AS K INCREASES THE NUMERATOR INCREASES OR DECREASES. AND THIS ALSO HAPPENS FOR THE DENOMINATOR.

THE BEST K (THEORETICALLY) IS THE ONE WITH THE HIGHEST CH VALUE

S' IS THE IS STARTING ARRAY TO CALCULATE THE CENTERS

CK IS THE ACTUAL CENTERS,

A ARRAY TELLS YOU WHAT GROUP EACH DATA GROUP WILL BE.

MACHINE LEARNING 9/16

- HOW TO DECI THE NUMBER OF K IN CLUSTERING

WITH CLUSTER VARIATION THERE IS A PROBLEM BECAUSE IT JUST KEEPS DECREASING.

BETWEEN CLUSTER VARIATION

$$B = \sum_{k=1}^{K} n_k \| \bar{x}_k - \bar{x} \|_2^2$$

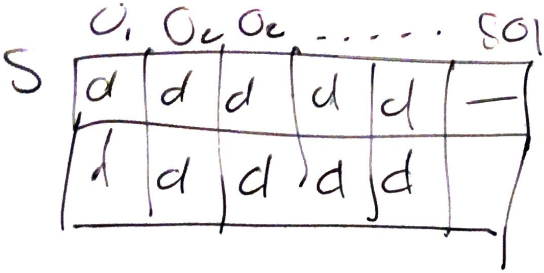
n = CAN BE WEIGHTED TO MATCH THE # OF SAMPLES

PROBLEM BECAUSE AS K KEEPS INCREASING SO DOES CLUSTER VARIATNN

$W(K)$ - WITHIN CLUSTER VARIATION

$B(K)$
BETWEEN CLUSTER VARCATIN

$$CH(K) = \frac{B(K)/(K-1)}{W(K)/(n-K)}$$



DOWNLOAD IS A WRITE OPERATION

OWNER CAN CHANGE THE PERMISSION OF A FILE. OWNER RIGHT TO CHANGE OF OWNER. ONLY ROOT CAN CHANGE THE RIGHT OF OWNERS.

MAKE SURE CONGURENCY COULD HAPPEN. RACE CONDITION .... ETC. MANAGER CALLS INCTR OR DECRTR AND THOSE IMPLIMENT THE FUNCTIONS.

ACCESS CONTROL MECHANISM

ACCESS CONTROL LIST

CAPABILITIES

RING BASED ACCESS CONTROL

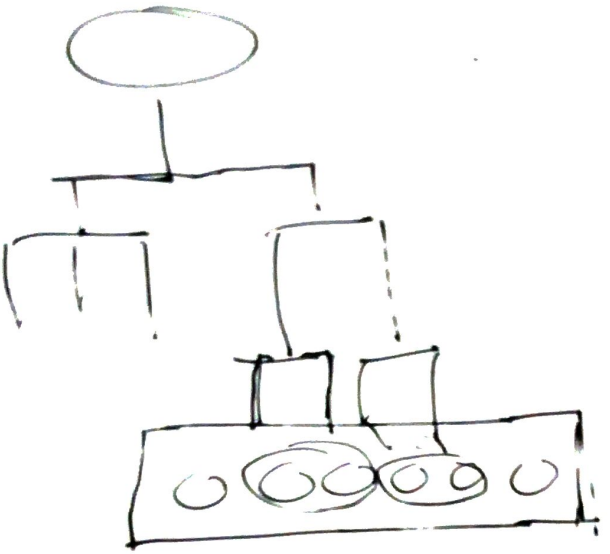
IMPLEMENTING

- DEFINE SECURITY POLICY
- CREATE ACCESS CONTROL MATRIX
- IMPLEMENT AND PUT ACM IN STORAGE
- DEFINE PROCEDURE TO

ACCESS CONTROL LIST

EACH FILE HOLDS ITS OWN PERMISSIONS.

HIERARCHICAL CLUSTERING 9/6



MERGE THE CLOSEST TWO POINTS TOGETHER THEN GROUPS = M-1

REDEAT IT AND YOU WILL HAVE M-2 GROUPS

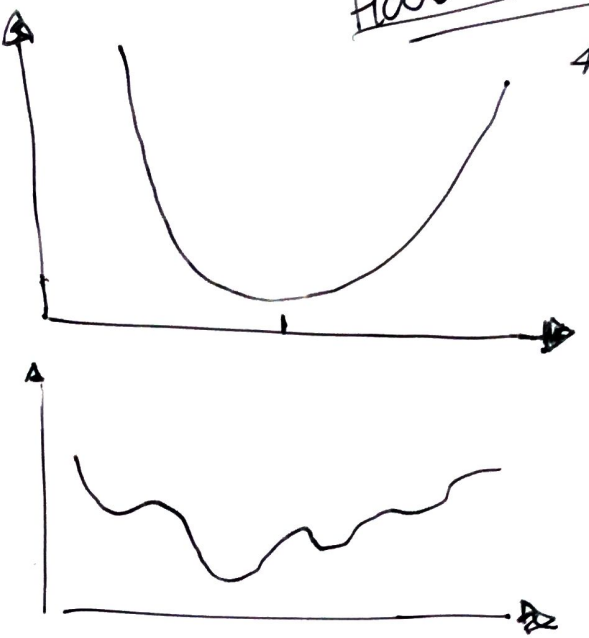
CONTINUE DOING THIS UNTIL YOU HAVE ON GROUP.

GIVE K AND WE WILL "CHOP" THE RESULTS GIVING US THE # OF GROUPS AND DATA POINTS IN THEM.

CONVEX

HOW TO CHOOSE K

ELBOW



- FINDING THE LOCAL MINIMUM MIGHT BE RETURNED BY THE K MEANS ALGORITHM
- TO AVOID THIS RUN THE ALGORITHM MULTIPLE TIMES.

K MEANS IS SENSATIVE TO OUTLIERS.

K MEANS VS K-MEDOIDS

K MEDOIDS GETS NEW CENTER AS A POINT.

2) REPEAT UNTIL CONVERGENCE

{ FOR EVERY i, SET

$$C := \text{argmin}_j ||x^i - M_j||^2$$

x_i	x_2 - - - x_n	C (CLUSTER)
n {		

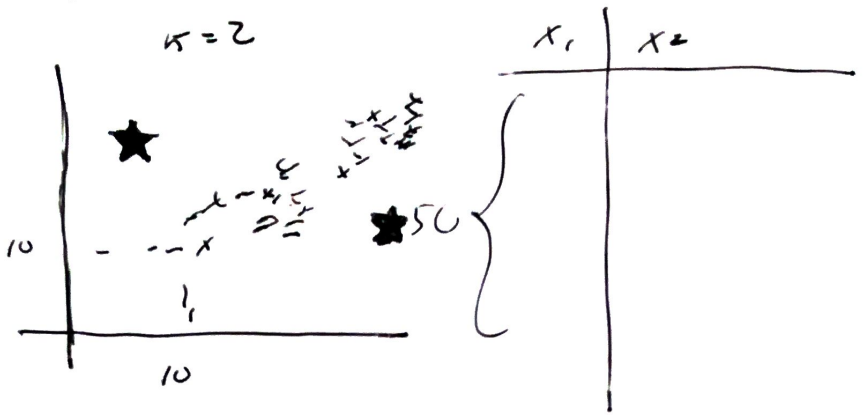
$\text{argmin}\,(x-2)^2$ WHAT VALUE OF x WILL MAKE

3) FOR EVERY j, SET

$$M_j := \frac{\sum_{i=1}^{m} 1.\{c^{(i)} = j\}\, x^{(i)}}{\sum_{i=1}^{m} 1.\{c^{(i)} = j\}}$$

MACHINE LEARNING

FOR THE K ALGORITHM YOU MUST STATE THE NUMBER OF CLUSTERS YOU WANT TO GROUP THE DATA IN.



- EACH DATA POINT WILL CALCULATE THE DISTANCE TO EACH RANDOM POINT.
- THE CLOSEST RANDOM POINT WILL BELONG TO THAT GROUP.
- THE NEW CENTER WILL BE CALCULATED FOR EACH GROUP
- THEN THE PROCESS STARTS AGAIN AND CONTINUES UNTIL THE CENTERS ARE NO LONGER CHANGED.

K-MEANS:

INPUT $\{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\}$

① INITIALIZE (CLUSTER CENTROIDS)

$M_1, M_2 \ldots M_K$ ERR

DISTANCE MATRIX

	(1,1) C_1	(1,2) C_2	GROUP
(1,1)	0	1	C_1
(2,1)	1	0	C_2
(4,3)	3.61	2.81	C_2
(5,4)	5	4.24	C_2

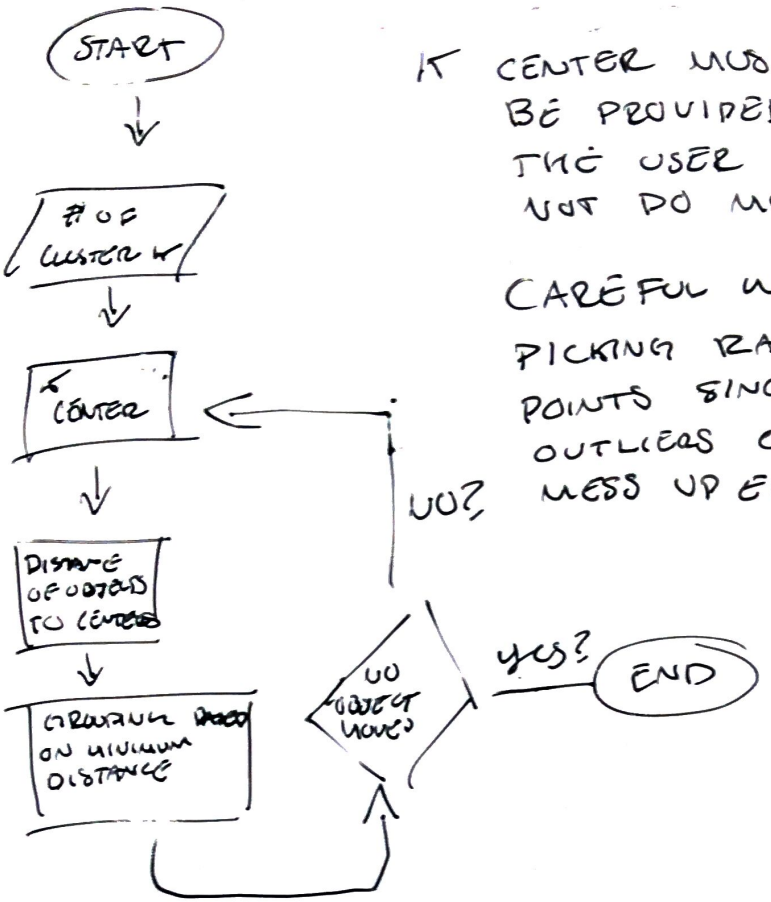
NEW CENTER IS AVERAGE OF EACH FEATURE DEVIDED BY # OF POINTS

$$C_2 = \left(\frac{2+4+5}{3}, \frac{1+3+4}{3}\right)$$

$$C_1 = \left(\frac{1}{1}, \frac{1}{1}\right) = (1, 1)$$

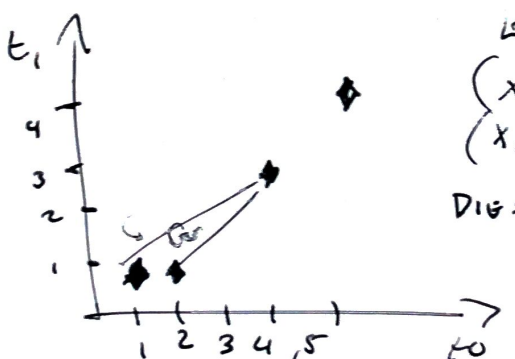
THE PROCESS REPEATS WITH THE NEW CENTERS.
UNTIL THE CENTERS DO NOT CHANGE.

THIS METHOD MAY NOT WORK WITH DATA THAT HAS MULTIPLE DIMENSIONS.



IT CENTER MUST BE PROVIDED BY THE USER AND NOT DO MUCH.

CAREFUL WHEN PICKING RANDOM POINTS SINCE OUTLIERS CAN MESS UP EVERYTHING



$K = 2$

(x_1, y_1)

(x_1, y_2)

$DIS = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$

t_0	t_1
1	1
2	1
4	3
5	4

APPLICATIONS

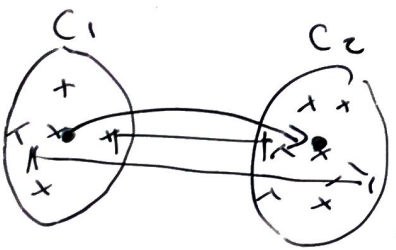
- PATTER RECOGNITION

WHAT IS GOOD CLUSTERING?

- HIGH QUALITY CLUSTERING
  - HIGH INTER CLASS SIMILARITY
  - LOW BETWEEN CLASS SIMILARITY

- QUALITY IS A RESULT ON SIMILARITY MEASURE USED AND ITS IMPLEMENTATION.
- ALSO ABILITY TO DISCOVER SOME OR ALL PATTERNS.

CALCULATE THE CENTER OF THE CLUSTERING.



DISTANCE BETWEEN CLUSTERS

- CENTER OF BOTH
- THE TWO CLOSEST
- THE TWO FURTHEST POINTS

PARTITIONING METHOD

K-MEANS ALGORITHM.

K - # OF CLUSTERS OR GROUPS

MACHINE LEARNING 9/4

TRAINING DATA

YEAR	MILAGE	. . . .
1998	200,000	

m

n

SUPERVISED
- PROVIDE THE ANSWER

UNSUPERVISED
- NO ANSWER IS PROVIDED

m - # OF TRAINING SAMPLES

n - DIMENSION OF DATA SAMPLES. (NO ANSWER)

X - TRAINING DATA SET

$X^{(i)}$ - TRAINING SET FOR A SAMPLE

Y - LABEL / ANNOTATION

$X \longrightarrow Y$ RELATIONSHIP (SUPERVISED)

X_j - j^{th} FEATURE

WHAT IS CLUSTERING?
- SIMILAR DATA GROUPED TOGETHER.

X_j^i - CONSTANTLY USED TOGETHER TO PINPOINT A CERTAIN PIECE OF DATA

FIND THE BEST FIT LINE TO FIND EQUATION AND USE THAT TO PREDICT SELLING PRICE.

"REGRESSION PROBLEM" USED WHEN PREDICTED VALUE IS CONTINUES

CAN IT USE WITH DESCRETE VALUES

CLASSIFICATION PROBLEM

CANCER - PROVIDE DATA WITH RIGHT ANSWER AND THEN USE THAT TO PREDICT WHAT KIND OF CANER.

+ BENIGN — MALIGMENT

DESCRETE VALUES

UNSUPERVISED LEARNING

- DATA WITH NO LABEL BUT THE ALGORITHM WILL CLUSTER TOGETHER SO IT MAKES SENSE.
- NO ANSWER IS GIVEN

ROLE OF STATISTICS: INFERENCE FROM A SAMPLE

ROLE OF COMPUTER SCIENCE: EFFICIENT ALGORITHMS TO: MAKE INFERENCES.

SOME TYPES

LEARNING ASSOCIATIONS: RELATIONSHIP IN THE DATA

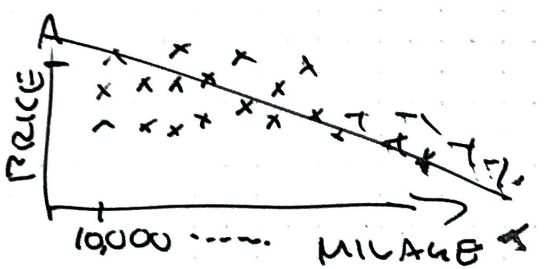
SUPERVISED: GIVE EXAMPLES WITH LABELS

UNSUPERVISED: GIVE DATA AND THE MACHINE FINDS PATTERS.

REINFORCEMENT LEARNING: LEAR A POLICY THAT REWARDS FOR BEING CORRECT.

SUPERVISED

- SELL CAR: COLLECT SELLING DATA FROM PAST (RIGHT ANSWER MUST BE PROVIDED)



CS 4347: MACHINE LEARNING

MACHINE LEARNING WHEN TO USE IT?

CAN USE TO PREDICT STOCKS TAKING INTO ACCOUNT PAST STATES.

WHY LEARN?

- PROGRAM TO OPTIMIZE A PERFORMANCE CRITERION USING DATA OR PAST EXPERIENCE
- DO NOT NEED TO LEARN TO CALCULATE PAYROLL
- LEARNING USED
  - HUMAN EXPERTISE DOES NOT EXIST
  - HUMANS ARE UNABLE TO EXPLAIN THEIR EXPERTISE
  - SOLUTION CHANGES IN TIME
  - SOLUTION NEEDS TO ADAPT TO PARTICULAR CASES. (FACE RECOGNITION)

WHAT WE TALK ABOUT WHEN SAYING LEARNING

- DATA IS CHEAP + ABUNDANT
  KNOWLEDGE IS EXPENSIVE + SCARCE
- LEARNING GENERAL MODELS FROM A DATA OF PARTICULAR EXAMPLES.
- BUILD MODEL THAT IS GOOD AND USEFUL APPROXIMATION TO THE DATA.

DATA MINING – APPLICATION OF MACHINE LEARNING

MACHINE LEARNING AUG, 28

DEEP LEARNING - IAN GOODFELLOW
PATTERN RECOGNITION AND MACHINE LEARNING

MACHINE LEARNING - ARTHUR SAMUEL (1959)
FIELD OF STUDY THAT GIVES COMPUTERS THAT ABILITY TO LEARN WITHOUT BEING EXPLICITLY PROGRAMED.

TOM MITCHELL (1998) WELL-POSED LEARNING PROBLEM: A COMPUTER PROGRAM IS SAID TO LEARN FROM EXPERIENCE E WITH RESPECT TO SOME TASK T AND SOME PERFORMANCE MEASURE P, IF IT IS PERFORMANCE ON T AS MEASURED BY P, IMPROVE WITH EXPERIENCE E.

SUPERVISED LEARNING

- MOST POPULAR
- LOTS OF DATA WITH THE ANSWER.

UNSUPERVISED LEARNING

- DATA BUT NO RIGHT ANSWER IS PROVIDED