NAIVE BAYES CLASSIFIER

K-NEAREST NEIGHBOURS

LOGISTIC REGRESSION

LINEAR REGRESSION

THESE ARE ALL EXAMPLES OF

SUPERVISED LEARNING

ALL SUPERVISED LEARNING TECHNIQUES
HAVE A STAGE CALLED

TRAINING THE MODEL

TRAINING DATA

WE START WITH A SET OF DATA
FOR WHOM THE OUTPUT OF THE
ALGORITHM IS ALREADY KNOWN

TRAINING INVOLVES STUDYING THE
UNDERLYING PATTERNS/REGULARITIES
IN THE TRAINING DATA

THE ALGORITHM ANALYZES
THE TRAINING DATA AND
PRODUCES AN INFERRED FUNCTION

THESE PATTERNS ARE CAPTURED IN
A FUNCTION THAT CAN THEN BE APPLIED
ON A NEW PROBLEM INSTANCE

THE FUNCTION THAT IS DERIVED
AFTER ANALYZING THE
TRAINING DATA IS CALLED THE

MODEL

A "GOOD" MODEL DOES TWO THINGS

IT IS TYPICALLY IMPOSSIBLE TO DO BOTH OF THESE THINGS

CAPTURES ALL THE PATTERNS IN THE TRAINING DATA

THIS IS MEASURED BY HOW ACCURATELY THE MODEL PREDICTS/CLASSIFIES THE TRAINING DATA

CORRECTLY COMPUTES THE OUTPUT VALUE FOR A NEW INSTANCE

THIS IS MEASURED BY HOW WELL IT CAN PREDICT/CLASSIFY A PREVIOUSLY UNSEEN INSTANCE

THE MORE COMPLEX THE MODEL THE BETTER IT REPRESENTS THE TRAINING DATA. BUT, THERE IS A BALANCE TO BE ACHIEVED HERE

IF ITS NOT COMPLEX ENOUGH, IT MIGHT MISS OUT ON AN IMPORTANT DYNAMIC PRESENT IN THE DATA

UNDERFITTING

IF THE MODEL IS TOO COMPLEX IT WILL PICK UP SPECIFIC RANDOM FEATURES IN THE TRAINING DATA, I.E. NOISE

OVERFITTING

THIS PROBLEM GETS TO THE HEART OF SOMETHING CALLED THE

BIAS - VARIANCE TRADEOFF

THIS PROBLEM GETS TO THE HEART OF SOMETHING CALLED THE

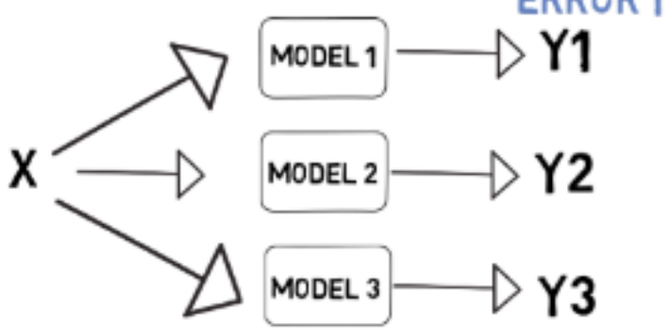
BIAS - VARIANCE TRADEOFF

SAY YOU HAVE SEVERAL TRAINING DATA SETS, ALL EQUALLY GOOD IE. THEY ARE REPRESENTATIVE OF THE ENTIRE POPULATION

A LEARNING ALGORITHM WILL USUALLY PRODUCE A SLIGHTLY DIFFERENT MODEL WITH EACH TRAINING DATASET

NOW LET'S SAY YOU HAVE A NEW INPUT X

TYPICALLY THERE ARE TWO TYPES OF ERRORS



ERROR 1

EACH MODEL PRODUCES A DIFFERENT OUTPUT. THIS ALGORITHM HAS A HIGH

VARIANCE ERROR

HIGH VARIANCE MEANS THE ALGORITHM PRODUCES A MODEL THAT IS TOO SPECIFIC TO THE TRAINING DATA

THIS IS A TYPICAL CASE OF

OVERFITTING

ERROR 2

Y

Y

Y

IT IS NOT POSSIBLE TO MINIMIZE BOTH THESE ERRORS SIMULTANEOUSLY. HIGH BIAS USUALLY MEANS LOW VARIANCE AND LOW BIAS MEANS HIGH VARIANCE

EACH MODEL SYSTEMATICALLY PRODUCES THE SAME INCORRECT OUTPUT. THIS ALGORITHM HAS A HIGH

BIAS ERROR

HIGH BIAS MEANS THE ALGORITHM HAS NOT PICKED UP SOME DEFINING PATTERN PRESENT IN THE DATA

THIS IS A TYPICAL CASE OF

UNDERFITTING

THIS IS EXACTLY WHAT THE

BIAS-VARIANCE TRADEOFF IS ALL ABOUT