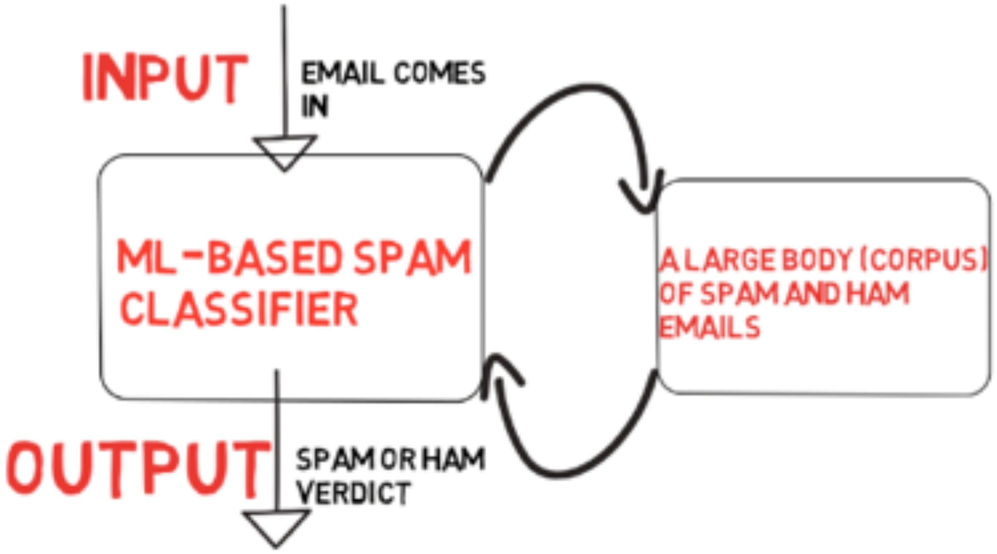
PLUNGING IN – MACHINE LEARNING APPROACHES TO SPAM DETECTION

THE ML-BASED APPROACH



INPUT

EMAIL COMES IN

ML-BASED SPAM CLASSIFIER

A LARGE BODY (CORPUS) OF SPAM AND HAM EMAILS

OUTPUT

SPAM OR HAM VERDICT

AS USUAL WITH AN ML-BASED TECHNIQUE, WE HAVE A LARGE CORPUS OF SPAM AND HAM EMAILS

FROM THIS, CALCULATE FOR EACH WORD IN EACH EMAIL, THE NUMBER OF TIMES THAT WORD APPEARS IN SPAM AND HAM EMAILS

FOR EACH WORD, CALCULATE A SPAMMINESS MEASURE FOR EACH WORD

$$\mathbf{S}[T] = \frac{C_{spam}(T)}{C_{spam}(T) + C_{ham}(T)}$$

HOW SPAMMY IS THE WORD T?

HOW MANY SPAM MESSAGES CONTAIN THE WORD T?

HOW MANY NON-SPAM MESSAGES CONTAIN THE

NOW LET'S SAY A NEW MESSAGE M COMES IN, CONSISTING OF WORDS T1, T2,...TN

1. LOOK UP THE SPAMMINESS OF EACH WORD T1, T2,...TN

2. FIND THE TOTAL SPAMMINESS OF THE MESSAGE, SAY BY MULTIPLYING THE SPAMMINESS OF EACH WORD

3. THEN FIND THE TOTAL HAMMINESS OF EACH WORD, BY MULTIPLYING (1-SPAMMINESS) OF EACH WORD

CALL THIS H[M]

M]

M], THEN

GE M IS SPAM.

THIS EXAMPLE IS A REAL ONE - THERE ARE LOADS OF SPAM DETECTORS THAT USE THIS BASIC

AS USUAL WITH AN ML-BASED TECHNIQUE, WE HAVE A LARGE CORPUS OF SPAM AND HAM EMAILS

FROM THIS, CALCULATE FOR EACH WORD IN EACH EMAIL, THE NUMBER OF TIMES THAT WORD APPEARS IN SPAM AND HAM EMAILS

FOR EACH WORD, CALCULATE A SPAMMINESS MEASURE FOR EACH WORD

$$\mathbf{S}[T] = \frac{C_{spam}(T)}{C_{spam}(T) + C_{ham}(T)}$$

HOW SPAMMY IS THE WORD T?

HOW MANY SPAM MESSAGES CONTAIN THE WORD T?

HOW MANY NON-SPAM MESSAGES CONTAIN THE WORD T?

NOW LET'S SAY A NEW MESSAGE M COMES IN, CONSISTING OF WORDS T1, T2,...TN

1. LOOK UP THE SPAMMINESS OF EACH WORD T1, T2,...TN

2. FIND THE TOTAL SPAMMINESS OF THE MESSAGE, SAY BY MULTIPLYING THE SPAMMINESS OF EACH WORD

CALL THIS S[M]

3. THEN FIND THE TOTAL HAMMINESS OF EACH WORD, BY MULTIPLYING (1-SPAMMINESS) OF EACH WORD

CALL THIS H[M]

IF S[M] > H[M], THEN THE MESSAGE M IS SPAM, ELSE ITS HAM

THIS EXAMPLE IS A REAL ONE - THERE ARE LOADS OF SPAM DETECTORS THAT USE THIS BASIC IDEA

NOTICE HOW THE DETECTOR HAD 2 DISTINCT PHASES

FIRST THE DETECTOR DID A BUNCH OF STUFF WITH THE PRE-EXISTING CORPUS OF SPAM AND HAM MAILS

"TRAINING THE MODEL"

THEN IT STARTED TO ACTUALLY ACCEPT REAL EMAILS AND MAKE SPAM/HAM DECISIONS

"RUNNING THE MODEL"

MACHINE LEARNING TECHNIQUES THAT EXPLICITLY HAVE A "TRAINING THE MODEL" STAGE ARE EXAMPLES OF **SUPERVISED LEARNING**

ALSO THE PROBLEM OF HAVING TO DECIDE HOW SOME ENTITY SHOULD BE CLASSIFIED IS A CLASSIC USE-CASE OF MACHINE-LEARNING

CLASSIFICATION PROBLEMS

THE METHOD WE JUST SAW IS SOMETHING KNOWN AS

A NAIVE BAYES CLASSIFIER

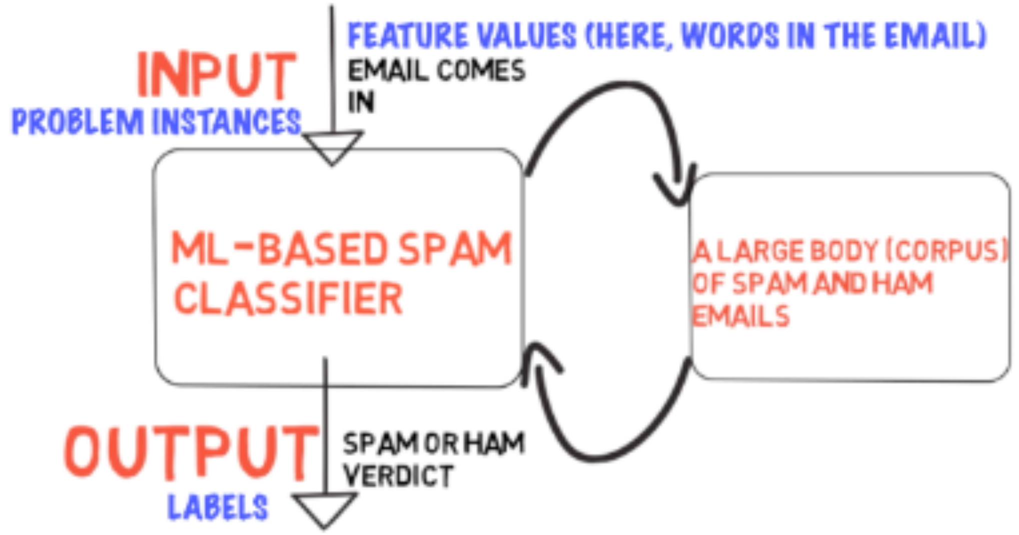
THE ENTITIES THAT WE ARE SEEKING TO CLASSIFY ARE CALLED **PROBLEM INSTANCES**

(IN OUR EXAMPLE, EMAILS ARE PROBLEM INSTANCES)

EACH PROBLEM INSTANCE IS A **VECTOR OF FEATURE VALUES**

(VECTOR LOOSELY MEANS LIST, OR TUPLE)

THE ML-BASED APPROACH



OF HAVING TO DECIDE HOW SOME ENTITY SHOULD BE CLASSIFIED IS A CLASSIC USE-CASE OF MACHINE-LEARNING

CLASSIFICATION PROBLEMS

THE METHOD WE JUST SAW IS SOMETHING KNOWN AS

A NAIVE BAYES CLASSIFIER

THE ENTITIES THAT WE ARE SEEKING TO CLASSIFY ARE CALLED PROBLEM INSTANCES

(IN OUR EXAMPLE, EMAILS ARE PROBLEM INSTANCES)

DON'T BE FOOLED BY THE NAME, NAIVE BAYES CLASSIFICATION IS AN EXTREMELY POWERFUL TECHNIQUE

THE "NAIVE" IN THE NAME IS SIMPLY BECAUSE THIS METHOD ASSUMES THAT FEATURE VALUES ARE INDEPENDENT OF EACH OTHER -

EACH PROBLEM INSTANCE IS A VECTOR OF FEATURE VALUES

(VECTOR LOOSELY MEANS LIST, OR TUPLE)

FEATURE VALUES IN OUR EXAMPLE? THE WORDS

THE CATEGORIES WE SEEK TO CLASSIFY INTO ARE CALLED LABELS

("SPAM" AND "HAM" IN OUR EXAMPLE)

OK! SO A NAIVE BAYES CLASSIFIER
IS A SUPERVISED MACHINE-LEARNING
BASED APPROACH TO SPAM DETECTION

(MEMORIZE THAT AND REPEAT AT EVERY COCKTAIL PARTY YOU EVER ATTEND)

SERIOUSLY THOUGH - MACHINE LEARNING IS NOT ROCKET SCIENCE - IT JUST HAS A LOT OF INTIMIDATING TERMS WE HAVE TO GET USED TO USING WITH CONFIDENCE