First, a true story, from
Greenwich Connecticut, 2007

First, a true story, from Greenwich Connecticut, 2007

Financial markets were at all-time highs (this is before the Great Financial Crisis)

Mr V worked at a quant hedge fund as a trader of credit derivatives.

Mr V was paid to build financial models, convince the hedge fund's owner that they were good models, and then trade them with the HFs' money

First, a true story, from Greenwich Connecticut, 2007

Financial markets were at all-time highs (this is before the Great Financial Crisis)

Mr V worked at a quant hedge fund as a trader of credit derivatives.

HOW?

Mr V was paid to build financial models, **convince the hedge fund's owner that they were good models**, and then trade them with the HFs' money

Using Backtests of course!

Using Backtests of course!

A backtest "runs" the model on recent market data, and tells how it performed.

Easy as Pie!!

Err..wasn't the model also built using recent market data?

Err..Yes..

Its really not an exaggeration that Overfitting ML models directly contributed to causing the GFC.

OVERFITTING

IS THE BUGBEAR OF MACHINE LEARNING

SO WHAT IS OVERFITTING? AND WHY IS IT SUCH A PROBLEM?

CROSS VALIDATION

REGULARIZATION

ENSEMBLE LEARNING

SOME OF THE WAYS TO MITIGATE THIS PROBLEM

FRODO AND SAM ATE AT A RESTAURANT EVERY DAY LAST WEEK AND RATED IT ON EACH DAY

MONDAY	GOOD
TUESDAY	BAD
WEDNESDAY	GOOD
THURSDAY	GOOD
FRIDAY	GOOD
SATURDAY	BAD
SUNDAY	GOOD

FRODO AND SAM ATE AT A RESTAURANT EVERY DAY LAST WEEK AND RATED IT ON EACH DAY

MONDAY	GOOD
TUESDAY	BAD
WEDNESDAY	GOOD
THURSDAY	GOOD
FRIDAY	GOOD
SATURDAY	BAD
SUNDAY	GOOD

AT THE END OF THE WEEK,

FRODO SAYS

THE FOOD IS GOOD AT THIS RESTAURANT

SAM SAYS

THE FOOD IS GOOD AT THIS RESTAURANT ON ALL DAYS EXCEPT TUESDAYS AND SATURDAYS

WHICH ONE OF THEM IS RIGHT?

WHICH ONE OF THEM IS RIGHT?

HOW DO WE MEASURE THIS?

WE COULD CHECK EACH OF THEIR

STATEMENTS

MODELS

AGAINST THE DATA WE ALREADY HAVE

TRAINING SET

WHICH ONE OF THEM IS RIGHT?

	TRAINING SET	FRODO'S MODEL	SAM'S MODEL
MONDAY	GOOD	GOOD	GOOD
TUESDAY	BAD	GOOD	BAD
WEDNESDAY	GOOD	GOOD	GOOD
THURSDAY	GOOD	GOOD	GOOD
FRIDAY	GOOD	GOOD	GOOD
SATURDAY	BAD	GOOD	BAD
SUNDAY	GOOD	GOOD	GOOD
		71%	100%

WE COULD CHECK EACH OF
THEIR STATEMENTS
AGAINST THE DATA WE ALREADY HAVE

ACCURACY

WHICH ONE OF THEM IS RIGHT?

		FRODO'S MODEL	SAM'S MODEL
MONDAY	GOOD	GOOD	GOOD
TUESDAY	BAD	GOOD	BAD
WEDNESDAY	GOOD	GOOD	GOOD
THURSDAY	GOOD	GOOD	GOOD
FRIDAY	GOOD	GOOD	GOOD
SATURDAY	BAD	GOOD	BAD
SUNDAY	GOOD	GOOD	GOOD
		71%	100%

ACCURACY

ON THE TRAINING SET, FRODO'S MODEL HAS 71% ACCURACY AND SAM'S MODEL HAS 100% ACCURACY

FROM THIS, IT SEEMS LIKE SAM'S MODEL IS BETTER.

SAM AND FRODO GO BACK TO THE RESTAURANT NEXT WEEK

WHICH ONE OF THEM IS RIGHT?

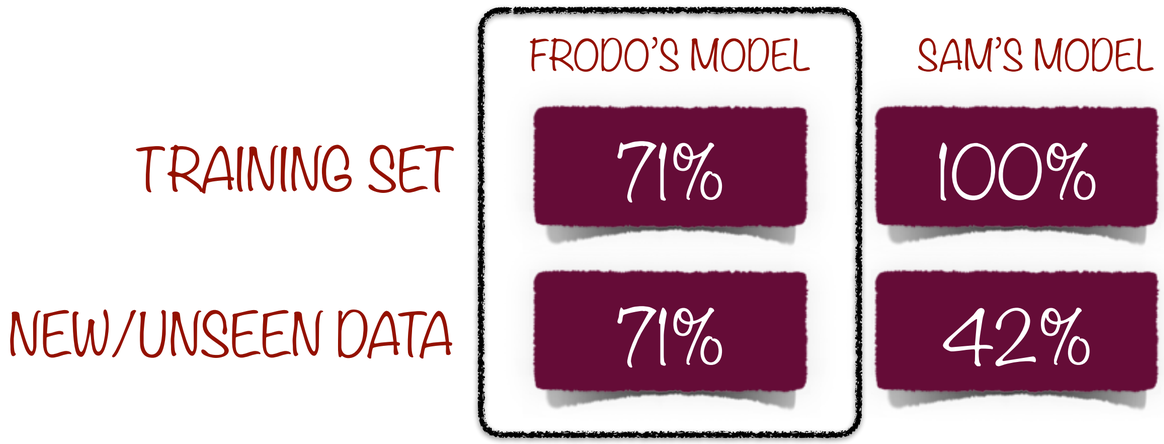
			FRODO'S MODEL	SAM'S MODEL
WEEK 1	MONDAY	GOOD	GOOD	GOOD
	TUESDAY	BAD	GOOD	BAD
	WEDNESDAY	GOOD	GOOD	GOOD
	THURSDAY	GOOD	71%	100%
	FRIDAY	GOOD	GOOD	GOOD
	SATURDAY	BAD	GOOD	BAD
	SUNDAY	GOOD	GOOD	GOOD
WEEK 2	MONDAY	GOOD	GOOD	GOOD
	TUESDAY	GOOD	GOOD	BAD
	WEDNESDAY	BAD	GOOD	GOOD
	THURSDAY	GOOD	71%	42%
	FRIDAY	GOOD	GOOD	GOOD
	SATURDAY	GOOD	GOOD	BAD
	SUNDAY	BAD	GOOD	GOOD

ON THE TRAINING SET, FRODO'S MODEL HAS 71% ACCURACY AND SAM'S MODEL HAS 100% ACCURACY

SAM AND FRODO GO BACK TO THE RESTAURANT NEXT WEEK

ON NEW DATA, FRODO'S MODEL HAS 71% ACCURACY AND SAM'S MODEL HAS 42% ACCURACY

WHICH ONE OF THEM IS RIGHT?



WHAT HAPPENED HERE?

FRODO'S MODEL IS THE BETTER MODEL

IT GENERALIZES WELL

FRODO'S MODEL PERFORMS WELL ON BOTH TRAINING AND NEW/UNSEEN DATA

WHAT HAPPENED HERE?

	FRODO'S MODEL	SAM'S MODEL
TRAINING SET	71%	100%
NEW/UNSEEN DATA	71%	42%

THE FOOD IS GOOD AT THIS RESTAURANT

THE FOOD IS GOOD AT THIS RESTAURANT ON ALL DAYS EXCEPT TUESDAYS AND SATURDAYS

FRODO'S MODEL IS SIMPLER ("DUMBER", IN FACT), YET IT PERFORMS BETTER

SAM'S MODEL IS MORE COMPLEX, AND MORE ACCURATE ON THE TRAINING SET

YET, IT PERFORMS BADLY ON NEW DATA

IE, SAM'S MODEL DOES NOT GENERALIZE WELL

THE FOOD IS GOOD AT THIS RESTAURANT ON ALL DAYS

EXCEPT TUESDAYS AND SATURDAYS

SAM'S MODEL PICKS UP ON A RELATIONSHIP BETWEEN THE WEEKDAY AND THE QUALITY OF FOOD

THIS RELATIONSHIP HOWEVER, IS SPECIFIC TO THE TRAINING SET, AND NOT TRUE IN GENERAL

SAM'S MODEL IS A PERFECT EXAMPLE OF

OVERFITTING

OVERFITTING OCCURS WHEN A MODEL PICKS UP ON RANDOM PHENOMENA OR NOISE PRESENT IN THE TRAINING SET

INSTEAD OF THE UNDERLYING RELATIONSHIP BETWEEN THE INPUT AND OUTPUT

OVERFITTING

BUT WHY IS OVERFITTING SUCH A COMMON PROBLEM?

THE TRAINING SET IS ONLY PART OF A MUCH LARGER SET

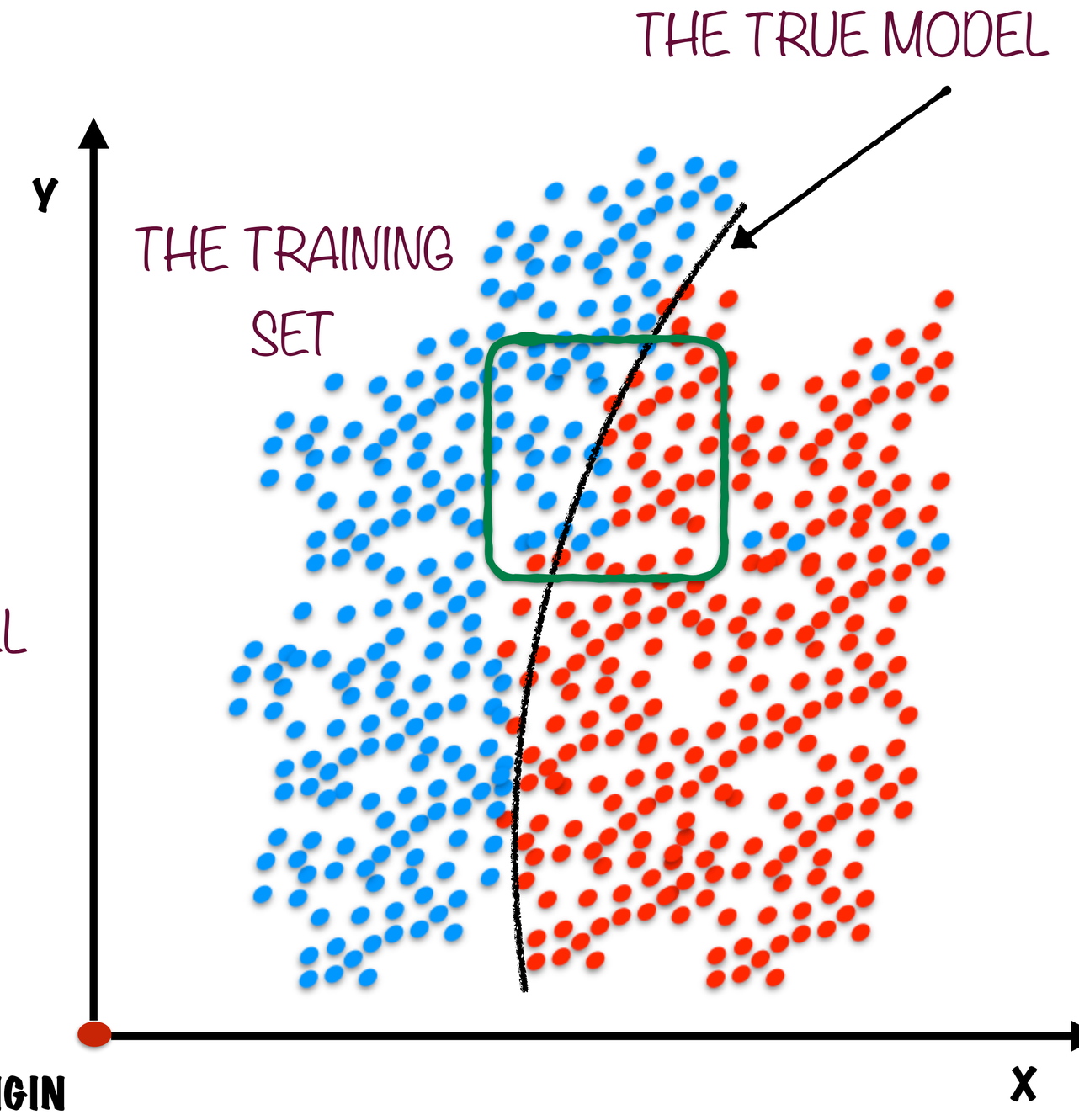
WE ARE TRYING TO FIND A MODEL, THAT DESCRIBES THIS MUCH LARGER SET

IT'S LIKE TRYING TO DESCRIBE PHOTOGRAPH, BUT YOU ARE ONLY SHOWN A SMALL, ZOOMED IN PORTION OF THE PHOTOGRAPH

OVERFITTING

YOU WANT TO CLASSIFY EMAILS AS SPAM OR HAM

THESE ARE ALL THE EMAILS IN ALL INBOXES IN THE WORLD (BOTH PAST AND FUTURE)

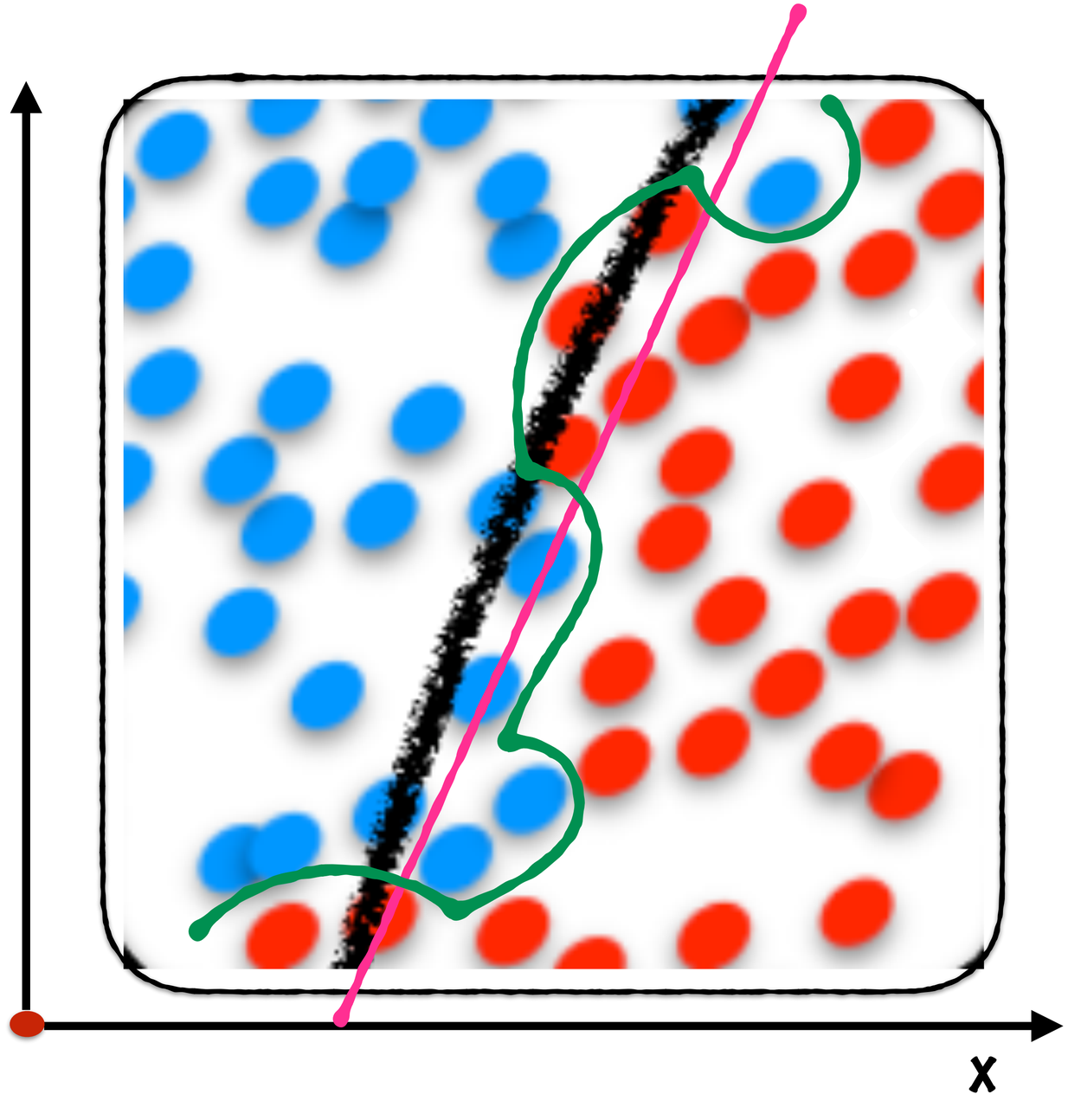


OVERFITTING

1. A SIMPLE LINEAR MODEL

2. AN OVERFITTED MODEL

(USUALLY A POLYNOMIAL OF EXTREMELY HIGH ORDER)



BECAUSE THE TRAINING DATA IS ONLY A PART OF THE PICTURE

WE CAN'T TELL FOR SURE WHAT IS RELEVANT AND WHAT'S NOT

OVERFITTING

IS A PRETTY DIFFICULT PROBLEM TO SOLVE

BY AVOIDING OVERFITTING, WE CAN END UP WITH THE OPPOSITE ERROR OF UNDERFITTING

THIS IS THE FAMOUS BIAS-VARIANCE TRADEOFF

OVERFITTING

IS THE BUGBEAR OF MACHINE LEARNING

SO WHAT IS OVERFITTING? AND WHY IS IT SUCH A PROBLEM?

CROSS VALIDATION

REGULARIZATION

ENSEMBLE LEARNING

SOME OF THE WAYS TO MITIGATE THIS PROBLEM

OVERFITTING

IS THE BUGBEAR OF MACHINE LEARNING

SO WHAT IS OVERFITTING? AND WHY IS IT SUCH A PROBLEM?

CROSS VALIDATION

REGULARIZATION

ENSEMBLE LEARNING

SOME OF THE WAYS TO MITIGATE THIS PROBLEM

CROSS VALIDATION

IS A TECHNIQUE FOR MODEL SELECTION

PERFORMING WELL ON TRAINING DATA IS NO GUARANTEE FOR A GOOD MODEL

IN ORDER TO TEST THE PERFORMANCE OF A MODEL, IT WOULD BE NICE IF WE CAN

A GOOD MODEL IS ONE THAT PERFORMS WELL ON DATA IT HAS NOT SEEN BEFORE → GET SOME DATA THAT WE MIGHT SEE IN THE FUTURE (SOME NEW DATA)

GET MULTIPLE TRAINING DATA SETS

A GOOD MODEL DOES NOT OVERFIT → WE CAN THEN FIND A MODEL THAT PERFORMS WELL ACROSS TRAINING DATA SETS, AND NOT JUST ON ONE TRAINING SET

IN ORDER TO TEST THE PERFORMANCE OF A MODEL, IT WOULD BE NICE IF WE CAN

GET SOME DATA THAT WE MIGHT SEE IN THE FUTURE (SOME NEW DATA)

GET MULTIPLE TRAINING DATA SETS

WE CAN THEN FIND A MODEL THAT PERFORMS WELL ACROSS TRAINING DATA SETS, AND NOT JUST ON ONE TRAINING SET

CROSS VALIDATION IS A COMBINATION OF THESE TWO IDEAS

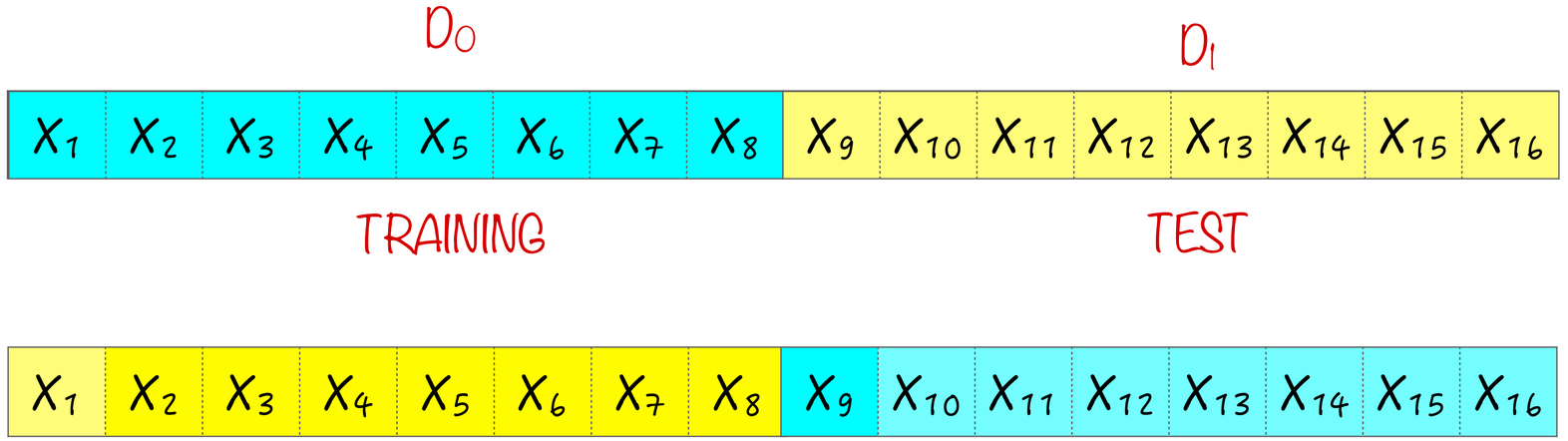
KEEP SOME DATA ASIDE FOR PERFORMANCE TESTING

CREATE MULTIPLE TRAINING SETS - EACH ONE A SUBSET OF THE ORIGINAL TRAINING SET

THE BELOW TABLE REPRESENTS THE ENTIRE TRAINING DATA SET

1. DIVIDE THE TRAINING SET RANDOMLY INTO TWO EQUAL PARTS - D_0 AND D_1

2. USE D_0 TO TRAIN THE MODEL AND D_1 TO TEST THE PERFORMANCE



3. THEN, USE D_1 TO TRAIN THE MODEL AND D_0 TO TEST THE PERFORMANCE

THE BEST MODEL IS THE ONE WITH BEST AVERAGE PERFORMANCE

THIS TECHNIQUE IS CALLED

2-FOLD CROSS VALIDATION

WHEN DO YOU USE CROSS VALIDATION?

1. TO CHOOSE BETWEEN DIFFERENT ALGORITHMS

SUPPORT VECTOR MACHINES VS K-NEAREST NEIGHBOURS

2. TO TUNE THE PARAMETERS OF THE ALGORITHM

THE VALUE OF K IN K-NEAREST NEIGHBOURS,
THE MAX DEPTH OF A DECISION TREE

3. TO IDENTIFY THE FEATURES THAT ARE RELEVANT

IF YOU HAVE 20 FEATURES, SHOULD YOU USE ALL OF THEM? OR A SUBSET?

OVERFITTING

IS THE BUGBEAR OF MACHINE LEARNING

SO WHAT IS OVERFITTING? AND WHY IS IT SUCH A PROBLEM?

CROSS VALIDATION

REGULARIZATION

ENSEMBLE LEARNING

SOME OF THE WAYS TO MITIGATE THIS PROBLEM

OVERFITTING

IS THE BUGBEAR OF MACHINE LEARNING

SO WHAT IS OVERFITTING? AND WHY IS IT SUCH A PROBLEM?

CROSS VALIDATION

REGULARIZATION

ENSEMBLE LEARNING

SOME OF THE WAYS TO MITIGATE THIS PROBLEM

REGULARIZATION

PENALIZES MODELS WHICH ARE TOO COMPLEX

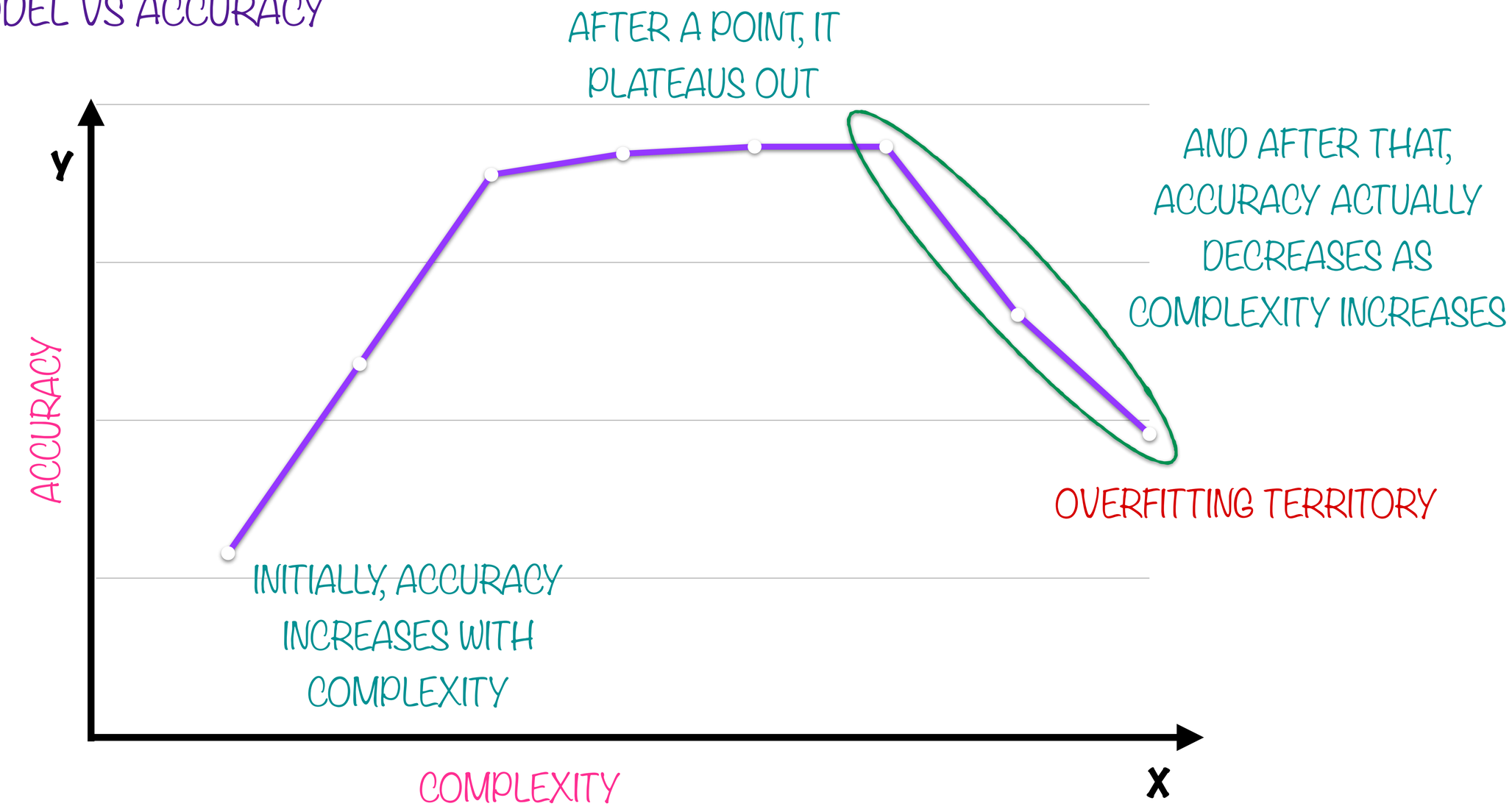
OVERFITTING OCCURS BECAUSE THE MODEL HAS BECOME NEEDLESSLY COMPLEX

EXAMPLES OF COMPLEXITY MEASURES

(THE NUMBER OF BRANCHES IN A DECISION TREE (OR) THE ORDER OF THE POLYNOMIAL USED TO REPRESENT A CURVE)

LET'S SAY YOU PLOTTED COMPLEXITY OF A MODEL VS ACCURACY

LET'S SAY YOU PLOTTED COMPLEXITY OF A MODEL VS ACCURACY



REGULARIZATION

PENALIZES MODELS WHICH ARE TOO COMPLEX

FINDING A MODEL USUALLY INVOLVES MINIMIZING AN ERROR FUNCTION

FOR EXAMPLE, THE ERROR FUNCTION COULD BE THE SUM OF SQUARES OF DISTANCES BETWEEN THE PREDICTED POINTS AND THE ACTUAL POINTS IN THE TRAINING SET

LET THE ERROR FUNCTION BE E(f) FOR A MODEL f

LET THE ERROR FUNCTION BE E(f) FOR A MODEL f

A REGULARIZATION TERM IS
ADDED TO THIS FUNCTION

$$E'(f) = E(f) + \lambda R(f)$$

NEW ERROR FUNCTION THAT NEEDS TO BE MINIMIZED

A PARAMETER THAT CONTROLS THE IMPORTANCE OF THE REGULARIZATION TERM

REGULARIZATION TERM THAT INCREASES WITH COMPLEXITY

LET THE ERROR FUNCTION BE E(f) FOR A MODEL f

A REGULARIZATION TERM IS ADDED TO THIS FUNCTION

$$E'(f) = E(f) + \lambda R(f)$$

NEW ERROR FUNCTION THAT NEEDS TO BE MINIMIZED

A PARAMETER THAT CONTROLS THE IMPORTANCE OF THE REGULARIZATION TERM

REGULARIZATION TERM THAT INCREASES WITH COMPLEXITY

WE GET A MODEL THAT GIVES LOW ERROR ON THE TRAINING SET, WHILE KEEPING THE COMPLEXITY LOW AS WELL