GET THE LAY OF THE LAND: TYPES OF ML PROBLEMS

PROBLEMS WHERE ML IS OFTEN APPLIED

CLASSIFICATION

CLUSTERING

ASSOCIATION DETECTION

ANOMALY DETECTION

DIMENSIONALITY REDUCTION

TECHNIQUES TO SOLVE THOSE PROBLEMS

NAIVE BAYES

K-NEAREST NEIGHBOR

SUPPORT VECTOR MACHINES

NEURAL NETWORKS

DECISION TREES

LINEAR REGRESSION

LOGISTIC REGRESSION

APPLICATIONS OF THESE SOLVED PROBLEMS

SPAM DETECTION

TOPIC MODELING

SENTIMENT ANALYSIS

RECOMMENDATIONS

GENRE CLASSIFICATION

QUANT TRADING

CLASSIFICATION PROBLEMS

IS AN EMAIL SPAM OR HAM?

WE HAVE A POPULATION (ALL EMAILS)

THAT POPULATION IS DIVIDED INTO CATEGORIES (SPAM AND HAM)

WE HAVE A SET OF INSTANCES FOR WHICH THE CORRECT CATEGORY MEMBERSHIP IS KNOWN (TRAINING DATA - EMAILS ALREADY CORRECTLY MARKED AS SPAM OR HAM)

WE ARE GIVEN A PROBLEM INSTANCE (A NEW EMAIL COMES IN)

WE NEED TO ASSIGN A CATEGORY TO THE PROBLEM INSTANCE (IS THE NEW EMAIL SPAM OR HAM?)

AN ALGORITHM THAT IMPLEMENTS CLASSIFICATION IS CALLED **A CLASSIFIER**

WE HAD VERY BRIEFLY TOUCHED UPON THREE TYPES OF CLASSIFIERS -

A NAIVE BAYES CLASSIFIER
A K-NEAREST NEIGHBOR CLASSIFIER
A SUPPORT VECTOR MACHINE CLASSIFIER

CLASSIFICATION IS A FORM OF SUPERVISED LEARNING

BECAUSE A SET OF CORRECTLY CLASSIFIED INSTANCES IS AVAILABLE (THE TRAINING DATA)

CLUSTERING

GIVEN A SET OF INSTANCES (ALL FACEBOOK USERS)

DIVIDE THOSE INSTANCES INTO CLUSTERS, SO THAT INSTANCES WITHIN A CLUSTER ARE MORE SIMILAR TO EACH OTHER THAN TO INSTANCES IN OTHER CLUSTERS (DISJOINT COMMUNITIES OF FACEBOOK USERS)

CLUSTERING IS VERY CLOSELY RELATED TO CLASSIFICATION -

BOTH CLUSTERING AND CLASSIFICATION DIVIDE A SET OF INSTANCES INTO DISJOINT GROUPS

CLASSIFICATION IS A BIT MORE FOCUSED ON CLASSIFYING A PROBLEM INSTANCE

(A NEW USER HAS SIGNED UP - WHAT COMMUNITY WILL SHE MOST LIKELY BELONG TO?)

CLUSTERING ON THE OTHER HAND IS LARGELY FOCUSED ON THE PROCESS OF DIVVYING UP THE INSTANCES WE ALREADY HAVE

CLUSTERING IS A PROTOTYPICAL EXAMPLE OF

UNSUPERVISED LEARNING

CLUSTERING ALGORITHMS

K-MEANS CLUSTERING

HIERARCHICAL CLUSTERING

DENSITY-BASED CLUSTERING

DISTRIBUTION-BASED CLUSTERING

ASSOCIATION RULE LEARNING

LET'S SAY YOU WORK AT AN ECOMMERCE COMPANY AS A CATEGORY MANAGER

YOU ARE IN CHARGE OF SELLING MOBILE ACCESSORIES - THINGS LIKE CELLPHONE CASES, CHARGERS ETC

YOUR JOB IS TO SELL A LOT OF STUFF, AND AT PRICES AS HIGH AS POSSIBLE, AND SPEND AS LITTLE AS POSSIBLE ON MARKETING

WHAT IF YOU COULD FIGURE OUT, SOMEHOW,
THAT FOLKS WHO BOUGHT ADAPTERS AND EARPLUGS
WERE MORE LIKELY TO BUY CELLPHONE CHARGERS –

THAT INFORMATION COULD REALLY HELP –
YOU COULD PERHAPS "BUNDLE" ADAPTERS
AND CELLPHONE CHARGERS, OR DISPLAY
PROMOTIONAL PRICING, OR OFFER QUANTITY
DISCOUNTS

{Adapter, Earmuffs} -> {Cellphone Charger}

IDENTIFYING RULES OF THIS SORT
IS EXACTLY WHAT **ASSOCIATION RULE LEARNING**

ANOMALY DETECTION

SAY YOU ARE THE NETWORK ADMINISTRATOR AT A UNIVERSITY RESEARCH LAB

YOU MIGHT HAVE TO DEAL WITH – LITERALLY – HUNDREDS OF ATTEMPTED HACKER ATTACKS A DAY

HOW WOULD YOU KNOW WHAT INCOMING TRAFFIC ON YOUR NETWORK IS INNOCUOUS, AND WHAT IS POTENTIALLY HARMFUL?

CHANCES ARE THAT YOU'D RELY ON AN INTRUSION DETECTION SYSTEM –

WHICH IN TURN WORKS USING ANOMALY DETECTION TECHNIQUES

WE WON'T SPEND A LOT OF TIME ON ANOMALY DETECTION, BUT DO TAKE A MOMENT TO PONDER THAT -

(SUPERVISED LEARNING APPROACH)

ANOMALY DETECTION COULD BE VIEWED AS A **CLASSIFICATION PROBLEM**

WHERE WE SEEK TO LABEL NETWORK PACKETS AS "INNOCUOUS" OR "HARMFUL"

ANOMALY DETECTION COULD ALSO BE VIEWED AS A **CLUSTERING PROBLEM**

BY VIEWING INNOCUOUS TRAFFIC AS THE "NORM", AND SEEKING OUTLIERS FROM THIS NORM

(UNSUPERVISED LEARNING APPROACH)

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