**What is Machine Learning?**

Motivating example

• I am interested in finding out which UTD students get internships (or jobs).

• Think of it as a function: f: X -> Y **(the best function for entire UTD)**

X is an instance (student), Y is Boolean.

X has certain attributes X = <X1, X2, X3, … ,Xn>

X = e.g. X1= GPA, X2=Taken CS 6375?

X3=Years’ experience, …

It’s **a binary classification** problem. You will see lots of such cases in this course.

Learning Scenario

• You are given access to some training data: {(X1, Y1), (X2, Y2), …, (Xn , Yn)}

• Using this training data, you try to approximate the function f. This approximation is also called a **hypothesis** (h).

• Our aim is to design a learning model that can approximate f closely.

• If Y is Boolean (or has a fixed number of values), this is called a **classification** problem.

Chart, scatter chart

Description automatically generatedClassification

• Learning to differentiate between two or more classes.

• Learning the **separating boundary**

• Most common problem in ML

e.g. classes could be - spam or not spam - family or sports car - apple or orange

Chart, scatter chart

Description automatically generatedRegression

• Classification is different from regression.

• Regression is trying to fit the best model to approximate some real valued data.

• Common in statistics, not ML.

Difference in output type

• If the output was real-valued, this problem can be solved by **statistics**.

Chart, line chart, scatter chart

Description automatically generatedStatisticians love regression – linear or non-linear. It can be used to fit a curve to data points.

A picture containing graphical user interface

Description automatically generated**Classification**

• Computer scientists love all things Boolean

• So, we have Boolean attributes, Boolean output.

• How do we handle this scenario??

Components of ML (Task, Experience, error)

• What do you for a ML project?

• A well-defined **task** - what are you trying to teach to the machine?

• Training data that will be used by the machine – also called training **experience**.

• A way to check the performance of the machine – called **error** metric.

• A smart machine **learns** from the past errors and improves its performance.

What is ML?

• There is **a task** – generally involving prediction

-e.g.

. Will a student at UTD get internship?

. Will a stock go up?

. Will you play tennis?

. Will a person be approved for credit card?

. What is the price of a house?

. Which digit does a handwritten image represent?

• It has an associated **performance** measure i.e. how close is your prediction to actual value

• Error metric: If your hypothesis doesn’t match the actual value, penalize the model.

Text, whiteboard

Description automatically generatedThe above is the simplest error function.

**h(x)** denotes your hypothesis and

**y** is the actual value.

**D** is the dataset.

• The model learns from **experience** i.e. data.

- More data => Better performance - **Is it always true?** Only if data is meaningful i.e. it’s not **noise**

• If you could see the **entire population** (entire dataset), you would get the best learning model. - **Is it possible?** Can you poll the entire US population? Probably Not!! In this class, we will work with samples of data.

Definition “A computer program is said to learn from **experience E** with respect to some **task T** and some **performance measure P**, if its performance on T, as measured by P, **improves** with experience E.”

-- Tom Mitchell, Carnegie Mellon University

Simpler Definition

• Design of algorithms that **learn from data** and improve performance on a predictive task.

• Development of computational methods **using experience** to improve performance.

Common Theme of this class:

Chart, scatter chart

Description automatically generated- **More training** data is always good Provided data is **meaningful** and is **labeled** (in case of supervised learning). Think: Why does Google work so well?

When is ML a good choice?

• Learning used when

- A pattern exists

- There exists a correlation between predicted variable and features.

- We have data

•Supervised learning

- Unknown target function y = f(x) - You get data & labels (x1, y1) (x2, y2) ,…

- Learning algorithm picks a function g ≈ f (approximation)

Simple, but not obvious

• When can we use ML?

- There is some definite relation/correlation between X and Y.

i.e. there is a deterministic or highly probable function f

Chart, line chart

Description automatically generated- X comes from a well-defined distribution (we don’t need to know the details yet)

**=> X is not a set of random variables**.

Simple, but not obvious

• When can we NOT use ML?

- There is no correlation between X and Y. - There is no clear function f

- X is a set of random variables

=> Can ML predict the lottery? => Can you use it at the casino?

- Can we use it to predict the stock market?

Traditional Programming

. Getting computers to program themselves

. Writing software is the bottleneck, let data do the work

Diagram

Description automatically generated

Machine Learning

. In ML, The algorithm learns from data and outputs a program. For example, a self-programming robot.

Diagram

Description automatically generated

Types of Learning

. **Supervised (inductive) learning**

* Training data includes desired outputs

. **Unsupervised learning**

* Training data does not include desired outputs.
* Find hidden/interesting structure in data.

. **Semi – supervised learning**

* Training data includes a few desired outputs

. **Reinforcement learning** (Học tăng cường)

* The learner interacts with world via “actions” and tries to find an optimal policy of behavior with respect to “rewards” it receive from the environment.

Supervised Learning - Inductive

• **Inductive:** Tries to discover general concepts from a limited set of training examples. => Generalization

- Based on search of similar characteristics in different classes of examples. - e.g.

Given labeled examples, you find features that are common within each class.

• **Inductive**: goes from specific to general

- tries to obtain new knowledge. - new data points may force you to change old hypothesis

Calendar

Description automatically generated

Inductive vs Deductive

• **Deductive**: uses given premises and logical arguments to infer conclusions.

• tries to obtain knowledge that is implicit in original knowledge.

• Classic example:

\* All men are mortal. (**major premise**)

\* Socrates is a man. (**minor premise**)

\* Socrates is mortal. (**conclusion**)

• In this class, we will focus on **inductive** learning.

• Can you see any **issues** with **inductive** approach?

- What do you expect the data to be?

- How do you expect the learner to behave?

- When you make a conclusion, are you 100% sure or are you probably sure?

- Is it even possible to obtain 100% accuracy on training and test data? - …

Learning

• Inductive learning is **supervised learning** because the training data has the **class labels**.

=> That’s what you are trying to learn.

=> So you create an algorithm that separates the two classes based on features.

• What if you just want to find patterns in data i.e. how can you find similar items based on their attributes.

=> **Unsupervised learning**

=> **No labeling provided or you don’t care for labels**.

=> You want to create a natural grouping of people based on shared interests on FB.

Types of Learning task

. **Supervised**: correct output known for each training example

* Learn to predict output when given an input vector

. Classification: 1-of-N output (Speech recognition, object recognition, medical diagnosis)

. Regression: real – valued output (predicting market prices, customer rating)

. **Unsupervised** learning

* Create an internal representation of the input, capturing regularities/structure in data.
* Example: form cluster; extract features

. How do we know if a representation is good?

. **Reinforcement** learning

* Learn action to maximize payoff

Supervised Learning

Chart, scatter chart

Description automatically generated. **Classification**

You have N choices for output, learner needs to find best

. Outputs are categorical (1 -of -N)

. Inputs are anything

. Goal: select correct class for new inputs

Chart, scatter chart

Description automatically generated. EX: speech, object recognition, medical diagnosis

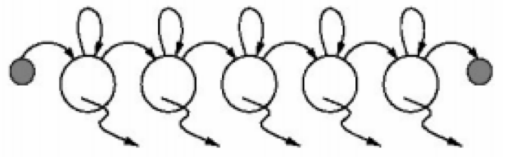
. **Regression**

You have infinite output choices

. Outputs are continuous

. Inputs are anything (typically continuous)

. Goal: predict output accurately for new inputs

. EX: predicting market prices, customer rating of movie.

**. Temporal Prediction**:

. Goal: perform classification/ regression on new input sequence values at future time points

. Given input values and corresponding class labels/ output at some previous time points.

Applications of ML

• Pattern Identification: - facial, fingerprint, gene sequence

• Differentiation: spam or non-spam cancerous or non-cancerous cells normal traffic or hacker traffic

• Finding association rules • Recommender Systems • … Many more

The machine learning framework

• Apply a prediction function to a feature representation of the image to get the desired output:

A picture containing shape

Description automatically generated

Diagram

Description automatically generatedThe machine learning framework

• **Training**: given a training set of labeled examples **{(x1,y1), …, (xN,yN)},** estimate the prediction function **f** by minimizing the prediction error on the training set

• **Testing**: apply **f** to a never before seen test example **x** and output the predicted value **y = f(x).**

Diagram

Description automatically generated

Steps

Formal notation of learning

• You own a credit card company.

• You get a lot of applicants, your job is to design the best classifier for approving them.

• You have some historical data to rely upon.

x j = (x1, x2,..., xn ) T Input vector for customer xj

X = {x 1 , x 2 ,..., xN } set of all customers

y = {0,1} Binary Output

f : X− > Y **Ideal** target function i.e. if you **had entire** data in front of you

(x 1 , y 1 ),(x 2 , y 2 ),...,(xn , yn ) You have this data

H = {h1, h2,..., hN } The set of all possible hypothesis. Doesn’t matter if they are meaningful or not

g : X− > Y,

g ∈ H The best hypothesis you can come up with given the data

Can you really know f?

- **No**, you can only try to approximate it by **g**

- Your approximation is only as good as the data that you see.

Diagram

Description automatically generated

Really simple, right?

• All you have to worry about is the type of learning model and how to generate (and eliminate hypotheses)

• In this class, we will focus on these types:

- Linear separators (**Perceptron**:a computer model devised to represent the ability of the brain to recognize and discriminate.)

- Extension of linear to non-linear (**SVM, ANN, etc**)

- Tree-based classifiers (**Starting with decision trees**)

Linearly Separable Data

• Suppose you have following data: (This is a toy example J )

Graphical user interface, application, Word

Description automatically generated

• 2-D case

Graphical user interface

Description automatically generated

Text

Description automatically generatedPerceptron

• It’s a simple linear classifier -> straight line in 2-D and a plane in higher dimensions.

• Suppose each instance is represented by the vector xj

Example of instance can be a student, a customer, etc. Attributes are x1, x2, …, xd

• Each attribute can have different weights w1, w2, …, wd.

• Equation of separator is:

Chart, scatter chart

Description automatically generatedA picture containing text

Description automatically generated

An easy way to separate classes is:

If A picture containing text

Description automatically generatedassign class1

else assign class 2.

**Perceptron** : a computer model or computerized machine devised to represent or simulate the ability of the brain to recognize and discriminate.

• An easier way to present this is:

Graphical user interface, text, application

Description automatically generated

Learning

• Let us look at another way of classification. • Decision Trees.

Decision Tree

• Toy example again

Diagram

Description automatically generated

• 2-D case

Diagram

Description automatically generated

• What can you infer?

• Think about decision trees as a way to learn **classification** function.

• Can also be thought of as rules:

eg: X1 ^ X2 -> 0 where x1 is Boolean GPA and x2 is Boolean exp<3

• OK…but which attribute should I choose to be at the top i.e. how do I choose attibutes??

Sounds great, but …

•Can you make some guarantee about the **unknown** data / parameters?

Let’s do an experiment. You have a bin and want to estimate the **probability of** **red marbles (μ)**. You draw **N samples** and observe **fraction of red marbles = ν** Is there any relation between μ, ν, and N?

Diagram

Description automatically generated

Hoeffding's Inequality

• It turns out there is a relation: For **large values of N**, the **probability of large error (difference)** between **ν and μ is bounded**.

Think: What happens when N becomes large, and when N is small?

Text, letter

Description automatically generated

**Inductive Learning**

• Inductive learning - **generalize** from a limited set of training data

• Training data is labeled

• You would like to estimate true separating function f

• Attributes of data (i.e. features) are important.

Learning from data

• Given: a set of labeled training examples: <x,f(x)>

Global f(x) is unknown to us - Distribution of x is unknown to us

• Find: An approximation of f(x)

Appropriate situations

• **Credit risk assessment**

x:properties of the customer and proposed purchase

f(x): approve purchase or not

• **Disease diagnosis**

x: Properties of patient (symptoms, lab, tests)

f(x): Disease (or maybe, recommended therapy)

• **Face recognition**

x: bitmap picture of person’s face

f(x): Name of the person.

Learning

• Improving with experience (E) at some task (T) with respect to some performance measure (P).

• **Experience** = Training data

**Task** = Any classification task (for this class, at least)

**Performance Measure** = Error value -> difference between true value and predicted value.

Model Representation

**What are you given in supervised learning?** A set of training examples and their labels (x(i) , y(i) )

\*\* It is assumed y(i) is generated by a true function f(x) \*\*

**What do you do with the training data?** Feed it to a learning algorithm that learns a function h, that is an approximation to f

Learning Process

use it as a guide After each training instance, refine h so that value of error metric goes down.

Diagram

Description automatically generated

Model Representation

**How do you know if h is good?**

Line chart

Description automatically generated with medium confidenceWe measure the error (overall) by using h Example:

Error = | f(x) – h(x) | or Error = 1/(2m)\* (f(x) – h(x))2 Think: Is more training data good? Always?

**LINEAR REGRESSION**

Learning a linear function

• Suppose we want to learn a function of the form:

ℎ θ (x) = θ0 + θ1x

to represent house price. Let's say it's a one-D problem and only independent attribute is house size x.

Linear Regression

• Linear Regression – find best model that fits a continuous (i.e. real valued) output

Chart, scatter chart

Description automatically generated

Error Function

• Our aim can be stated as: Choose parameters θ0& θ­1 ( such that our hypothesis ℎ θ (x) is as close to y for our training examples.

• Mathematically, choose parameters such that the following is minimized (called error or cost function). m is the number of training instance

Chart, surface chart

Description automatically generatedA picture containing text, clock

Description automatically generated

Error function

• How does J vary wrt the parameters

• Contour plot

• We are looking for the minima

• How do we get there?

Gradient Descent

• Given a function J of parameters ϴ, how do we find its minimum or maximum.

• Gradient Descent is a very powerful and popular algorithm.

• Widely used in machine learning

Chart, surface chart

Description automatically generated• In many cases, analytical solution is not possible, so we have to randomly take steps in search of minimum.

• Aim: We have a function 𝐽 (𝜃0, 𝜃1) , and we want

argmin 𝐽( 𝜃0 , 𝜃1)

𝜃0 𝜃1

**STEPS:**

• Start with some random values

• Keep changing these values such that you achieve a reduction in J

• Imagine a man at a random point on the mountains.

• He needs to reach the city by walking randomly

Text, letter

Description automatically generated

α is called the learning rate Intuition: It is how big a step you are taking.

Chart

Description automatically generatedIllustration

• In the curve on the right, imagine you are at point A

• The slope there is positive

• Update rule:

𝜃1 = 𝜃1 – 𝛼\* 𝜕𝐽/𝜕𝜃1

since 𝜕𝐽/𝜕𝜃1 is positive and α is always positive, we would move towards left.

Local Minima can be a problem

Chart, line chart

Description automatically generated

Gradient Descent for Linear Regression

Table

Description automatically generated

Gradient Descent for Linear Regression

**Hypothesis representation**

Diagram

Description automatically generated

**Gradient Descent for multiple variables**

Text

Description automatically generated with low confidence

**Partial derivative of cost function for multiple variables**

Text, letter

Description automatically generated

**Gradient descent for multiple variables**

Diagram

Description automatically generated with low confidence

Regression Evaluation Metrics

• Suppose we propose a linear model:

𝑌 = 𝛽0 + 𝛽1X + ϵ where 𝜖 represents the error.

• The coefficients 𝛽0 and 𝛽1 need to be estimated from the data (using gradient descent or other computational techniques).

• Let’s suppose our estimates are ^𝛽0 and ^𝛽1 ( , then the predicted value would be: ^𝑦 = ^𝛽0 + ^𝛽1 𝑋

• 𝑒i = ^y­i − yi represents the residual or error for the ith data point.

• Residual sum of square (RSS) is defined as:

𝑅𝑆𝑆 = e12 + e22 + …. + en2,

• By minimizing the RSS, we can arrive at the estimates ^𝛽0 and ^𝛽1.

Another evaluation metric

• We would like to check what fraction of data variance is explained by the model.

• R2 statistic measures this:

R2 = 1 – RSS/TSS

where **RSS** is the **residual sum of squares** (defined earlier) and **TSS** is **the total sum of squares**:

𝑇𝑆𝑆 = A picture containing text

Description automatically generated

Practice question:

. Consider the problem of predicting the number of A grades that a student a UTD will obtain in second year of M.S. based on the number of A grades obtained in the first year of M.S. course.

Below is the data:

A picture containing text, light, roof

Description automatically generated

x represents the number of A grades in 1st year

y represents the number of A grades in 2nd year

You decide to use a hypothesis of the form ℎ θ (x) = θ0 + θ1x where θ0 = 0 and θ1 = 1 . Find the value of the square error? <https://www.coursehero.com/qa/wait/43823259/>

**DECISION TREE**

**What is a DT?**

.

Decision Tree

■ 2-D case