No: Date: / /

Week 2

Define ML as a practice

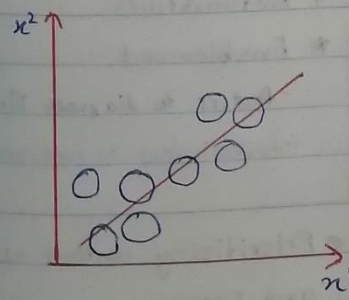
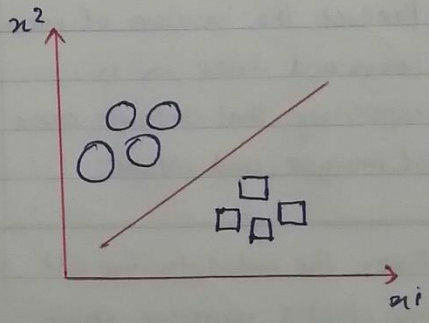
Introduction I.

* Common ML problem types
  1) supervised
  2) Unsupervised

Supervised learning

* Supervised learning problems can categorized into two problems.
  1) Regression :- When we trying to predict continous numeric value, we use regression.
     Ex:- predicting the price of a house

  2) Classification :- When we are trying to predict a distinct category or class
     Ex:- predicting whether something is true of false

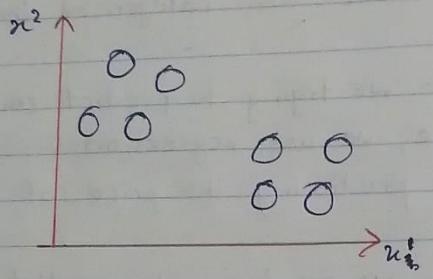
Regression	classification
Predict numeric values	Predict distinct categories
	

Unsupervised Learning

* This did not involve predicting something specific
* This is all about uncovering patterns or structure in your data. It means, seeing your data naturally falls into different groups or clusters.

Clustering

* Uncover patterns or structure in data



Examples

Problem statement	Type	Benefit
Predict how many cars will pass through a specific intersection between 8 and 9 PM	Regression	* Optimize the traffic lights * Offer rerouting option for GPS apps
Predict the location of each image and video in a repository that contains 1000s of images and videos	classification	* Automation * Enablement - Doctors to diagnose illness - Investigators to find missing things
Given the metadata for all our classes, uncover which classes have a significant content overlap	Clustering	* Prioritizing content updates and purging * Identify new content needs

No:

Date: __/__/__

* Sometimes, problems that we are focusing on is really big. To solve it we have to break it down into smaller problem statements.

Example :- how do we stop COVID-19?

Problem Statement	Type	Benefit
Predict whether or not a patient is possitive for covid-19 using CT scans / imagery	classification	* Efficient diagnostic tool; instant test result (testing kit takes few hours.
Predict the number of patients that will be infected with the virus in a specific region	Regression	* Preparing health care messures, optimize staffing and room availability
Uncover similarities or differences between those who have already tested possitive	Clustering	* To identify whether or not there are different of the virus. -how virus affect to different groups, regions, Etc.

Standard Algorithm and Data

* Machine learning uses standard algorithms to analyze data to derive predictive insights and make repeated decisions.

* Above definition applies to the regression and classification types of machine learning problems.

* A single ML algorithm can use to solve different type of problems. It happens because of our data.

* A algorithm that trained with data is called a "trained model"

* To get accurated moden in regression and classification problems, you need lot of ~~data~~ labled data.

Data Quality

* Bugs in ML are often cause by bugs in the data.

* ~~A bug~~ In software development, a bug is a mistake in the code that causes unexpector or undesired behavior.

* In ML, there also can be bugs in the implementation of an algorithm and bugs in data are the most common.

* Qualities of ~~good do~~ good data
  1) has coverage
  2) Clean
  3) Complete

1) Data coverage :- Rerers to the domain scope and all possible scenarios the data can account for.

:- All possible input and output data

2) Data cleanliness :- Dirt in data refers to anything that can detract the model from making accurate predictions or understanding data behavior.

:- Sometimes called "data consistency"

5) Data completeness :- Refers ~~tho~~ to the avilability of sufficient data about the world to replace human ~~kodo~~ knowledge.

* Data is the only tunneled through which your model views the world

Q1) Use machine learning to predict staffing to predict staffing requirments per retail branch. To ~~acur~~ accuratly predict the number of employees you will need per branch, ~~what data would you need~~?

I) what data would you need?
II) How would you collect it?
III) How might you broaden the coverage, cleanliness or completeness?

I) Store size
Average number of customers.
Number of different departments.
Number of employes per department.
Number of self checkoutr stations.
Average of returns at the store
Average wait time at customer service

II) Manager and Data analyst

III) Consistency in data format
Completeness :- no empty fields.
Coverage :- inputs and outputs related to refuel.

Predictive Insights and Decisions

Predictive Insights

* ML is a way to derive predictive insights from data
  - Dashboard and reports are backward-looking
  - Predictive analytics are forward looking

Example:- A business analyst reviews a report an sees that demand is increasing for a specific product in a specific region. The analyst then suggest a new price for that product in that region that would increase profits. We can tell bussiness analyst made predictive insight.

* But, above example is not scalable. he made one predictive insight. it cannot apply to every products. there for we need ML to do that.

Decision

* Can use only repeated model in frequantly

Ex:- makin decisions using weather data. (weather data genarating daily)

* There we can select whether ML model suitable or not by analyzing decisions.

Building and Evaluating ML Models

Introduction

Features and Labels

* An input data or example has ~~tree~~ three parts
  1) Features of the example.
  2) The resulting label or classification.
  3) The label type.

Features

* Features are the attributes of an "example."
* It gives content or meaning to a piece of data.

Ex: features of a leaf are yellow, small, spoty, Etc.

* We used to features in the content of products.
  ex: a new camera feature in your phone
* But in ~~the~~ this content, feature simply means a distinctive attribute. ~~Lab~~ resulting labels
* Features are used to then identify the resulting label or classification.

if our leaves are in good condition we can label tha data set as "healthy". Unless we can name it as "ill".

Label types

* Label types can be numbers. or categories, or even ~~phas~~ phrases

ex

Label data type	Example
Nemeric	$10,000 (amount)
Categorial	High (amount high or Low)
Phrase	Let me know if you have any questions?

Q1) Use machine learning to predict the price of a house

I) What is the label?
Price

II) What is the label type?
Numeric

III) What are some relavant features?
Location and number of rooms, home style, school distric, basement, Etc.

IV) What ML problem is this?
Regression (Label is numeric value)

* Using wrong labels will lead to failiur of our ML model.

Building labeled datasets

* Suppose you own an online stock photography collection, and you want to make it easier for your users to search for relavant photos.

* In above case, you can use ML to label the photograph for catelog search ~~pupo~~ purposes.

* Google cloud vision API offers powerful, full trained ML models that assign multiple labels to images and quickly classify them into millions of predefined categories.

Google Cloud Vision API

* Google cloud vision API is available directly through a browser.
* You can visit the site using cloud.google.com/vision address.
* To use this feature you will need to upload an image from your computer.

No:

Date: __/__/____

* If a image can't identify by a service (Google cloud vision API, Etc) we need to train an ML model. For that you need a labeled dataset.

* There are few ways to obtain or build a labeled dataset.

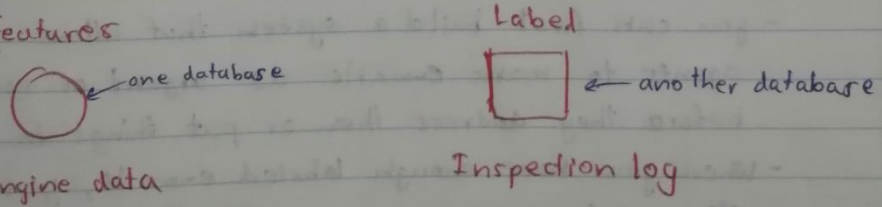
1) Use labels from historical (joined) data.
2) Use a proxy label
3) Build a labeling sistem
4) Use a labeling service

} if you don't have historical data

Use labels from historical data

* Let's assume we need to creat ML moded to predict whether a jet engine will require mechanical inspection.

step 1 :- look into historical data to find the labels and features. But it is possible the data you need is in two different places.



step 2 :- ther for you need to make sure you data must be from same place. we can call that place as "Data Warehouse."

Use a proxy label

* When we havn't historical data we can use proxy label.
* Suppose we have to build a recommendation engine for personalized shopping experience. but we don't have any historical data.
* Proxy labels measures something related to the label.
  Ex:- if you don't have customer ratings, the number of warranty claims or support phone calls might serve as a proxy for customer ratings of a product.

* Another way to get a proxy is to train and ML model to produce it for you
  Ex:- suppose you want to predict future demand for retail product. One great feature would be the amount of attention that in-store customers pay to that product

Build a Labeling System

* This method is also use when we don't have historical data.
* Suppose you are building a system to identify automatically identify incoming customer e-mails as urgent or not urgent. but you don't have historical data
  - you can first build a system that allows your triaging agents to mark e-mails as urgent or not urgent before they address them or put things in a queue.
  - When you have enough labeled e-mails, you can train ML model

No:

Date: __/__/__

Use a labeling service

* There are companies that specialize in manually labeling documents or images based on the criteria that you specify.

* It is much better to incorporate labeling into the workflow of the humans who will make decisions based on your model

Capabilities of Vision API versus Auto ML

Vision API	Auto ML vision
Trained on Google's data set with Google's labels	Train on your data set with your labels
if this can't predict the labels that you need for your data, then use Auto ML version	

* But those tools don't work for numeric data.

No:

Training An ML Model

* There are three steps to formulating the ML Problem
  1) choose an objective.
  2) choose input features.
  3) Get labels.

Exercise

- Choose the right objective - What are you optimizing for in your ML model?

Q1) You work for an online retail store and want to use ML to serve product recommendations to customers. You ~~can~~ choose to optimize recommendation system toward products "the user is ~~likely~~ likely to purchase the item". What is the risk of optimizing for purchase likelihood?

* The system might recommend only popular items.

if	result
1) The user is likely to purchase the item	⊗ Recommend popular items only
2) The user might not have otherwise bought the item	⊗ Low conversion rate
3) The purchased item is rated high	⊗ Popular items migh not be recommended.

* Evaluate ML model on a held-out portion of training data

1) Take your label data and split it into two portions. May be 80% for training and 20% for testing (evaluating)

2) Now we will feed the model the training data so it learns from that.

3) After model trained, evaluate it using remaining 20% of your data that hasn't use before.

4) Next, to evaluate the model we use a confusion matrix

Ex:

Example	Label	Prediction
0	Good	Good

		Prediction	
		Good	Fractured
Labels	Good	1	
	Fractured		

Ex:

		Prediction	
		Good	Fractured
Label	Good	893	232
	Fractured	46	675

$$\text{accuracy} = \frac{893 + 675}{893 + 232 + 675 + 46} = 0.8494$$

* When the model training is finished, you have a file that has a set of parameters representing your trained model

* The model has learned from the data and now the model's capable of doing what you trained it to do.

* All the information that learned in that model file.

* The purpose of evaluating a model is to determine whether it's ready to move to the deployment phase.

* You can deploy your model in two ways.
  1) You can download the code for your trained ML model to your local machine and Run it. But it's difficult to manage and it's definitly not scalable

  2) You can store your dataset into cloud and simply do the training job. This is call ML as a service.

Precision :- messures how precise things are. take the ratio of the true positives to the set of all

Recall :- messure of how enhaustive your model is. looks at the set of true positives that was recognized compared with the set of all possitives.

* Confidence Threshold ↑ Precision ↑ Recall ↓
* Confidence Treshold ↓ Precision ↓ Recall ↑