DIMENSIONALITY REDUCTION

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

WHAT DOES THIS MEAN?

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

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

(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)

A 3-DIMENSIONAL SPACE IS REPRESENTED BY A CUBE

A 2-DIMENSIONAL SPACE IS REPRESENTED BY A RECTANGLE

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 VALUES

ANYTHING - ANYTHING - CAN BE REPRESENTED AS A POINT IN A HYPERCUBE

LET'S SEE HOW WE COULD REPRESENT AN EMAIL IN AN N-DIMENSIONAL HYPERCUBE

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

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

(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

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

TAKE THE SET OF ALL WORDS
THAT CAN POSSIBLY APPEAR IN ANY EMAIL

W1, W2,...Wn

ANY EMAIL WILL CONTAIN SOME
SUBSET OF THESE WORDS

M1, M2,..Mi

REPRESENT EACH MESSAGE AS A TUPLE

(X1,X2,X3...Xn)

WHERE A PARTICULAR ELEMENT XJ
IS 1 IF WORD J APPEARS IN THE EMAIL,
ELSE IS 0

BASICALLY - ANY INSTANCE CAN BE
REPRESENTED USING A FEATURE VECTOR
- A LIST OF NUMBERS THAT DESCRIBE THAT
INSTANCE

WE CAN THEN DO ALL KINDS OF COOL
THINGS TO THAT INSTANCE - FINDING
OTHER VECTORS THAT ARE "LIKE" THIS
ONE, FINDING ITS DISTANCE FROM
OTHER INSTANCES, AND SO ON

NOW, THE THING IS, THAT AS AN INSTANCE
GETS COMPLICATED, ITS FEATURE VECTOR
STARTS TO GET REALLY LONG

IN THE EXAMPLE ABOVE, WE
WERE SUGGESTING A FEATURE VECTOR
FOR AN EMAIL WHERE EVERY POSSIBLE
WORD WAS REPRESENTED WITH A 1 OR A 0

THIS FEATURE VECTOR WOULD BE
INFINITELY LONG, AND IMPOSSIBLE TO
DO ANYTHING WITH!

THIS GETS TO THE HEART OF
SOMETHING CALLED

THE CURSE OF DIMENSIONALITY

THE CURSE OF DIMENSIONALITY

ON THE ONE HAND

ANY RICH REPRESENTATION OF A COMPLEX INSTANCE REQUIRES A LOT OF FEATURES

ON THE OTHER HAND

WE ARE NOT SET UP TO EITHER VISUALIZE OR EFFICIENTLY PROCESS DATA OF VERY HIGH DIMENSIONALITY

THE SOLUTION?

DIMENSIONALITY REDUCTION TECHNIQUES

WHICH EFFECTIVELY REDUCE THE NUMBER OF DIMENSIONS THAT WE NEED TO EXPRESS OUR DATA IN

DIMENSIONALITY REDUCTION TECHNIQUES

WHICH EFFECTIVELY REDUCE THE NUMBER OF DIMENSIONS THAT WE NEED TO EXPRESS OUR DATA IN

SELECT FROM EXISTING FEATURES

FEATURE SELECTION TECHNIQUES

THESE ATTEMPT TO FIND A SUBSET OF FEATURES THAT RETAIN ALL THE NECCESSARY INFORMATION BUT LOSE THE JUNK

A SIMPLE EXAMPLE IS STEPWISE REGRESSION

GIVEN MANY TIME SERIES IN A REGRESSION, INCLUDE ONLY THOSE THAT, IN SOME SENSE, ADD INFORMATION TO THE REGRESSION

CREATE NEW FEATURES

FEATURE EXTRACTION TECHNIQUES

THESE SEEK TO RE-EXPRESS THE DATA IN A LOWER DIMENSIONALITY FORM

THE MOST FAMOUS FEATURE EXTRACTION TECHNIQUE IS PRINCIPAL COMPONENTS ANALYSIS (PCA)

GIVEN A LARGE NUMBER OF CORRELATED TIME SERIES, PCA WILL FIND 2-3 UNDERLYING CAUSES THAT EXPLAIN MOST OF THE MOVEMENTS