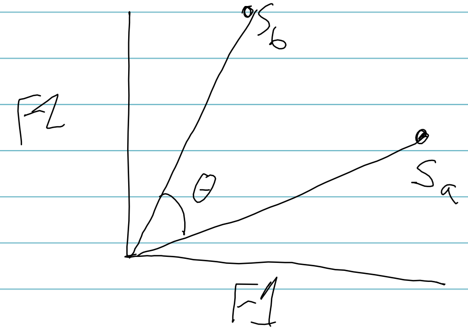
18 SEP 2015 SEM. WEB MINING

REAL VAL MEAS TO DIS. FEATURE

REAL MEAS	FEAT
$m(X)$	YES IF $m(X) > 50$ NO OW YES IF $m(X) < 20$ NO OW
$m(x)$	$m(X) \in [0,10]$ $m(X) \in [10,50]$ $\vdots$

COSINE SIM.



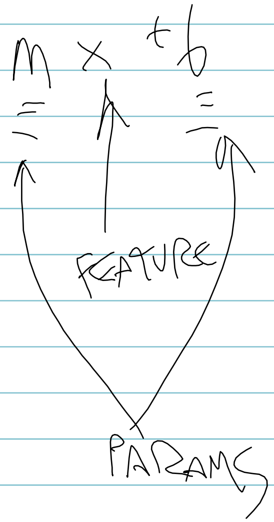
$\cos\theta$ IS "SIMILARITY"
MEAS BETWN S_a & S_b

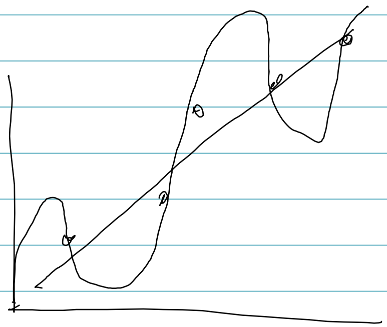
UNSUPERVISED

DESCRIBING A DATASET

SUPERVISED

TRAINING DATA — GIVES US AN
INTUITION
ON FUTURE DATA


mx + b
FEATURE
PARAMS



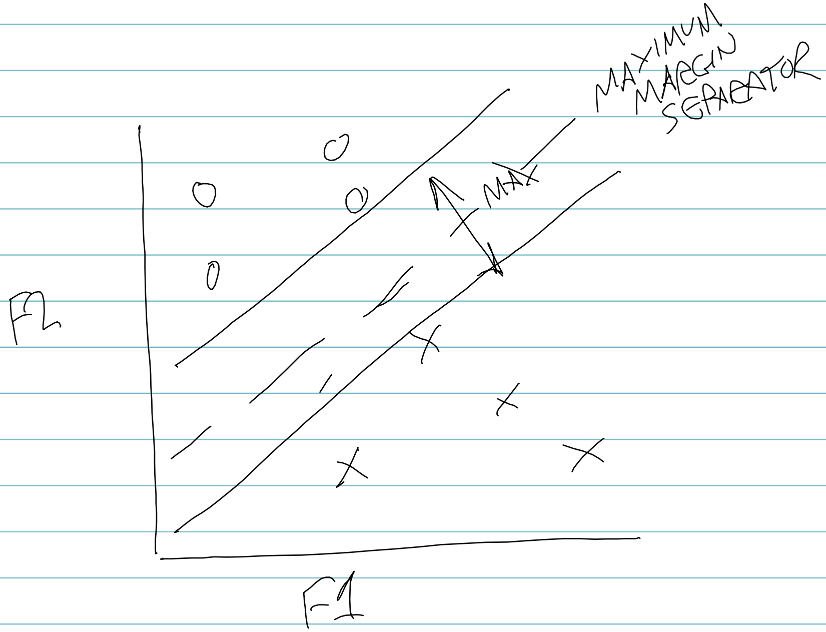
PARAMETRIC

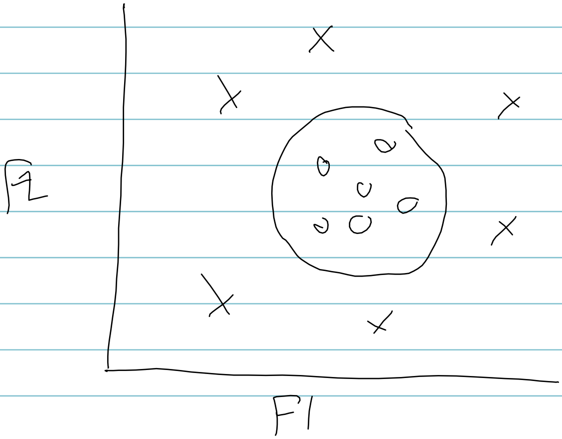
NUMBER OF PARAMETERS IS FIXED (NOT RELATED TO x)

NON-PARAMETRIC

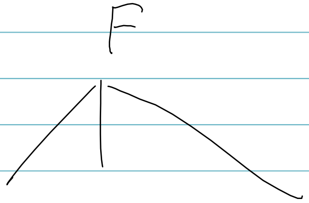
NUMBER OF PARAMETERS IS REL'D TO TNG DATA

SVM



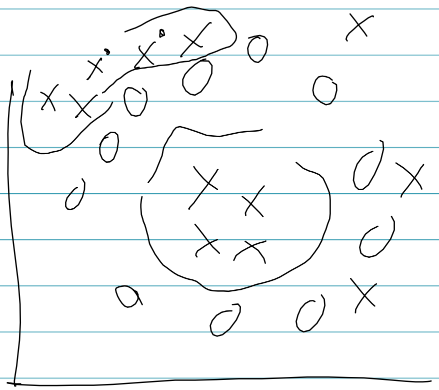


DECISION STUMP



PICKE BASED ON IG FOR A SAMPLE OF THE TNG DATA.

DATA COMPLEXITY / IMBALANCE



$$P = \frac{\text{CORR CL. SAMPLES}}{\text{SAMPLE CLASS BY LEARNER}}$$

$$R = \frac{\text{CORR CL SAMPLES}}{\text{ALL SAMPLES OF CLASS}}$$

DATA SET D

LEAVE-ONE-OUT CROSS VAL.

n SAMPLES IN D

EACH TEST IS DONE LEARNED

W. $D - \{x\}$

AND TESTED ON x

10-FOLD CROSS VAL

90% TNG

10% TEST

REPEAT 10x

REPORT AVG METRICS (PREC, RECALL, AUC) FOR EA. TRIAL

VALIDATION EXPERIMENT

NEW DATA THAT WASNT COLL W. INIT DATASET.