BAYES THEOREM

P(Dog ate Trash given that Trash is on floor)

AFTER-THE-FACT

=

THAT'S EXACTLY WHAT BAYES' RULE DOES FOR US

LIKELIHOOD

P(Trash was on floor given that Dog ate trash)

x P(Dog ate Trash)

BEFORE-THE-FACT

P(Trash is on floor)

EVIDENCE

SIMILARLY..

P(Bear ate Trash given that Trash is on floor)

=

P(Trash was on floor given that Bear ate trash)

x P(Bear ate Trash)

P(Trash is on floor)

THEN, TO KNOW IF THE TRASH WAS EATEN BY DOG OR BEAR, WE SIMPLY NEED TO COMPUTE THE NUMERATORS - THE DENOMINATOR IS NOT REQUIRED

THIS IS PRECISELY THE FOUNDATION OF A

NAIVE BAYES CLASSIFIER

APPLE OR BANANA?

LET'S SEE HOW WE COULD BUILD A NAIVE BAYES CLASSIFIER TO CLASSIFY FRUITS INTO APPLES AND BANANAS

WE CAPTURE 3 ATTRIBUTES OF EACH FRUIT - LENGTH, BREADTH, AND COLOR

TRAINING DATA

WE HAVE INFORMATION ABOUT A LARGE NUMBER OF CORRECTLY CLASSIFIED FRUITS

PROBLEM INSTANCE

NOW, GIVEN A FRUIT, WE HAVE TO OPINE: IS THAT FRUIT AN APPLE, OR A BANANA?

THIS IS A CLASSIFICATION PROBLEM -

FRUITS ARE THE INSTANCES

LENGTH, BREADTH AND COLOR ARE FEATURES

(THUS EACH FRUIT HAS A FEATURE VECTOR OF LENGTH 3)

"APPLE" AND "BANANA" ARE LABELS

(THERE ARE JUST 2, SO THIS IS A BINARY CLASSIFICATION PROBLEM)

APPLES

IN OUR TRAINING SET, 55% OF THE FRUITS ARE APPLES, AND 45% ARE BANANAS

LENGTH: NORMALLY DISTRIBUTED, MEAN = 5 INCHES, STANDARD DEVIATION = 1 INCH

BREADTH: NORMALLY DISTRIBUTED, MEAN = 5 INCHES, STANDARD DEVIATION = 1 INCH

COLOR: GREEN 30% OF THE TIME, RED 50% OF THE TIME, YELLOW 20% OF THE TIME

WE GET A GREEN FRUIT, 6 INCHES LONG, 3.5 INCHES BROAD

APPLE OR BANANA?

BANANAS

LENGTH: NORMALLY DISTRIBUTED, MEAN = 5 INCHES, STANDARD DEVIATION = 1.5 INCHES

BREADTH: NORMALLY DISTRIBUTED, MEAN = 2 INCHES, STANDARD DEVIATION = 0.3 INCHES

COLOR: GREEN 50% OF THE TIME, YELLOW 50% OF THE TIME

$$P(f \text{ is Apple} / \{\text{length} = 6, \text{breadth}, \text{color} = \text{green}\}) = \frac{55\% \; P(f \text{ is Apple}) \times P(\text{length} = 6 / f \text{ is Apple}) \; 0.24 \times P(\text{breadth} = 3.5 / f \text{ is Apple}) \; 0.13 \times P(\text{color} = \text{green} / f \text{ is Apple}) \; 30\%}{P(\text{length} = 6, \text{breadth} = 3.5, \text{color} = \text{green})} = 0.005$$

$$P(f \text{ is Banana} / \{\text{length} = 6, \text{breadth}, \text{color} = \text{green}\}) = \frac{45\% \; P(f \text{ is Banana}) \times P(\text{length} = 6 / f \text{ is Banana}) \; 0.21 \times P(\text{breadth} = 3.5 / f \text{ is Banana}) \; 0.000004 \times P(\text{color} = \text{green} / f \text{ is Banana}) \; 50\%}{P(\text{length} = 6, \text{breadth} = 3.5, \text{color} = \text{green})} = 0.00000018$$

THE DENOMINATORS ARE THE SAME,
SIMPLY CALCULATE THE NUMERATORS,
AND CHOOSE LABEL WHERE
PROBABILITY IS HIGHER

NOW FOR THE LENGTH AND BREADTH,
USE STANDARD PROBABILITY TABLES
TO GET PROBABILITY GIVEN THE
MEAN AND STANDARD DEVIATION

THE "AFTER-THE-FACT" PROBABILITY
THAT THIS IS AN APPLE IS HIGHER
THAN THE "AFTER-THE-FACT" PROBABILITY
THAT THIS IS A BANANA -

SO LABEL OUR PROBLEM INSTANCE
AS AN APPLE

YOU CAN SEE FROM THIS EXAMPLE
THAT WE DID NOT TAKE INTO ACCOUNT
ANY "JOINT PROBABILITIES"

THE PROBABILITIES AROUND EACH
OF THE FEATURES WERE MEASURED
INDEPENDENTLY

(FOR INSTANCE, HAD WE DONE A JOINT
PROBABILITY CALCULATION AROUND A FRUIT
WITH LENGTH:BREADTH RATIO OF 6:3.5, WE
MIGHT HAVE CONCLUDED THIS IS A BANANA)

NAIVE BAYES IS CALLED NAIVE
BECAUSE IT ASSUMES THE FEATURES
ARE INDEPENDENT IN THEIR
PROBABILITY DISTRIBUTIONS

BTW, YOU MIGHT BE WONDERING
HOW WE CALCULATED THE MEAN
AND STANDARD DEVIATIONS OF THE
LENGTHS AND BREADTHS OF APPLES
AND BANANAS FROM THE TRAINING
DATA

IN OUR TRAINING SET, 55% OF THE FRUITS ARE APPLES, AND 45% ARE BANANAS

APPLES

LENGTH: NORMALLY DISTRIBUTED, MEAN = 5 INCHES, STANDARD DEVIATION = 1 INCH

BREADTH: NORMALLY DISTRIBUTED, MEAN = 5 INCHES, STANDARD DEVIATION = 1 INCH

COLOR: GREEN 30% OF THE TIME, RED 50% OF THE TIME, YELLOW 20% OF THE TIME

WE GET A GREEN FRUIT, 6 INCHES LONG, 3.5 INCHES BROAD

APPLE OR BANANA?

HOW ARE THESE ESTIMATED?

BANANAS

LENGTH: NORMALLY DISTRIBUTED, MEAN = 5 INCHES, STANDARD DEVIATION = 1.5 INCHES

BREADTH: NORMALLY DISTRIBUTED, MEAN = 2 INCHES, STANDARD DEVIATION = 0.3 INCHES

COLOR: GREEN 50% OF THE TIME, YELLOW 50% OF THE TIME

BTW, YOU MIGHT BE WONDERING HOW WE CALCULATED THE MEAN AND STANDARD DEVIATIONS OF THE LENGTHS AND BREADTHS OF APPLES AND BANANAS FROM THE TRAINING DATA

THEY ARE CALCULATED FROM THE TRAINING DATA, USUALLY USING A FAMOUS MATHEMATICAL TECHNIQUE CALLED

MAXIMUM LIKELIHOOD ESTIMATION