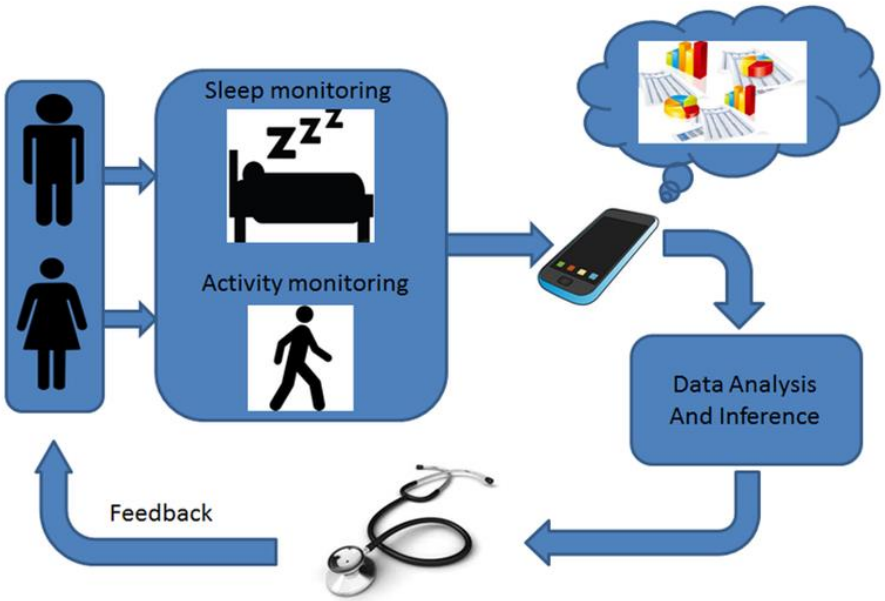
CS 7830
Machine Learning

Dr. Tanvi Banerjee

My Introduction

My Introduction



Data Types My Lab Works With

- Imagery
- Sensors (accelerometers and acoustic)
- Text

Data Fusion: Can we combine these heterogeneous data sources for actionable information?

Introduction

Are you familiar with Machine Learning?

If so, which topics have you learnt previously?

Introduction

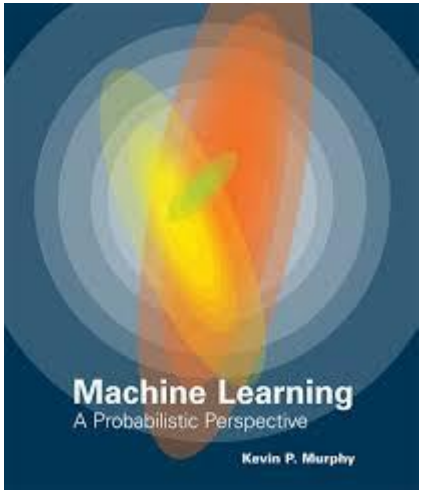
What are your research interests? Are you doing a thesis?

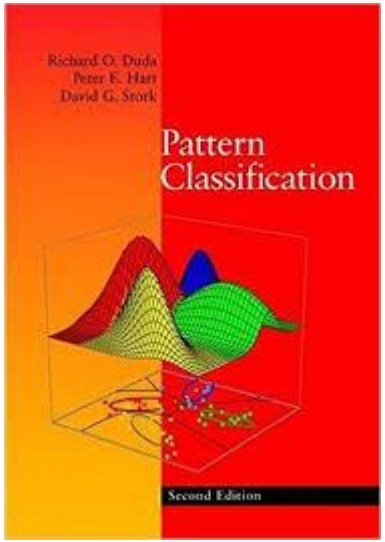
Course Outline

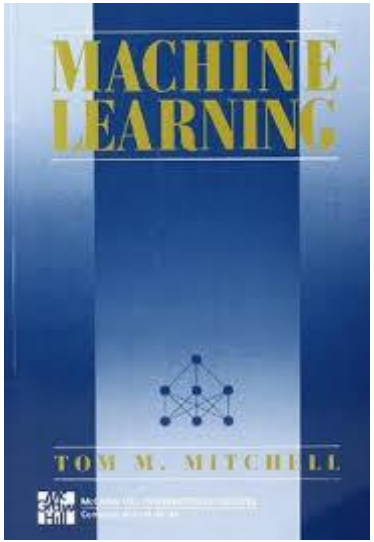
Assignments* : 30% of the overall grade
Participation**: 10% of the overall grade
Midterm : 25% of the overall grade
Project : 35% of the overall grade

Textbook

No textbook!







Course Outline

Reading Materials

- Slides will contain key concepts, but you need to read additional material
- The assignments and exams will be more concept based so choice of textbook is open
- Free book link
- Supplementary reading material will also be provided as we move through each topic

Mathematics and Statistics

Should have a preliminary understanding of probability theory and calculus

Fuzzy theory will build on the traditional set theory, so need to revise: boolean logic, as well as set theory notation

Participation

- Paper discussion
- **Pop quizzes**
- In-class discussion
- Attending the final presentations

Assignments

Please don’t copy. You won’t learn anything if you do.
Start your assignments and projects early.
Assignments will be thought provoking: the questions are meant to make you think and analyze, so not just coding

Programming Language

Any Programming language

- Need to have a strong background in that language

Will need to submit codes as well as results (graphs accuracy percentages, depending on assignment) and **analysis** along with the project report.

Final Project

Teamwork is OK (encouraged).

For projects, 3 is a good number. 2-4 are OK (scope has to be bigger for larger teams).

Start your projects early.

Ask for comments and feedback on projects. Can we beat the Stanford class?

[See sample project reports]

http://cs229.stanford.edu/projects2012.html

Final Project

Resources:

Data from your research

https://archive.ics.uci.edu/ml/datasets.html .

https://www.kaggle.com/

https://www.synapse.org/#!Synapse:syn8717496/wiki/422884

Learning Strategies

- Need basic understanding of Calculus and Algebra
- Group study
- Taking notes (slides will be made available)
- Regular reading of topics

  We will have discussions online as well as in class
- Feedback please!

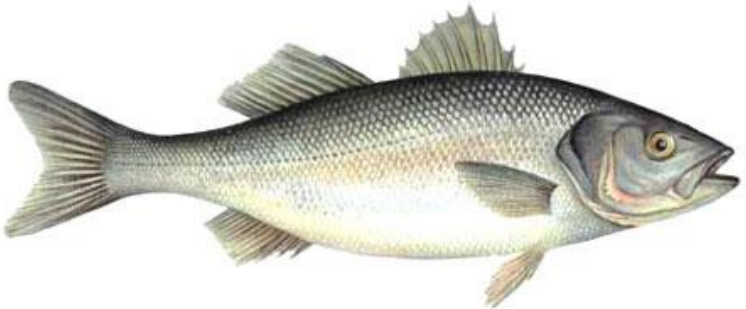
  Let me know if you have comments, concerns, or suggestions!

So, let’s begin!

What is Machine Learning?

What is Machine Learning?

Identify salmon.



What is Machine Learning?

If you want your program to identify salmon (task T), you can run it through a machine learning algorithm with data from the past that contains information from different types of fish (experience E) and, if it has successfully “learned”, it will then do better at identifying salmon (performance measure P). -- Tom Mitchell

Supervised & Unsupervised

- Supervised Learning
  - A model “learns” the data using known labels and is then used to classify the new data
- Unsupervised Learning
  - A model “groups” the data into different categories using the characteristic distribution but without any labels to help

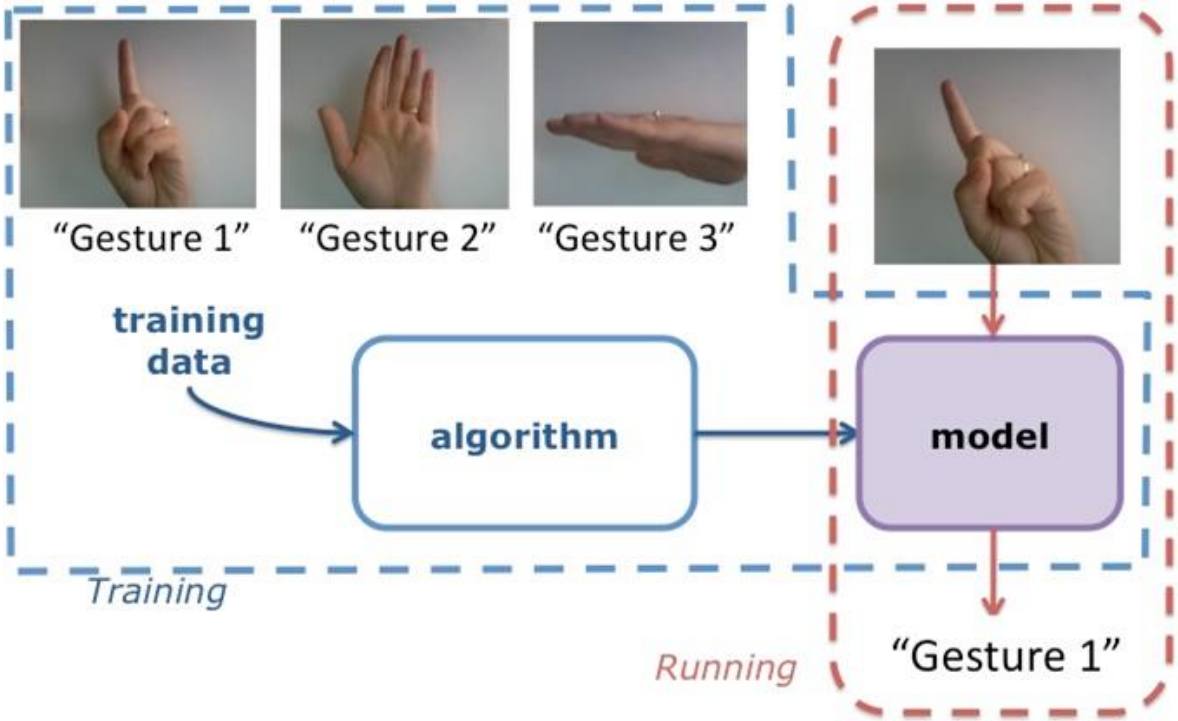
Categories Within Supervised ML

Classification machine learning systems: Systems where we seek a yes-or-no prediction, such as “Is this tumor cancerous?”, “Does this cookie meet our quality standards?”

a. Binary or Multiclass classifier : Output y in {-1, 1} or y in {1,..k}

Regression machine learning systems: Systems where the value being predicted falls somewhere on a continuous spectrum. These systems help us with questions of “How much?” or “How many?”.

Supervised Learning



source: https://code.google.com/p/wekinator/

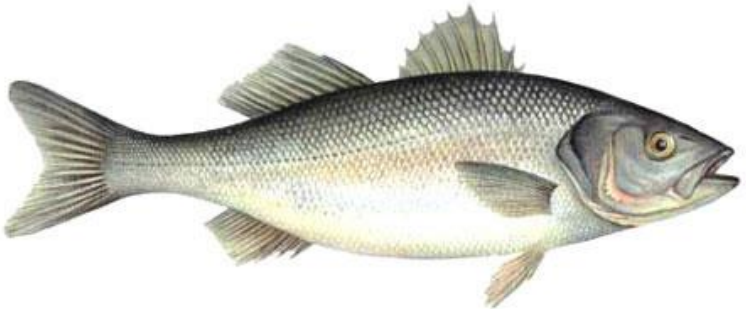
Some Terms in Supervised Learning

- Independent variables/ Input variables/ Features (X)

- Target variables/ Output variables/ Dependent variables (Y)

Example

Independent variable? Target variable?



Some Terms in Supervised Learning

Training Data

Data used to “learn” the relationship between the independent and target variables

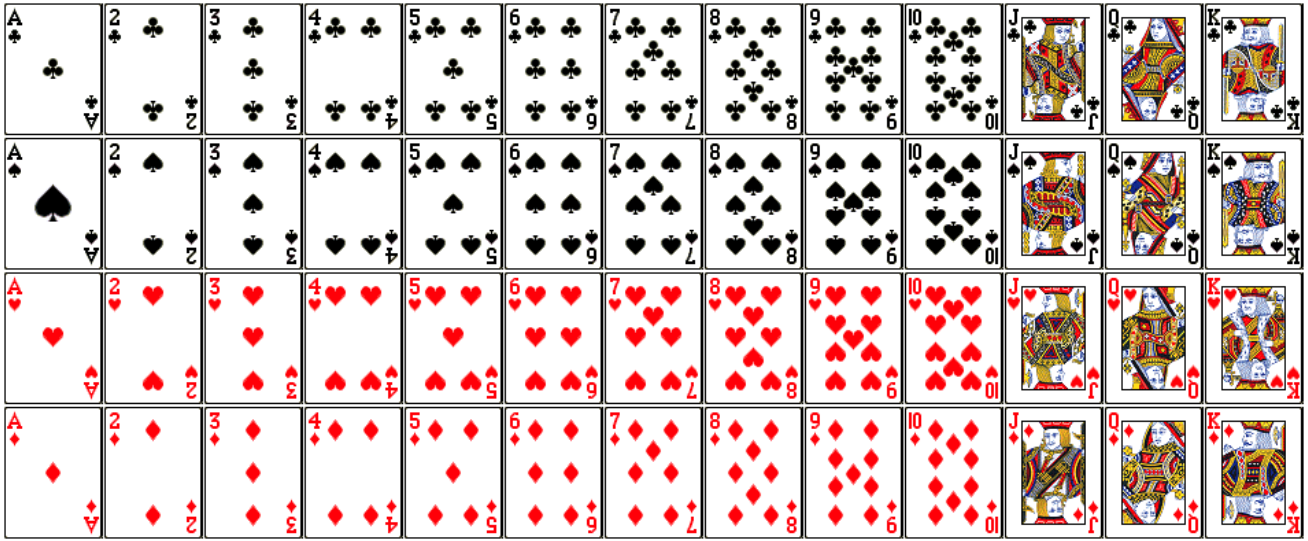
“gold standard”, contains labels

Test Data

Estimate the accuracy using the unseen data (validate using some performance metric P)

Unsupervised

How many groups are there in a deck of cards?

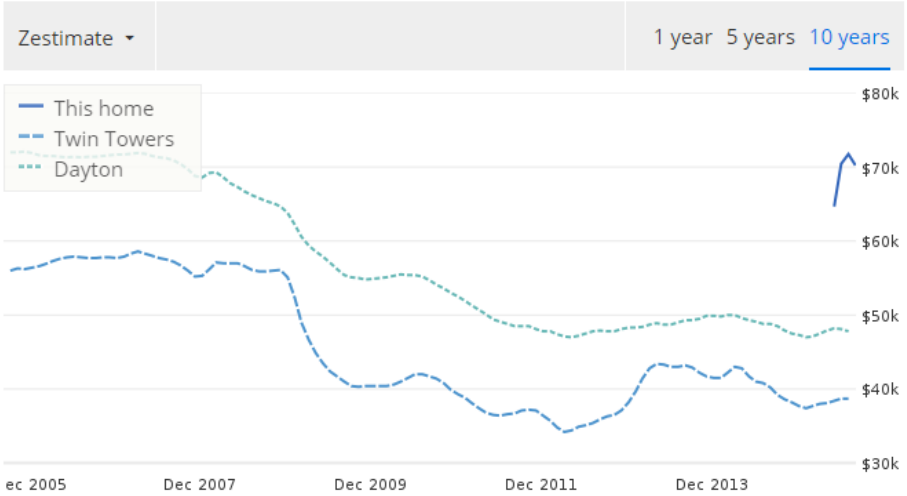


Other Approaches

1. Semi-supervised learning, in which only a subset of the training data is labeled
2. Time-series forecasting / prediction, such as in financial markets

Prediction

What will be the house price in 2016?



Other Approaches

1. Semi-supervised learning, in which only a subset of the training data is labeled
2. Time-series forecasting / prediction, such as in financial markets
3. Anomaly detection such as used for fault-detection in factories and in surveillance
4. Active learning, in which obtaining data is expensive, and so an algorithm must determine which training data to acquire and many others.