LET'S THINK ABOUT SHAPES FOR A BIT

WHAT REALLY IS A SHAPE? ITS A SET OF POINTS

"A LINE IS A UNI-DIMENSIONAL SHAPE"

ANY POINT ON A LINE CAN BE SPECIFIED WITH 1 NUMBER

(THAT 1 NUMBER IS THE DISTANCE FROM THE ORIGIN)

"A SQUARE IS A TWO-DIMENSIONAL SHAPE"

ANY POINT ON A SQUARE CAN BE SPECIFIED WITH 2 NUMBERS

(THOSE 2 NUMBERS ARE THE X AND Y COORDINATES)

"A CUBE IS A THREE-DIMENSIONAL SHAPE"

ANY POINT ON A CUBE CAN BE SPECIFIED WITH 3 NUMBERS

(THOSE 3 NUMBERS ARE THE X, Y AND Z COORDINATES)

WHAT DOES THIS MEAN?

SO - A POINT IN AN N-DIMENSIONAL SPACE NEEDS N COORDINATES TO BE REPRESENTED

A 3-DIMENSIONAL SPACE IS REPRESENTED BY A CUBE

A 2-DIMENSIONAL SPACE IS REPRESENTED BY A RECTANGLE

(WE MAY FIND IT HARD TO VISUALIZE 4 OR MORE COORDINATE SPACES, BUT THERE IS NO MAGIC ABOUT THEM - JUST THINK OF THE EACH POINT AS A TUPLE OF 'N' NUMBERS)

AN N-DIMENSIONAL SPACE IS REPRESENTED BY

AN N-DIMENSIONAL HYPERCUBE

THIS IS A FANCY WORD,
BUT DON'T BE INTIMIDATED -
IT JUST MEANS EACH POINT
IN THIS HYPERCUBE
IS A LIST OF N VALUE

OKEE - NOW WITH THIS IN MIND,
LET'S GET BACK TO OUR SPAM
CLASSIFICATION EXAMPLE

ANY EMAIL MESSAGE CAN BE REPRESENTED AS A POINT IN A HYPERCUBE

HOW?

REMEMBER WE ALREADY MENTIONED
THAT A PROBLEM INSTANCE CONSISTS
OF A FEATURE VECTOR

OUR PROBLEM INSTANCE WAS: AN EMAIL

OUR FEATURE VECTOR WAS: THE WORDS
IN THE EMAIL

LET'S SAY WE HAVE A LIST THAT REPRESENTS THE ENTIRE
UNIVERSE OF WORDS THAT CAN APPEAR IN AN EMAIL

$(W_1, W_2, \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots W_N)$

(hello, this, is, the, universe, of, all, words, in, emails, a, an, test, how, are, you, goodbye)

ANY MESSAGE WOULD ONLY CONTAIN A SUBSET OF THE WORDS IN
THE ABOVE LIST

EACH OF THESE MESSAGES CAN BE
REPRESENTED AS A TUPLE OF 1'S AND 0'S

MESSAGE 1: HELLO, THIS IS A TEST

(hello, this, is, the, universe, of, all, words, in, emails, a, an, test, how, are, you, goodbye)

(1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0)

MESSAGE 2: HOW ARE YOU

(hello, this, is, the, universe, of, all, words, in, emails, a, an, test, how, are, you, goodbye)

(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0)

MESSAGE 3: HELLO ALL! GOODBYE

(hello, this, is, the, universe, of, all, words, in, emails, a, an, test, how, are, you, goodbye)

(1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1)

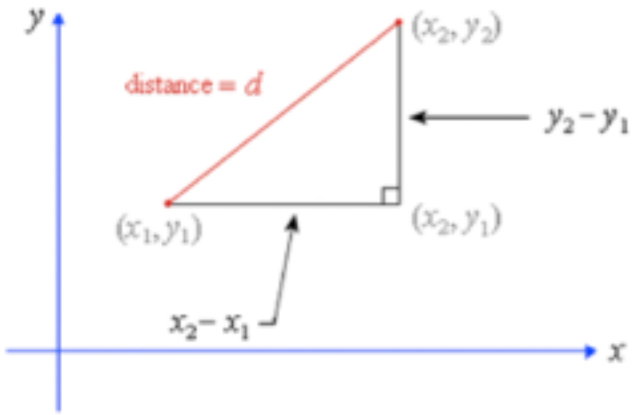
THESE TUPLES
ARE POINTS IN AN
N-DIMENSIONAL
HYPERCUBE

OK, NOW THAT WE HAVE REPRESENTED EACH EMAIL AS A POINT IN A HYPERCUBE - WHAT NEXT?

(THE PROBLEM INSTANCE)

DO THIS FOR THE EMAIL WE WANT TO CLASSIFY AS WELL AS FOR ALL THE EMAILS WE ALREADY HAVE INFORMATION ABOUT

THE TRAINING DATA

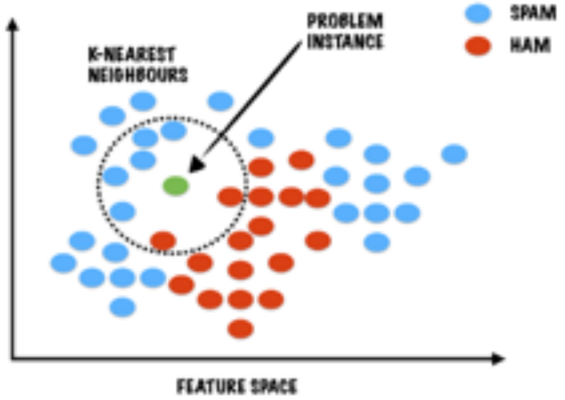


NOW THESE ARE ALL POINTS IN SPACE, SO WE CAN FIND THE DISTANCE BETWEEN THEM

(BUNCH OF WAYS TO CALCULATE DISTANCE BETWEEN 2 POINTS - INCLUDING THE SUPER-SIMPLE EUCLIDEAN DISTANCE FORMULA - MORE ON THIS IN A MINUTE)

FIND THE K NEAREST NEIGHBOURS OF OUR PROBLEM INSTANCE EMAIL

IF MORE THAN SOME THRESHOLD OF THEM ARE SPAM, THEN MARK THIS EMAIL AS SPAM TOO



FIND THE K NEAREST NEIGHBOURS OF OUR PROBLEM INSTANCE EMAIL

IF MORE THAN SOME THRESHOLD OF THEM ARE SPAM, THEN MARK THIS EMAIL AS SPAM TOO

THE SETUP ABOVE IS SLIGHTLY SIMPLER THAN REALITY, BUT NOT BY ALL THAT MUCH

THE FEATURE VECTOR WILL ALMOST NEVER CONTAIN A TUPLE OF 0,1 NUMBERS LIKE THAT DESCRIBED ABOVE, INSTEAD A SMART ALGORITHM THAT "HASHES" SUBSETS OF THE MAIL

THE DISTANCE WILL LIKELY NOT BE STRAIGHT EUCLIDEAN DISTANCE, BUT RATHER SOME MORE SOPHISTICATED FORMULA THAT IS FOUND TO WORK WELL IN TRAINING

ALSO, K-NEAREST NEIGHBOUR MAKES NO ASSUMPTION AT ALL ABOUT THE PROBABILITY DISTRIBUTIONS OF THE FEATURE VECTORS

FOR THIS REASON, THIS IS SAID TO BE A **NON-PARAMETRIC CLASSIFIER**

IN CONTRAST, THE NAIVE BAYES CLASSIFIER DOES INDEED MAKE ASSUMPTIONS ABOUT THE PROBABILITY DISTRIBUTION -

(ALTHOUGH WE DID NOT TOUCH UPON THOSE ASSUMPTIONS IN OUR OVERVIEW ABOVE)

SO, NAIVE BAYES IS SAID TO BE A **PROBABILISTIC CLASSIFIER**

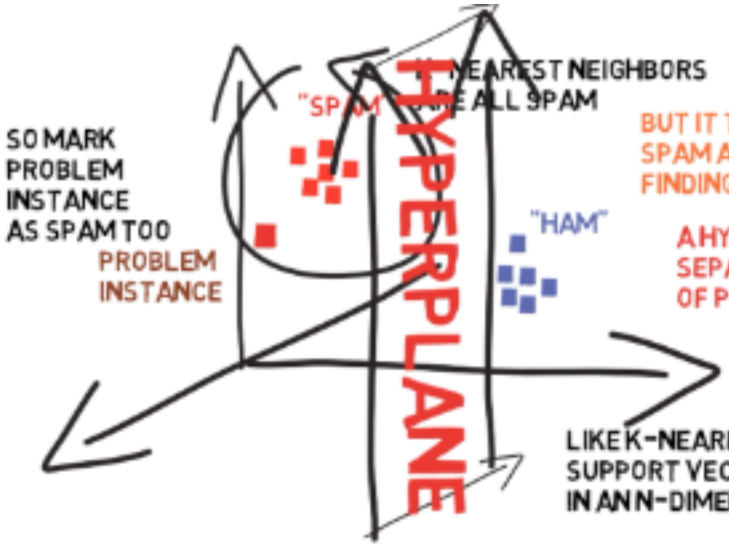
LET US, REALLY QUICKLY, ALSO SEE HOW YET ANOTHER MACHINE-LEARNING TECHNIQUE COULD BE APPLIED TO SPAM DETECTION

SUPPORT VECTOR MACHINES

ARE SUPERVISED-LEARNING MODELS THAT ARE USED TO BUILD NON-PROBABILISTIC CLASSIFIERS

THIS METHOD ONLY REALLY WORKS IF THERE ARE 2 LABELS ("SPAM" AND "HAM")

THUS SVMS ARE USED TO BUILD BINARY CLASSIFIERS



K NEAREST NEIGHBORS ARE ALL SPAM

SO MARK PROBLEM INSTANCE AS SPAM TOO

BUT IT TRIES TO SEPARATE THE SPAM AND HAM POINTS BY FINDING

A HYPERPLANE THAT NICELY SEPARATES THE TWO SETS OF POINTS

LIKE K-NEAREST NEIGHBOR MODELS, SUPPORT VECTOR MODELS ALSO WORK IN AN N-DIMENSIONAL HYPERCUBE

THIS HYPERPLANE THEN ACTS AS A BOUNDARY BETWEEN THE 2