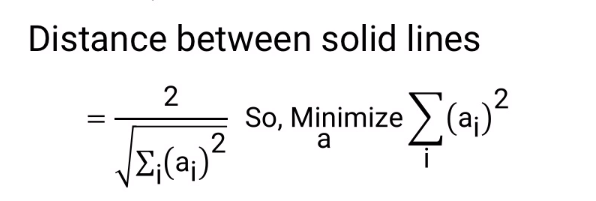
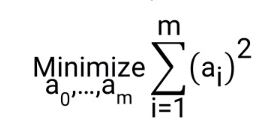
Parallel line are two lines with same a(i)x(i) but different A0 (intercept)

Middle dotted line is our classifer

(+m) line is for upper classifer(green) and (-m) is lower classifer (red)

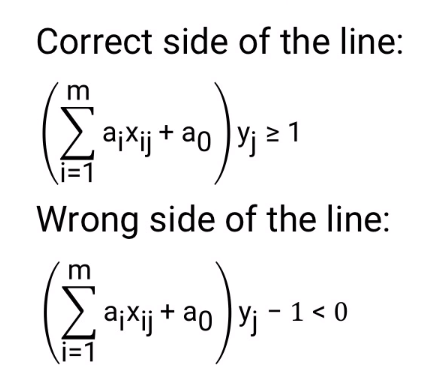
### Objective#1

-our **objective** is to maximize this distance(2/square root of sum of square of a), which is same as minimize the sum of squares of a

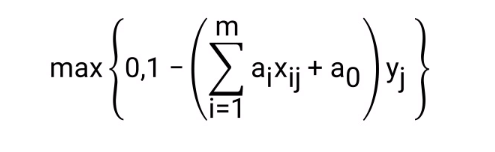
Minimize error and maximize margin

For each data point “J”, we have to find the below value. If on correct side of the line, each data point will get results of this equation as

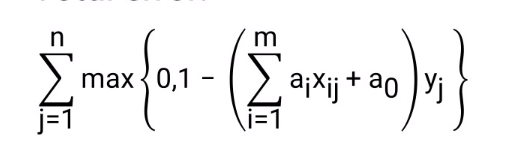
1.  If it's on the wrong side, then the sum minus 1 is less than zero and the amount that it's than zero is the amount of error.

### Objective#2

Error for each data point “j” can be written as below. Our objective is to minimize this error

 =max is because , it gives the farther distance the data point is from the classifer.

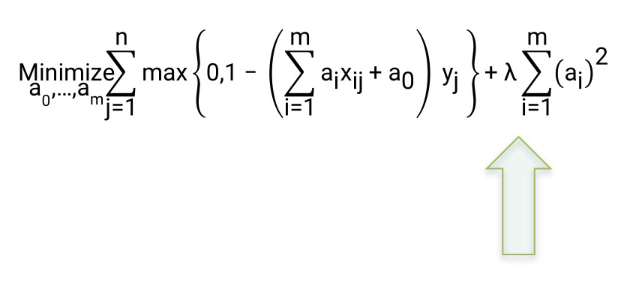
Converting above formula for all data points (1 thorugh n as n is number of data points)



### Overall objective (combining #1 and #2)

For setting priority:

1. as lambda gets large, second part of getting more distant gets larger which outweighs the errors of data points being assigned to wrong category.
2. As lamba is small, the distance is small between the upper and lower classifier.so the priority of assigning the data points to the corrrect category takes precedence.

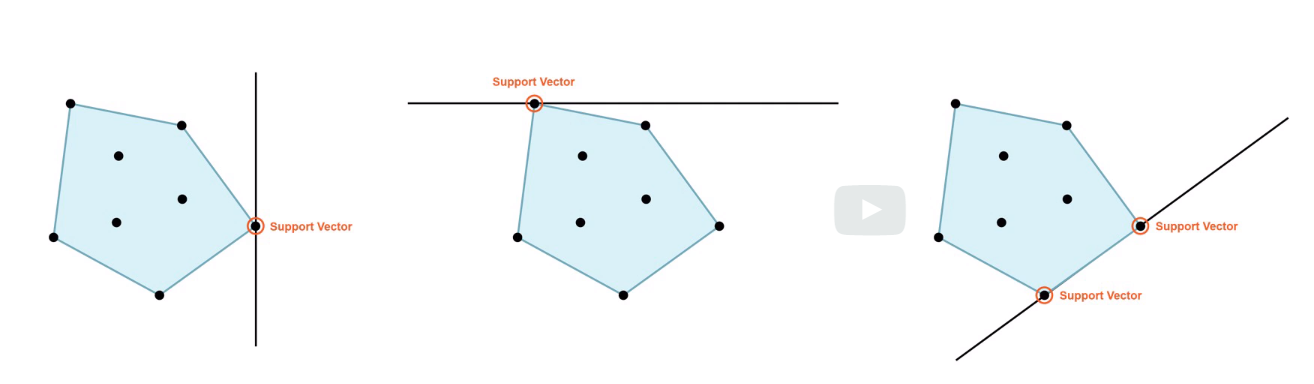


## Why name “Support vector”

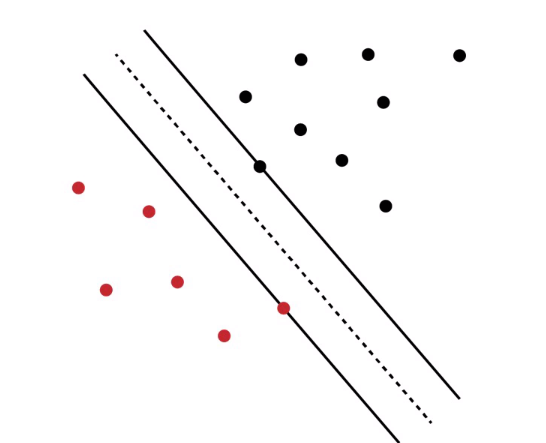
Convex hull : when you join the outer point to make a shape

Convex hull needs a support from bottom or top or side..there can be more than one support vector for a convex hull

Technically , it is support point, but each point is vector of data



In our algorithm, we are looking to draw a two parallel line that should be far away from each other and should touch one of the points of each category. Classifier passes through the middle of each support vector and it doesn’t touch both the support vectors



## Advanced SVM

**Agenda:**

extensiobn of SVm

Classification errors more costly than others

Prepare data before running SVM

Other classifier to use

### Scenario#1:

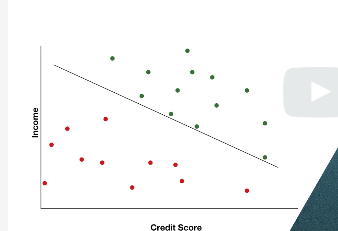
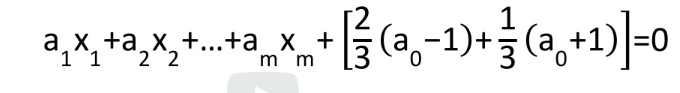
Classification errors more costly than others

Say, Giving loan to someone who wouldn’t pay is more costlier than withholding a loan from eligible candidate

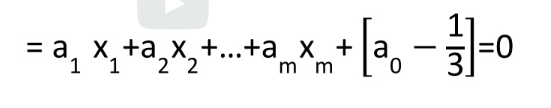
**Solution :** shift classifier away from the costlier category

**Logic:**

Technically intercept a(0) can extend from a(0-1) to a(0+1). In this case , we assign more weightage to a(0-1) than a(0+1). Like 2/3 of weightage

**Long Form** ****

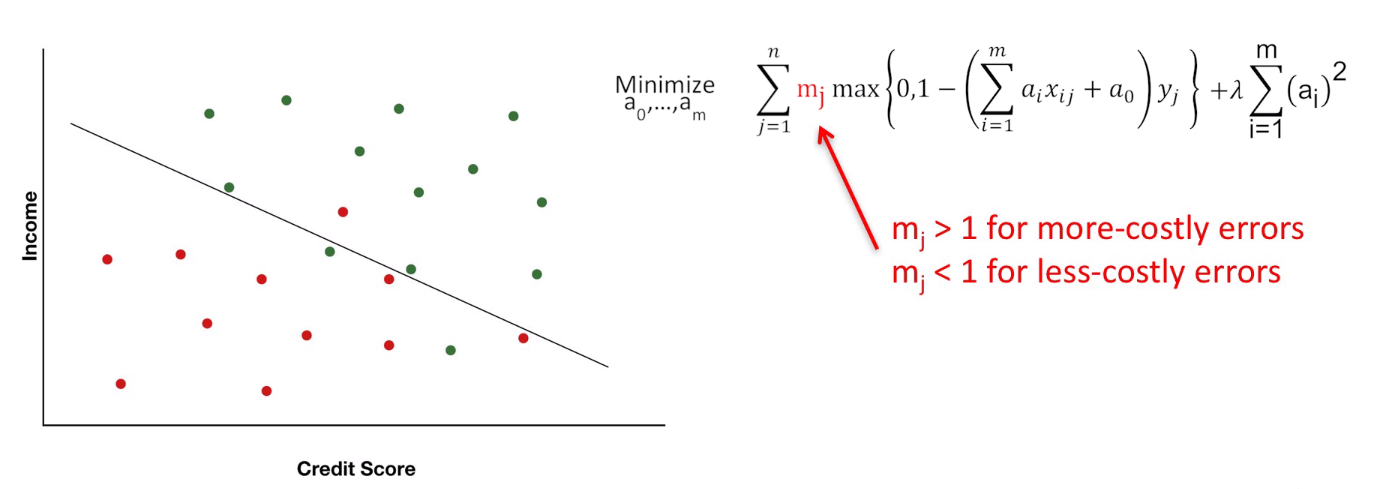
Short form after solving equation is

****

### Scenario#2:

Soft classification/Classification errors more costly than others

Add m(j) to errors.>1 for more costly errors and <1 for less costly erros



### Scenario#3:

Scaling issue

Credit score : range from 300-800.new ones in 100

Income : range in difference of millions

Our objective is to maximize the gap between lines , meaning minimize the sum of square of “a” co-efficients.

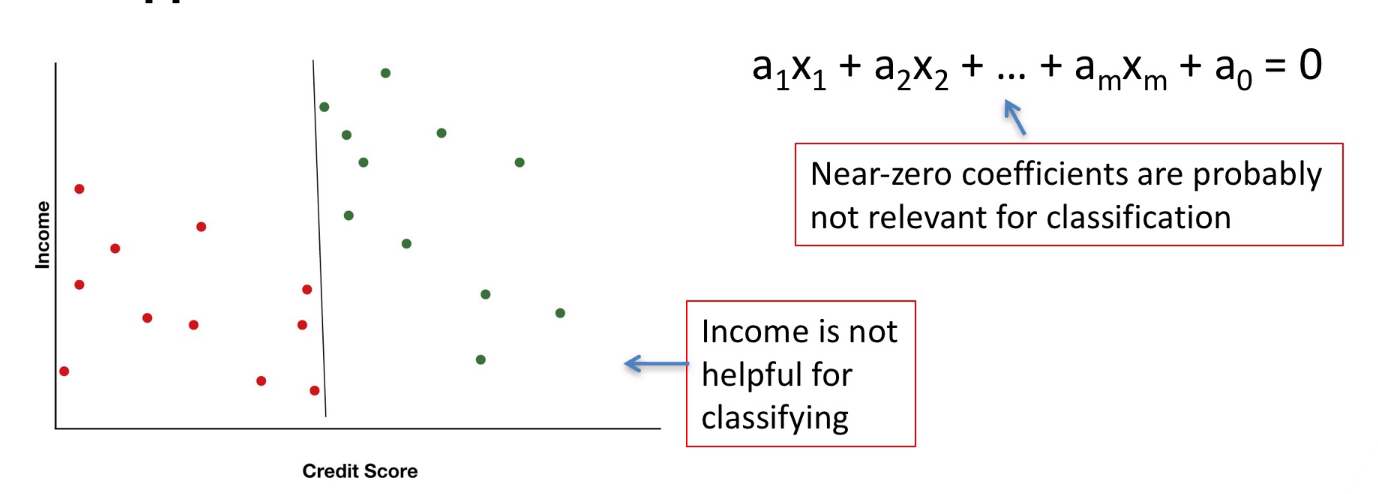
Since we are dealing with two attributes and squaring it, it will cause huge change. we need to set them up in the same scale.

**NEED TO SCALE DATA**

### Scenario#4:

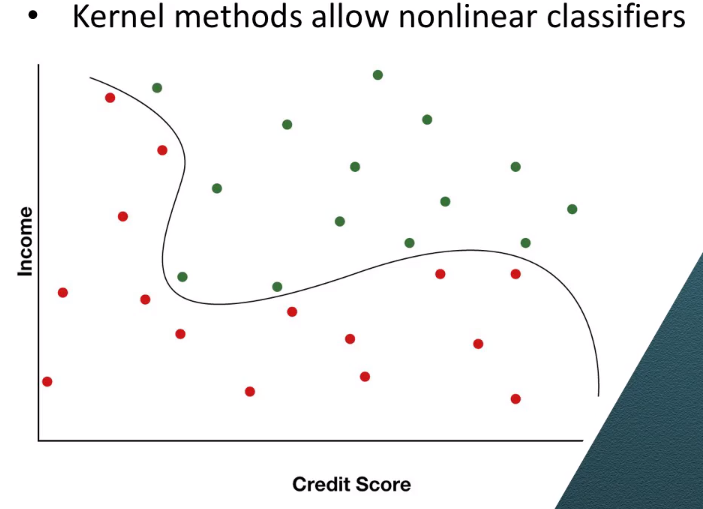
How to pick up attributes not needed for classification

Solution: for lot of attribtues.if co-efficients a1 to am is near zero, that is the indication that the coeffient corresponding to that attribute is not relevant to final result.



### Final Notes

1. SVM work the same for more than 2 dimension as well
2. SVM can be applied for non linear classification where the line is not a straight line . Usign **kernel (software like R has it)**



1. Can Some classification be answered as probability,

Say,instead of can we give loan or not; we can ask , what is the percetange that an applicant will pay back the loan. This is where “logistic regression” comes into play

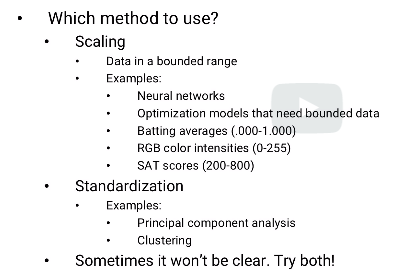
1. Other approach for classification. For More than 2 classes – yes –

## Scaling and standardization

Agenda: When to use scaling vs. stnaadardization?

### When to use what?

If some models have bounded range, then scaling is preferred..

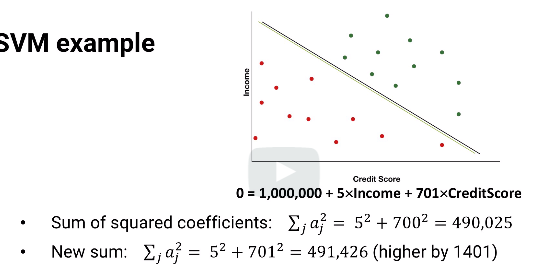


**ALWAYS SCALE THE DATA BEFORE generating model**

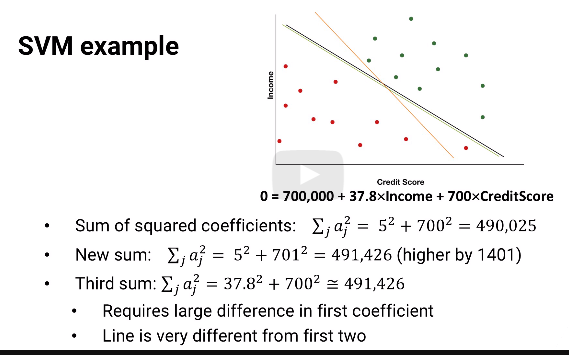
### Scaling

**Why should we do?**

**Initial co-efficients**

****

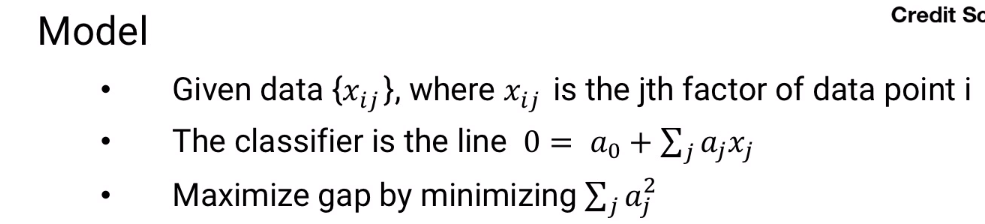
If credit score co-oefficient is changed from 700 to 701, we have to increase income from 5 to 37.8 (600% increase) which changes the classifier line



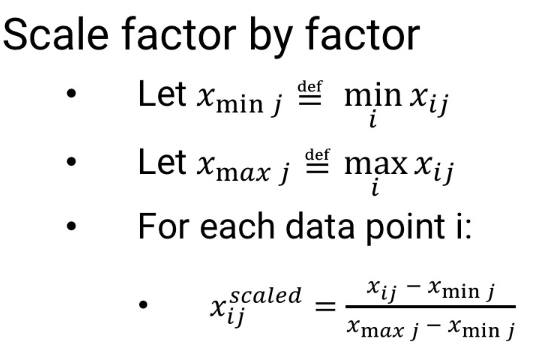
**How to scale:** Generally scaling demands we keep all attributes co-efficients between **0-1 as standard**

**Definition**

Original:



Scaled:

For each factor j, we set xminj and xmaxj to be the smallest and largest factor value and then for each point i, its new scaled factor value divided by xmaxj minus xminj.

Say, 4 values in attribute “j”

X1j=10

X1j=20

X1j=30

X1j=40

X(max) =40; x(min) =10 ;xmax-x(min)=40-10=30

**Scaling:**

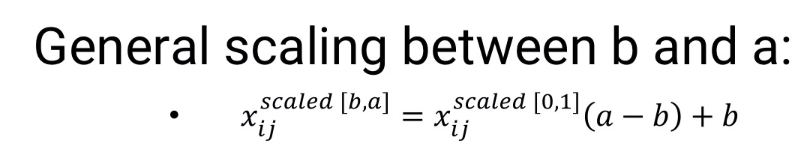
Xij(scaled)=(10-10)/30=0

X2j(scaled)=(20-10)/30=0.33

X3j(scaled)=(30-10)/30=0.66

X4j(scaled)=(40-10)/30=1

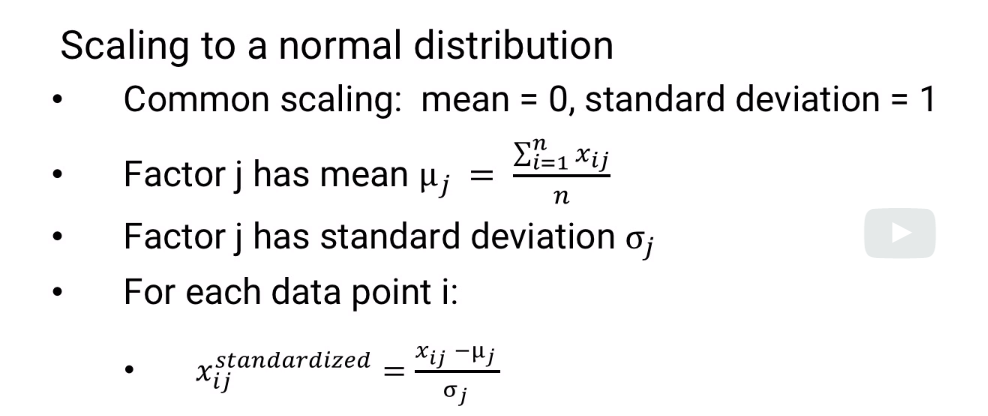
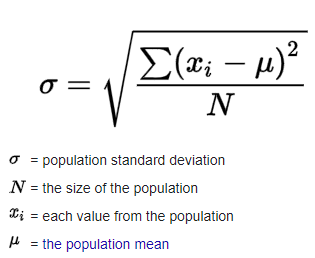
If scaling is not between 0 and 1, below is the formula for general scaling factor

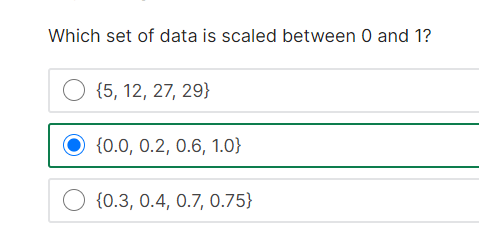


### Standardized:

Scale to a normal distribution

**Standard deviation formula:**

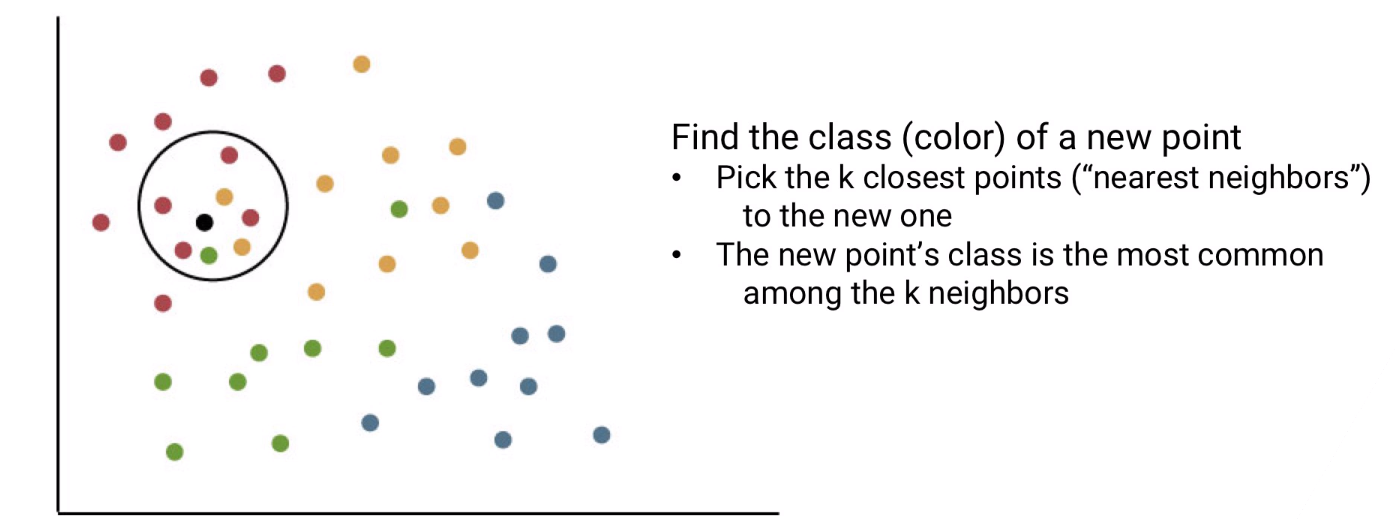
 this third option is between 1 and 1 but donot span the full range

## K-nearest neighbor classification (KNN) –simple model

-more than 2 classes are involved (SVM used when there is binary classification)

K: number of closet points to pick

**Logic:** pick the nearest neighbor. The neighbors that appear more are the ones that the new data point will be assigned to



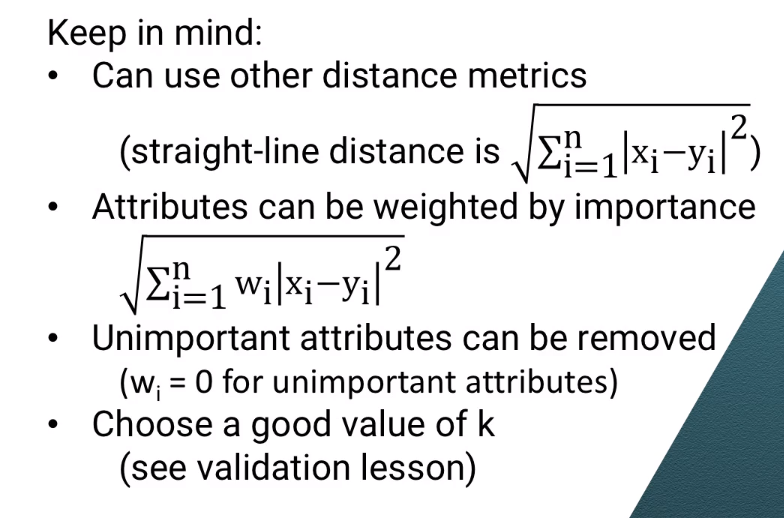
### Complexities to consider

1. **How to obtain the neighbor>** many way to getuse Distance: usually a straight line .Will see more in “distance metrics” topic
2. Some attributes weigh more than others. Larger the weight, greater the impact of the dimension’ distance.

Finding weights needs techniques from “regression”

1. Unimportant attributes/predictors can be removed. Will see more in “variable selection” topic
2. How to find right value of “k” : try and “validate” each “k” value and get to the right “k”

Distance is just pythagorus therem (distance between two points is square root of [(delta in x)^2 + delta in y)^2]



More notes (Self)

<https://www.analyticsvidhya.com/blog/2021/05/knn-the-distance-based-machine-learning-algorithm/>

\*K’ is the hyperparameter for KNN. For proper classification/prediction, the value of K must be fine-tuned.

### It is recommended to always select an odd value of K ~

* \***Larger K value:** The case of underfitting occurs when the value of k is increased. In this case, the model would be unable to correctly learn on the training data.
* **Smaller k value:**The condition of overfitting occurs when the value of k is smaller. The model will capture all of the training data, including noise. The model will perform poorly for the test data in this scenario.

# RESOURCES:

[http://www.statsoft.com/Textbook/k-Nearest-Neighbors](https://web.archive.org/web/20200121091131/http:/www.statsoft.com/Textbook/k-Nearest-Neighbors).

## Discussion forums

Please join the Piazza forums as follows:

Access code: edXVMMISyE6501FA21ESIW

Signup Link: [piazza.com/gatech/fall2021/isye6501verified](https://piazza.com/gatech/fall2021/isye6501verified)

Class Link: [piazza.com/gatech/fall2021/isye6501verified/home](https://piazza.com/gatech/fall2021/isye6501verified/home)

<https://commons.wikimedia.org/wiki/File:KnnClassification.svg>

# Homework

**Focus:**

The main focus of the homework will be your analysis of your results, **not your code**.  For a top score, you shouldn't just run some code and display some results; rather, you should also discuss the results qualitatively, point out anything surprising, and comment on possible explanations. The top recommended forms for your analysis, in order, are: pdf, html, txt, and then word doc.

We recommend **not** including your name / e-mail in your homework PDF.  We (TAs) and the system will know who you are, and submitting anonymously will not cause any problems assigning your grade later. It will also avoid potential problems of bias.

 100 is "All correct (perhaps except a few details) with a deeper solution than expected"

* This means that if students gave truly creative answers for optional portions or delves deeper into the analytical methods and modeling techniques in a manner not necessarily expected or gone over in this class, with appropriate solutions and explanations, then that constitutes a 100. All questions are completed. **When giving this grade, please comment what you believe the person did well and how they went above and beyond.**I cannot force you to leave a comment, but it will help stay consistent with the regrading policy that will be mentioned later.

 90 is for "most or all correct."

* This means that student provided code, answered the questions asked with code output, and provided reasonable explanations for their answers. There can be minor errors like slight inconsistencies or minor misunderstandings. All questions are answered. **This is the "default" grade. If a student does what is asked, and provides reasonable explanation for their work, but doesn't go above and beyond, this is their grade.** While comments here can be nice, if you have nothing else to add other than "they did everything right," you do not have to leave a comment.

 75 is "not correct, but a reasonable attempt."

* There is some code, solutions, and explanation, but the explanation is faulty or incorrect or the solution values are completely different than outlined in the homework solutions or solutions and explanation do not make sense. At least half of the questions are answered/attempted including the coding questions.  **When giving this grade, please comment what you believe the person did that was fundamentally incorrect.**This will help the student learn and will also help me in case of a regrade request.

 50 is "Not correct, insufficient effort."

* There is clearly no effort on homework like no code, no answers, no explanations, or little of any of these. Particularly, if a student only answers a descriptive problem (one which involves answering a prompt) and does none of the coding problems, then this would also be considered insufficient effort. **When giving this grade, please comment what you believe the person did that was fundamentally incorrect and explain what part of the homework they did not do.**This will help the student learn and will also help me in case of a regrade request.