Introduction to Machine Learning (MSC527)

Instructor: Dr. Rashmi Singh

Course Plan

Pre-requisites: It is good to be familiar with probability, linear algebra, matrices, calculus and programming.

Course Objective: In this course, students will be exposed to the fundamentals of machine learning and some popular machine learning algorithms. It will help students to understand the relevance of machine learning algorithms & its application in various domain. In this course, it is required that students should be familiar with the programming & coding (R, Matlab, Python, Octave) as it will be helpful to them for their assignments.

Learning Outcomes:

- To understand the concept of Machine Learning
- To develop understanding in various types of machine learning algorithm
- To develop the skill in application software like Python or R or Octave for solving business application problems through machine learning.

Logistics

Instructor: Rashmi Singh
Email: rashmis@iitism.ac.in
Office hours: Monday to Friday (10:00 AM-6:00 PM)

Reading: Mandatory reading materials will be uploaded on the MIS.

Textbooks:

- Understanding Machine Learning. Shai Shalev-Shwartz and Shai Ben-David. Cambridge University Press. 2017.
- The Elements of Statistical Learning. Trevor Hastie, Robert Tibshirani and Jerome Friedman. Second Edition. 2009.

Reference Books:

- Foundation of Data Science. Avrim Blum, John Hopcroft and Ravindran Kannan. January 2017.
- Pattern Recognition and Machine Learning. Christopher Bishop. Springer. 2006.
- Machine Learning. Tom Mitchell. First Edition, McGraw-Hill, 1997.

Course Syllabus

Module I: Introduction, Basic Principles, Applications, Challenges [4L]

Module II: Supervised Learning, Linear Regression (with one variable and multiple variables), Gradient Descent; Classification (Logistic Regression, Overfitting, Regularization, Support Vector Machines); Artificial Neural Networks (Perceptron, Multilayer networks, and back-propagation); Decision Trees [14L]

Module III: Unsupervised Learning, Clustering (K-means, Hierarchical); Dimensionality reduction; Principal Component Analysis; Anomaly Detection. [10L]

Module IV: Theory of Generalization, In-sample and out-of-sample error, VC inequality, VC analysis, Bias and Variance Analysis. [6L]

Module V: Applications, Spam Filtering, recommender systems, and others. [5L]

Evaluation Components (subject to change)

- Mid-Term Exam: 32%
- End-Term Exam: 48%
- Quizzes: (10%) Two quizzes will be conducted (subjective/multiple-choice questions) type.
- Assignments: (10%) deadline for submission of assignments will be announced with the assignment.
  - Assignments will contain programming questions

Academic Integrity Policy

- You are free to discuss the assignment problems with other students in the class. But all your code should be produced independently without looking at/referring to anyone else's code. Also add comments in the program to explain its purpose at each and every step.

- Python (preferred), and MATLAB or R can also be used for the programming in the course. You can use one of these for your assignments unless otherwise explicitly allowed.

- Honour Code: In cases of copying, students will be awarded a zero marks in the assignment or a penalty of -10. More severe penalties may follow.

- Late Policy: You are allowed to delay your assignment by 3 days at the max and you will get a zero in assignment once you exceed the (total) allowed exemption of 3 days.

- If you refer to any external material, always cite your sources.
  - Follow proper citation guidelines.

Sources: Datasets, Journals & Conferences

Datasets

- Kaggle: https://www.kaggle.com/datasets
- http://www.kdnuggets.com/datasets/index.html
- UCI Repository
- http://www.ics.uci.edu/~mlearn/MLRepository.html
- UCI KDD Archive:
- http://kdd.ics.uci.edu/summary.data.application.html
- Delve: http://www.cs.utoronto.ca/~delve

Journals

- Pattern Recognition Letters
- Nature Machine Intelligence
- Journal of Machine Learning Research www.jmlr.org/
- Machine Learning
- IEEE Transactions on Neural Networks
- IEEE Transactions on Pattern Analysis and Machine Intelligence

Conferences

- International Conference on Machine Learning (ICML)
- European Conference on Machine Learning (ECML)
- Neural Information Processing Systems (NIPS)
- Computational Learning
- International Joint Conference on Artificial Intelligence (IJCAI)
- ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)
- IEEE Int. Conf. on Data Mining (ICDM)

Introduction

- Motivation for Machine Learning
- What is Machine Learning
- Applications of Machine Learning
- Different Types of Learning Mechanism
- Challenges in Machine Learning

A Few Quotes

- “A breakthrough in machine learning would be worth ten Microsoft” (**Bill Gates, Chairman, Microsoft**)
- “Machine intelligence is the last invention that humanity will ever need to make” (**Nick Bostrom**)
- “A baby learns to crawl, walk and then run. We are in the crawling stage when it comes to applying machine learning” (**Dave Waters, Founder & Director, Paetoro’s**)
- Machine learning is going to result in a real revolution” (**Greg Papadopoulos, CTO, Sun**)
- “Machine learning is today’s discontinuity” (**Jerry Yang, CEO, Yahoo**)

Slide Credit: Pedro Domingos

What is Machine Learning?

- Ability of systems to “learn” directly from “example”, “data” or “past experience.”

“Machine Learning: field of study that gives computers the ability to learn without being explicitly programmed.”

- Arthur Samuel (1957)

“Learning is any process by which a system improves performance from experience.”

- Herbert Simon

What is Machine Learning? (Cont.)

Definition by **Tom Mitchell (1998)**:

Machine Learning is the **study of algorithms** that

- improve their performance P
- at some task T
- with experience E.

A well-defined learning task is given by <P, T, E>.

Tom M Mitchell et al. “Machine learning. 1997”. In: Burr Ridge, IL: McGraw Hill 45.37 (1997), pp. 870–877.

Question 1:

- Suppose your email program watches which emails you do or do not mark as spam, and based on that learns how to design spam filter. What is the task ‘T’ in this setting?

Options:

- Classify emails as spam or not spam
- Watching your label emails as spam or not spam
- The number (or fraction) of emails correctly classified as spam/not spam.
- None of the above, this is not a machine learning algorithm

Also mention ‘P’ and ‘E’ for the same problem.

When Do We Use Machine Learning?

- Human **expertise does not exist** (**navigating on Mars**)

- Humans **can't explain their expertise** (**speech recognition**): few tasks cannot be defined well
- Solution changes in time (**routing on a computer network**): **environment change over time**

- **Solution** needs to be adapted to particular cases (**user biometrics**)
- Models must be **customized** (**personalized medicine**)
- Models are based on **huge amounts of data** (**genomics**): relationships and correlations can be hidden within large amounts of data
- New **Knowledge can always be discovered** thus, it may be difficult to re-design systems by hand

Slide Credit: E Alpaydin

Different Types of Learning Mechanism

- **Supervised (Inductive) learning**

- Given: training data + desired outputs (labels)

- **Unsupervised learning**

- Given: training data (without desired outputs)

- **Semi-supervised learning**

- Given: training data + a few desired outputs (labels)

- **Reinforcement learning**

- Rewards from sequence of actions

Slide Credit: Pedro Domingos

Supervised or Inductive Learning

- **Given** examples of a function *(X, F(X))*

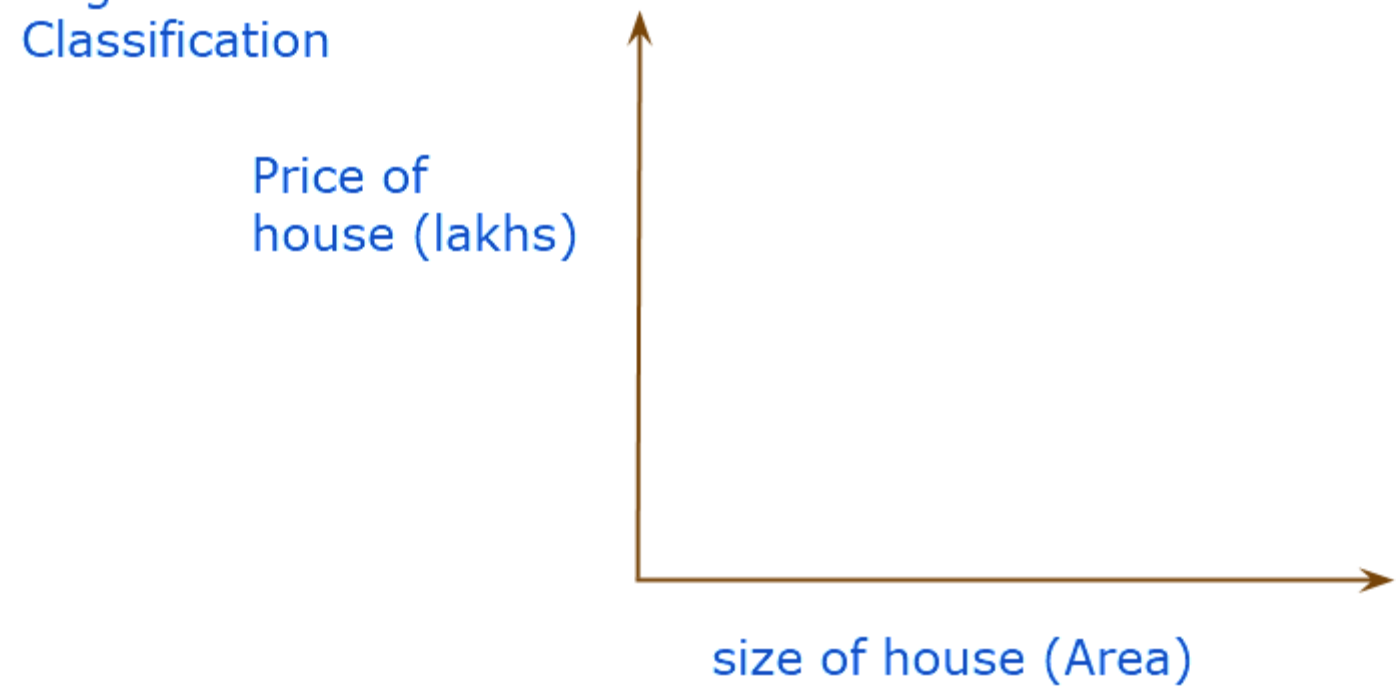
- **Predict** function *F(X)* for new examples *X*
  - **Discrete** *F(X)*: Classification
  - **Continuous** *F(X)*: Regression
  - *F(X)* = Probability(*X*): Probability estimation

Slide Credit: Pedro Domingos

Regression
Classification

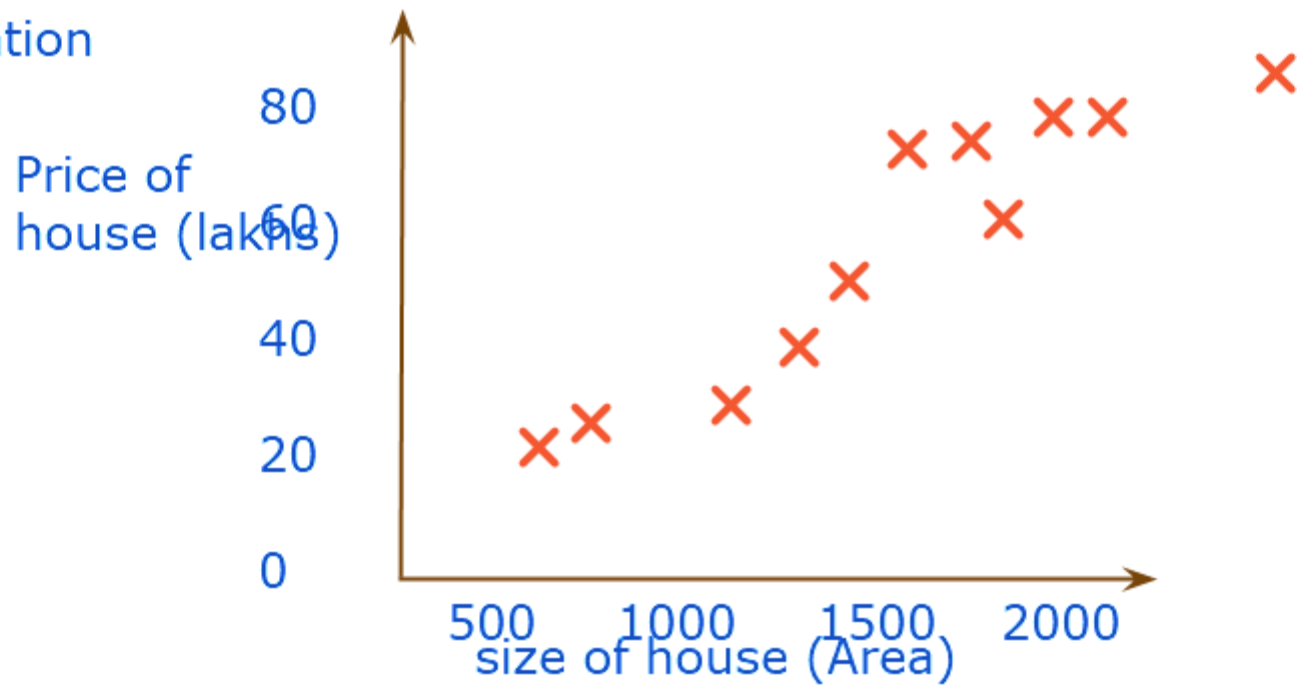
Regression
Classification

Regression
Classification



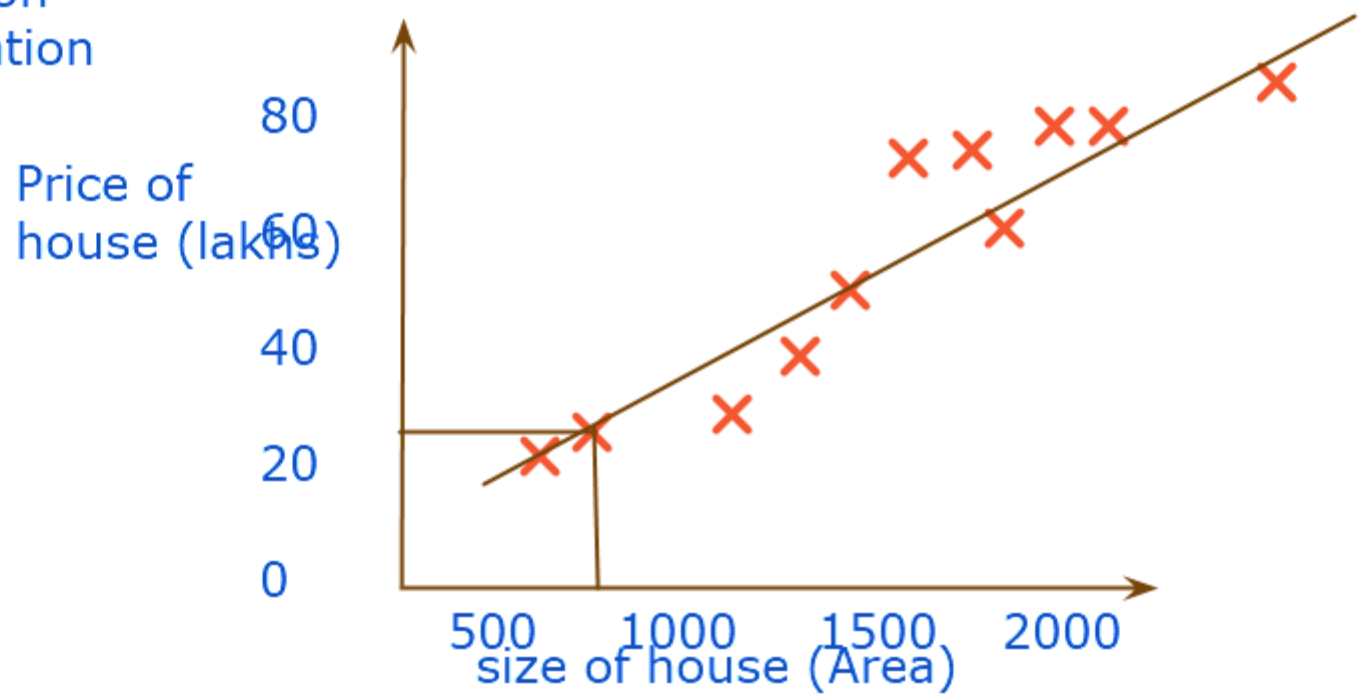
Regression
Classification

predict the price of the house



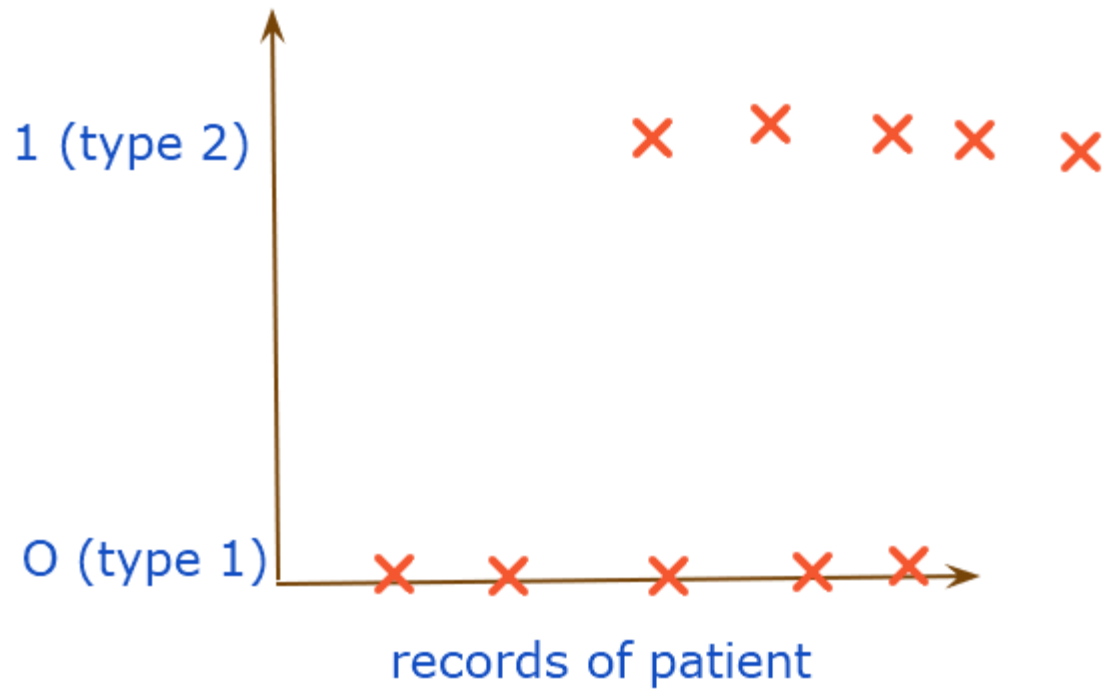
Regression
Classification

predict the price of the house



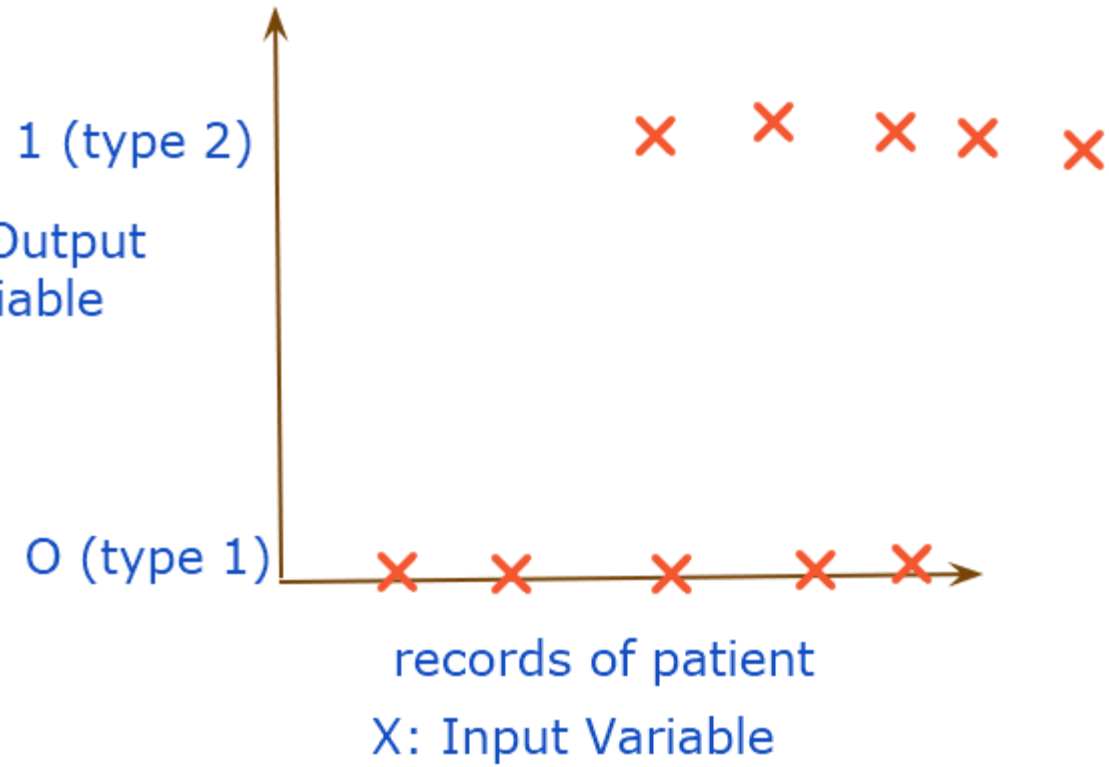
output variable = continuous variable:
Regression

output variable = continuous variable:
Regression



output variable = continuous variable:
Regression

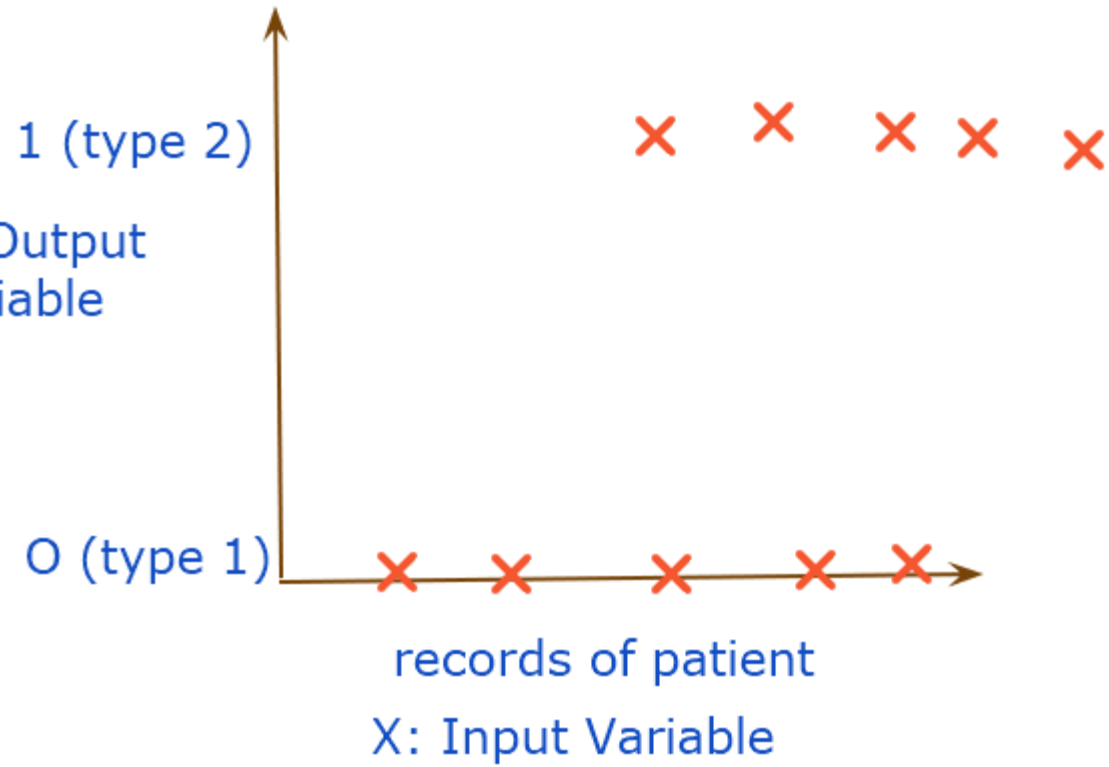
Output/Right answer



output variable = continuous variable:
Regression

Output/Right answer

output variable = discrete
variable



output variable = continuous variable:
Regression

Output/Right answer

output variable = discrete
variable

Stock price: Regression
Is it cat or dog: Classification
how's weather today :
classification
unhappy customers:
Classification